Title: "What is it like to be a bat?" - a pathway to the answer from the Integrated Information Theory

By Naotsugu Tsuchiya[1,2]

1. School of Psychological Sciences, Monash University, Melbourne, Australia, 3168

2. Monash Institute of Cognitive and Clinical Neuroscience, Monash University, Australia, 3168

May 10, 2016

Abstract:

What does it feel like to be a bat? Is conscious experience of echolocation closer to that of vision or audition? Or, do bats process echolocation non-consciously, such that they do not feel anything about echolocation? This famous question of bats' experience, posed by a philosopher Thomas Nagel in 1974 (Nagel, 1974), clarifies the difficult nature of the mind-body problem. Why a particular sense, such as vision, has to feel like vision, but not like audition, is totally puzzling. This is especially so given that any conscious experience is supported by neuronal activity. Activity of a single neuron appears fairly uniform across modalities, and even similar to those for non-conscious processing. Without any explanation on why a particular sense has to feel the way it does, researchers cannot approach the question of the bats' experience. Is there any theory that gives us a hope for such explanation? Currently, probably none, except for one. Integrated Information Theory (IIT), proposed by Tononi in 2004 (Tononi, 2004) has potential to offer a plausible explanation. IIT essentially claims that any system that is composed of causally interacting mechanisms can have conscious experience. And precisely how the system feels is determined by the way the mechanisms influence each other in a holistic way. In this article, I will give a brief explanation of the essence of IIT. Further, I will briefly provide a potential scientific pathway to approach bats' conscious experience and its philosophical implications. If IIT, or its improved or related versions, is validated enough, the theory will gain credibility. When it matures enough, predictions from the theory, including nature of bats' experience, will have to be accepted. I argue that a seemingly impossible question about bats' consciousness will drive empirical and theoretical consciousness research to make big breakthroughs, in a similar way as an impossible question about the age of the universe has driven modern cosmology.

1. Introduction

The title of Thomas Nagel's 1974 article "What is it like to be a bat?" articulates the immense difficulty of the mind-body problem. Bats sense the outside world by what is called "echolocation" (Jones, 2005). They produce a sound, receive its echo in virtue of which they detect the presence of a prey at a certain distance and direction. Despite extensive investigation into echolocation in terms of ecology and neural mechanisms, we have no idea what it is like to be a bat. Do bats experience echolocation as closer to their visual or auditory experience? Or do they not feel anything, like our non-conscious processing?

We know that bat brains and our brains are composed of neurons. Each neuron excites or inhibits other neurons that are connected via synapses. These biophysical principles are conserved across biological species. We also know that our conscious experience is generated by neurons in the brain, though we don't know exactly how. If we knew the principles for how various conscious experiences are generated in a human brain, it should be possible to understand[1] what kinds of experiences are generated in a bat brain.

Currently, most neuroscientists have no idea about what those principles could possibly be. To come closer to such principles, we might need to know a lot more about the brain. From microscopic to macroscopic levels, there are countless questions, and neuroscientists worldwide are tackling them everyday. It might take another 10, 20 or 100 years to come up principles for the mind-body problem. It is also plausible, however, that it is not the lack of knowledge that keeps us from a solution, but rather that there is a crucial idea that is missing: an idea that can dissolve the mystery that stands in between consciousness and the brain.

---

[1] I will clarify what I mean by "understand" here in later sections.

To address bat consciousness, what we need is a theory that can tell us “this is what it’s like to be a bat” if we understand all physical properties of the bat brain. Specifically, the theory should consist of a set of laws, which jointly translate information about the brain (connectivity and a pattern of neural activity) into a subjective experience. The theory should be empirically testable and falsifiable in some way.

Since it's impossible to become a bat, one may conclude any such theory is untestable and unfalsifiable. Surely, we cannot directly test a theory on bat consciousness, but neither can we directly test theories on how the universe started or how the life has emerged and evolved. The theories of the latter kind, however, are considered testable and falsifiable, because we have scientific constructs, such as relativity and quantum mechanics, biochemistry and DNA, and indirect evidence like astronomical observations and the fossil record, which give us answers to questions that are not directly testable. The only difference is that these theories have made many predictions in the past, and they have been supported by accumulated evidence over time, to the extent that seemingly untestable predictions are accepted. In this article, I propose to take a similar approach for consciousness research; to empirically test a promising theory and to refine it to the limit so that we can approach the seemingly untestable question of what it’s like to be a bat. I will focus on integrated information theory (IIT) (Tononi, 2004), which makes many qualitative predictions that are empirically testable.

What should we expect from a theory of consciousness and the brain as a starting point? Given that we can experience only our own consciousness, the theory has to explain all enigmatic features of the relationship between our own consciousness and the brain. To give examples of these mysteries: Why do I lose consciousness when I sleep or go under general anesthesia? Why is the activity in some parts of my brain (e.g. the cerebellum) seemingly irrelevant to what I am experiencing now? Why are any two moments of visual experience much more similar to each other than visual and auditory experiences?

Even if a theory provides answers to these problems, it is not enough. We should also expect the theory to explain and predict the conscious experience of other persons purely based on his/her neural connectivity and activity. To test the validity of the theory's explanation and prediction, rare forms of conscious experience will be most informative. For example, synesthetic experience (e.g., seeing color when hearing sound (V. Ramachandran & Hubbard, 2001) and substituted sensory experience (e.g., seeing through the auditory modality (Bach-y-Rita & S, 2003)) are hard cases to imagine what it feels like. So far, none of these phenomena have been theoretically explained based on connectivity and activation states of the neural system. An ideal theory should be able to predict who experiences what kinds of experience, just based on the brain data, without a need to ask their experience.

Further, the theory should explain and make predictions about animal consciousness, which is not directly verifiable by us. With certain animal species, however, there are strong cases to believe in what trained animals reports about their percept. For example, macaque monkeys can be trained to report on their own percept while viewing ambiguous stimuli, such as binocular rivalry. Of course, we cannot trust their reports as they are just by the fact that they can report percept in such a situation. However, when these ambiguous trials can be interleaved with unambiguous trials. In these unambiguous trials, stimulus characteristics are carefully manipulated to reveal highly homologous behavioral performance to humans'. Techniques such as this have strengthened the case to believe that monkeys and humans have similar visual experience in various situations (Leopold, Maier, & Logothetis, 2003; Wilke, Logothetis, & Leopold, 2006). Trained rats can show the evidence of their ability for "metacognition". In a sensory discrimination task, when they are given an option to "skip" a trial in addition to the two alternative forced choices, rats do skip more trials when the stimulus is ambiguous and decision is more difficult {Kepecs, 2008 #2307}. Based on fine details of neural connectivity and activity patterns in these animals, the theory of consciousness should predict conscious perception and metacognition in these animals, which fill strongly validate the theory.

We would gain more confidence in accepting what a theory predicts about bat consciousness if it can withstand the critical validations through human and animal testing suggested above. If successful, there would be no real difference between what we can accept about the predictions of the beginning of the universe, evolution of life and the consciousness of a bat.

2. Puzzles of consciousness and the brain

In this section, I will consider several candidate theories that aim to explain what we know about the relationship between consciousness and the brain. The more we learn about the facts about neurons and brains, the more puzzled we become about how brains generate consciousness. For example, people who know little about brains may assume that we lose consciousness when we sleep because the brain turns off like electrical equipment. However, according to various measures, brains during deep sleep without dreams are far from inactive (Dang-Vu et al., 2008; Schabus et al., 2007). Some brain-damaged patients who recovered from loss of consciousness may show very low metabolic activity, while other patients who remain unconscious can show high levels of metabolism. Thus, any theory that tries to explain consciousness simply based on the degree of neural activity fails to provide a reasonable explanation (Massimini & Tononi, 2015).

In the face of this puzzle, scientists have suggested that some form of "complexity" is necessary and sufficient for consciousness. But our experience is generated only by a part of the brain, and this fact is difficult to explain by complexity theories. When critical parts of the cortex and thalamus (a connectivity hub beneath the cortex) are impaired, we lose consciousness entirely (Bogen, 1995; Laureys, 2005). On the other hand, restricted injury to specific parts of the cortex can lead to loss of specific kinds of content, such as loss of some aspects of vision (V. S. Ramachandran & Blakeslee, 1998). Further, loss of a cerebellum, which contains four times more neurons than the cortico-thalamic system (Herculano-Houzel, 2012), hardly affects any aspects of consciousness (Lemon & Edgley, 2010; Yu, Jiang, Sun, & Zhang, 2015). As long as the cortico-thalamic and the cerebellar system cannot be distinguished in terms of "complexity", complexity explanations are far from satisfactory. Further, complexity theories also fall short of explaining contents of consciousness. Seeing, hearing, and touching are all supported by neurons in the cerebral cortex and the thalamus. Just how vision could be distinguished from audition in terms of complexity is very unclear: is vision more complex than audition? Less? Is it a different kind of complexity? If so, what could it be that makes both kinds of complexity different kinds of conscious?

As a neural correlate of consciousness, more specific forms of interactions between neurons have also been proposed, such as synchronous activity of neurons (Engel & Singer, 2001), global availability of information (Dehaene, 2014), and recurrent feedback activation (Lamme, 2010). These processes are all suited to sustain neural activity for a short-term and to facilitate communication across distant brain areas. However, these can be observed during loss of consciousness as in dreamless sleep or general anesthesia, and in the non-conscious cerebellum. More critically, they have no specificity to explain the distinct phenomenology between different senses. Why does vision feel like vision? Within vision, why does color feel different from shape, despite both being generated in the visual cortex? Whatever is critical for consciousness should be specific for the cortico-thalamic system during the awake and the dreaming state, and should be differentiable in ways that allow us to understand different modalities and their particular characteristics. What is this critical factor?

All the neurons in the brain operate under the same principle; they are connected with other neurons, receiving and sending electrical signals called spikes. The brain regions that are responsible for visual discrimination of colors and auditory discrimination of pitches both use the same spike mechanisms. Why do we experience these differently, if they are supported by the same mechanisms? Similarly puzzling is the fact that much of neural activity, even within the cerebral cortex, does not correlate with any aspects of phenomenology (Koch, 2004). What are the differences between neural activity resulting in consciousness and neural activity resulting in unconsciousness? Without a theory that can account for all these problems, we are very far from making reasonable predictions about what it is like to be a bat.

3. Integrated information theory in a nutshell

One of the most promising approaches available at the moment is the Integrated Information Theory (IIT), originally proposed by Tononi in 2004 (Tononi, 2004). IIT indeed has claimed that it would address the problem of bat consciousness, if sufficiently developed (e.g., p229 in (Tononi, 2008)). The original theory has undergone several revisions over the years, especially in its mathematical formulations (Balduzzi & Tononi, 2008, 2009; Oizumi, Albantakis, & Tononi, 2014; Tononi, Boly, Massimini, & Koch, 2016), but the core ideas remain the same.

IIT gives adequate explanations for all the problems raised in the previous section. In sum, IIT proposes that a system that is composed of multiple causal mechanisms that influence each other will experience something. Contents of consciousness (also known as qualia (Balduzzi & Tononi, 2009; Kanai & Tsuchiya, 2012)) are determined by the way these mechanisms causally interact with one another. This only gives an intuitive idea of IIT. To precisely understand the IIT, one needs to read through math-heavy papers (Balduzzi & Tononi, 2008, 2009; Hoel, Albantakis, & Tononi, 2013; Oizumi et al., 2014; Tononi et al., 2016). However, an intuitive understanding of IIT is enough for my purpose here, which is to provide a pathway to approaching the consciousness of bats.

IIT starts from seriously considering one's own phenomenology[2]. The theory identifies five fundamental properties of consciousness (Oizumi et al., 2014; Tononi, 2015): 1) *existence*: consciousness exists intrinsically and a conscious subject cannot doubt one's ongoing experience; 2) *composition*: any experience is composed of various modalities (e.g., vision, audition) and various aspects within each modality (e.g., visual motion, faces and objects, colors within objects); 3) *information*: one moment of consciousness is extremely "informative" and differentiated to an experiencing subject, in a sense that one experience excludes all other potential experiences that the subject could have had at that moment; 4) *integration*: parts of a conscious experience are bound together and experienced as a whole, that is, different aspects of an experience are not experienced separately but always as integrated parts of one unified whole (e.g., one cannot separate auditory experience and visual experience); and 5) *exclusion*: consciousness has a

---

[2] Starting from phenomenology and proposing neural mechanisms is a highly distinguishing strategy of IIT. Most other approaches for consciousness, as reviewed in section 2, start from observing the neural activity in experimental situations, then try to think how such neural activity gives rise to consciousness. That pathway of explanation (neuron -> consciousness) may be very Hard (Chalmers, 1996), but possibly not the other way around as taken by IIT.

definite spatiotemporal grain – it flows at a definite speed and has a definite scale – and no other overlapping conscious experience exists at another scale or speed. Any phenomenal distinction that does not meet the spatiotemporal grain (e.g., too fast or too slow) is excluded and not experienced.

IIT attempts to discover the physical mechanisms that can support these phenomenological properties. The exact forms of the postulated mathematical expressions of these mechanisms have evolved across the versions of the theory (Balduzzi & Tononi, 2008, 2009; Oizumi et al., 2014; Tononi, 2004). Common to all is that they involve the critical notion of integrated information, usually denoted as "ϕ" (phi). The more advanced and updated version has more sophisticated mathematical formulations, but it comes at a cost for intuitive understanding and for feasibility of validation through experiments. Below, I briefly explain the core features of IIT using a framework based on the second generation of the IIT (Balduzzi & Tononi, 2008; Oizumi, Amari, Yanagawa, Fujii, & Tsuchiya, 2016). The second generation lacks some theoretically important aspects (e.g., distinction between cause in the past and effects in the future) implemented in the third generation, but it has several advantages. Most importantly for our purposes is that it is easier to understand through simple numerical examples, such as the one given below. Further, it is much more feasible to compute integrated information patterns from empirical neuronal recordings (Haun et al., 2016). These properties make the second generation of IIT perfectly suitable for the purpose of this paper.

[Figure 1 around here]

To explain how the concept of integrated information, ϕ, captures the fundamental properties of consciousness (Oizumi et al., 2014; Tononi, 2004, 2008, 2012), let's consider the simple example given in Figure 1. Figure 1a depicts all four possible states (1-4) of a system, composed of two neurons. Each neuron is either "on" or "off" at any time. Each neuron copies the state of the other with a time delay (τ). In this situation (with the connectivity and the rule for each neuron's firing), if the present state is "off-on" (Figure 1b, right), then the past state (Figure 1b, left) must have been "on-off".

In an information theoretic jargon, the present state is said to remove uncertainty about the past. If the present state is unknown, uncertainty about the past state is maximal; the four states of this system (i.e., on-on, off-off, on-off, or off-on) are equally likely. We can quantify the degree of uncertainty with a concept of entropy (H). Entropy quantifies possible variability of the system (usually with the logarithm with a base of 2 of the number of possible states of the system). Here, $H=\log_2(\text{the number of possible states})=\log_2(4)=2$. The remaining uncertainty after knowing the present state is called conditional entropy (H*) and $H^*=\log_2(1)=0$.

Now, the concept of "information" can be defined as reduction of uncertainty. The more information you have, the less uncertain you are. Another mathematical concept, called mutual information, I, is defined as H-H* to capture this idea formally. In the above case, mutual information, I, between the present state and the past state is $I = H - H^* = 2$.

Integrated information (ϕ) is the difference between the information derived from the whole system (I) compared with the sum of the information arising from its parts (I^*): $\phi = I - I^*$. In the above example, if the system is cut into two parts (Figure 1c), each part cannot specify its past state even if its present state is known, thus $I^*=0$, and $\phi=2$. In other words, ϕ quantifies how much information is lost if the whole system is cut into its constituent parts.

Importantly, ϕ can be exhaustively computed for any subset in the system. For a system of three neurons, A, B, and C (Figure 1d), ϕ is defined for all subsets, including AB, AC, BC as well as ABC. Once we exhaustively compute integrated information for all subsets, there is a hierarchical pattern of integrated information. Say, AB and ABC are high, BC is low, and AC is 0. This nested and compositional structure of integrated information is postulated to correspond to compositionality of experience. When we experience a face, it is composed of experience of parts, such as eyes, nose and mouth. An experience of a face is also a subset of larger experience of vision, composed of other objects and background. Visual experience is also embedded in an experience composed of all sensory modalities.

[Figure 2 around here]

In addition to the basics described above, there are two concepts that are crucial to understand how IIT treats non-conscious processing: the Minimum Information Partition (MIP) and the exclusion principle.

The MIP can be considered as the most appropriate way to cut the system when one tries to compute I* in the step depicted in Figure 1c. In Figure 2, we consider 2 pairs of 2 neurons, where there is an interaction within each pair, but not at all between the pairs. If we compute ϕ of the entire system with a cut between the left and the right pair, ϕ for the entire system is correctly identified as 0 (Figure 2a). (This cut "minimizes" information between the cut parts, thus it is called the MIP.) But if we cut it through the interacting pairs between the upper and the lower half (Figure 2b), ϕ is overestimated as non-zero.

The exclusion principle relates how to find the most critical subset of the system. According to the exclusion principle, which IIT postulated based on the exclusive property of phenomenology, the subset that has the largest ϕ, which is called "complex" in IIT, only matters for consciousness. In Figure 2c, we consider an example of 3 strongly interacting neurons ABC and additional neuron D. In this case, the maximal interactions can be identified within ABC. Any cut introduced to ABC always reduces integrated information. Further, adding D to ABC will introduce a very weak link to the system. In this case, the cut between ABC and D will make ϕ_{ABCD} to be nearly 0. Any neural interaction outside of the complex corresponds to non-conscious processing. Here, IIT predicts that interaction between C and D is not experienced by the complex, ABC.

Important for the discussion in this paper is IIT's explanation on how uniqueness of each sensory modality arises. According to IIT, the uniqueness arises from the way each mechanism in the complex causally interacts with others, constructing a specific pattern of integrated information. For example, the "visualness" of visual experience is determined not only by the way visual neurons interact with other visual neurons, but it also depends on how the visual neurons interact with auditory neurons and other neurons within the complex (Figure 2d). Likewise, within visual quality, patterns of integrated information for color should have unique properties, which distinguish them from patterns of integrated information for shape. Relationships between these patterns define quality of color and shape. In other words, the meaning of neural interactions, or quality of experience for which they are responsible, can be determined only by the interactions with other neural interactions in a holistic manner.

This intuitive summary of IIT will be our guide for the rest of the paper. According to these principles, IIT explains the known neural basis of consciousness and makes further predictions. The more variable interactions a system can have, the richer conscious phenomenology it can entertain. Not all interactions matters, as any interactions that are outside of the complex have no effects on the complex, leading to non-conscious processing. It is the connectivity and the activation patterns that eventually determine exactly what types of conscious experience a system has at each moment. The theory, in principle, can get us closer to approaching bats' experience.

4. A framework for empirical testing of IIT towards understanding bats' consciousness

4.1 Computing integrated information patterns from neural activity

Applying these IIT concepts as they are to a real human brain, which is composed of 10^{11} neurons and 10^{14} synaptic connections, is currently impossible for practical purposes. Thus, we need some gross approximations for these concepts when we empirically test explanations and predictions from IIT (Barrett & Seth, 2011; Chang et al., 2012; Lee, Mashour, Kim, Noh, & Choi, 2009; Oizumi, Amari, et al., 2016; Oizumi, Tsuchiya, & Amari, 2016; Tegmark, 2016). With approximations, our research group has computed patterns of integrated information from real neural activities recorded in awake human patients while they reported what they see in each trial in several tasks (Haun et al., 2016) (Figure 3). The result is consistent with an idea and prediction from IIT, which is that patterns of particular types of neural interactions determine quality of a particular aspect of experience. While this research program is still at an early stage, we can now compute patterns of integrated information based on neural recordings and test if such patterns correspond to what subjects experience.

[Figure 3 around here]

4.2 No-report paradigms to understand consciousness in non-speaking animals

Contents of consciousness at perceptual thresholds would require us to test if patterns of integrated information correspond to perceptual reports in a trial-by-trial manner (Haun et al., 2016). However, the act of perceptual reports may activate various brain areas that are neither necessary nor sufficient for conscious experience per se. Recently developed "no-report" paradigms remove strict requirements of perceptual reports from subjects by manipulation of their conscious experience through instructions/expectations or by reliable inference of conscious contents through bodily signals, such as eye movements (Tsuchiya, Wilke, Frässle, & Lamme, 2015). No-report paradigms have implied that certain parts of the brain areas, such as the prefrontal areas, may not be related to consciousness, but more to do with the act of the reports (Koch, Massimini, Boly, & Tononi, 2016).

No-report paradigms are especially powerful to infer the nature of experience in animals, as they remove difficulties associated with training animals to reliably reports their percepts. No-report paradigms for simple perceptual discriminations, such as discriminations of visual and auditory stimuli would be feasible to develop for humans and various animals, especially without any perceptual masking. Once we establish no-report paradigms and record neuronal activities, we can then compare the structure of conscious experience and the patterns of integrated information across various sensory modalities and animal species, which brings us closer to bat consciousness. A remaining difficulty is comparing structures of consciousness and patterns of integrated information. Such comparisons can be formally achieved by a mathematical formalism, called category theory.

4.3 Category theory to link consciousness and patterns of integrated information across different modalities and animal species

Towards empirical studies of bat consciousness, we have to examine what types of relationships exist among completely distinct domains. We need to compare visual and auditory consciousness, consciousness for humans and bats, and crucially, the domain of conscious experience and the domain of mathematics (integrated information). Mathematical formalism, called category theory (Awodey, 2010; Mac Lane, 1998) is a powerful tool to achieve such a goal (Tsuchiya, Taguchi, & Saigo, 2016).

Category theory can be thought of as a more flexible version of set theory. It can precisely characterize relationships between two completely different domains of knowledge to the extent that what types of mathematical conclusions can be transferred from one domain to the other. Unlike set theory, category theory is developed to characterize the nature of "relationships" between objects (Tsuchiya et al., 2016). The category theory's focus on relationships rather than objects is very well suited for its application to the problems of consciousness as well as IIT, as the "relationships" are critical for both, as outlined above.

So far, category theory has been applied mainly in mathematics and physics. For example, by establishing a certain similarity between geometry and algebra, a very difficult theorem in geometry can be easily solved in algebra, which can be used as a proof of the theorem in geometry. Also, quantum mechanics, logic, and computation can be formally shown to be similar in some sense (Baez & Stay, 2009), which allows proofs in one of these domains to be directly applied to the problems in the others. Importantly, category theory offers precise definitions about "similarity" in different degrees (e.g., a very strong similarity of "isomorphism" is weaker than "identity", (Tsuchiya et al., 2016)). Different levels of transfer of knowledge between the categories can be achieved at different levels of similarity between categories.

For our purpose, we need to formally compare across categories of consciousness, which varies in modalities (e.g., vision, audition), animal species (e.g., humans, bats), and categories of mathematical structures, such as integrated information patterns. Being mathematical objects, integrated information patterns as a category should be relatively easy to deal with in category theory. To characterize categories of consciousness, some framework in mathematical psychophysics (Hoffman, 1966) combined with no-report paradigms in animals will be useful.

Once these domains are characterized as categories, we can investigate the nature of the relationships among these categories (Figure 4). How are visual and auditory consciousness different and similar to each other? Under no-report paradigms, is visual experience in humans comparable with those in monkeys (Crick & Koch, 2003; Leopold et al., 2003), rats, and bats? What about auditory experience? We can ask the corresponding questions with respect to patterns of integrated information across animals and modalities. Of course, it is critical to ask if the domain of consciousness and integrated information patterns correspond at each level. If not, it implies the mathematical structures proposed by IIT are wrong, a potentially powerful way to reject IIT in the currently proposed format.

Now, coming back to bats, what happens if we apply the same IIT analyses to bats' brain? If patterns of integrated information from their echolocation area are more similar to those generated in the visual than the auditory cortex, then, the theory predicts that experience of echolocation should be similar to those of visual experience. If they are closer to those originating from the auditory areas, quality of echolocation is closer to sense of sound. If bats are not really experiencing anything with echolocation, much like non-conscious processing, IIT would predict that patterns from the echolocation area is very low in magnitude without much variety, residing mainly outside of the complex. Perhaps, processing modules for echolocation may be parallel and independent, like those of our non-conscious cerebellar system. The outline above is a potential pathway to understand bat consciousness.

5. Concluding remarks

Some explanations and predictions from IIT are still not yet developed and most of them have not been directly empirically tested. Some of them are even untestable. However, similar to the age of the universe or the evolutionary theory, the theory can be grounded by available evidence and make progress. There are growing interests in empirically testing the theory and the tools that enable testing are being developed. Rejecting IIT as non-testable theory would be premature.

If IIT is validated, it will have significant philosophical implications. IIT is unlikely to be easily categorized as one of the traditional options in philosophy, be it physicalism, dualism, panpsychism or others. IIT starts from the phenomenology, acknowledging that one cannot doubt one's own ongoing conscious experience. But its essence is to try to find physical substrates of consciousness. Note that essential relationships in IIT are those between consciousness and mathematical structures derived from the physical substrates, not between consciousness and matter as is usually debated in philosophy. This means that two distinct physical substrates can generate identical consciousness. Also, IIT does not assume everything is conscious (Tononi & Koch, 2015), which is a direct consequence of the exclusion principle, which says that only the local maxima of integrated information is relevant for consciousness. In other words, if a neuron (or a fundamental particle, or whatever) participates in my current consciousness, it cannot participate in any consciousness at smaller or bigger scales. This seems a feature that is present in most versions of panpsychism (Skrbina, 2003)[3]. It would be an interesting project in philosophy to clarify theoretical issues surrounding IIT

[3] This exclusion principle solves the "combination problem" in panpsychism.

and how it fits (or not) with the traditional classifications and options available in philosophy of mind.

An approach outlined in this article is neuroscientific and empirical, allowing us to attack the problem of bat consciousness. Especially with ever-advancing techniques in identifying anatomical connections and recording from many neurons simultaneously as well as in manipulating the connections and states of neurons, this line of research will be highly possible and fruitful, especially when combined with more sophisticated computational analyses (Tononi et al., 2016; Tsuchiya, Haun, Cohen, & Oizumi, 2017). Starting with one's own phenomenology, the theory tries to come up with a mathematical framework, which explains the quality of consciousness based on neural connectivity and activity. The theory would start explaining one's own phenomenology, but should be gradually extended and confirmed to other humans who can report. Then, to animals who are trained to report with careful manipulations (Kepecs, Uchida, Zariwala, & Mainen, 2008; Kiani & Shadlen, 2009; Leopold et al., 2003), and through to humans in no-report paradigms including people without report capability (e.g., babies, injured subjects) (Tsuchiya et al., 2015). Across various modalities and animals, we need to verify if the structure of consciousness corresponds to that of the proposed mathematical structure, such as integrated information patterns. Category theory (Awodey, 2010; Mac Lane, 1998) is a powerful mathematical tool to bridge these two distinct domains of knowledge (Tsuchiya et al., 2016). If IIT makes a highly counter-intuitive prediction, yet empirical tests confirm it, IIT will gain the credibility. As the credibility of IIT gradually builds up, we can gradually increase our trust in the theory to infer conscious experience in animals, eventually in bats.

Although it may be practically impossible to understand bats' phenomenology in every detail, the research project I outlined above would be sufficient to give a highly credible answer as to whether the bat's echolocation is closer to audition, vision or non-conscious processing. Identifying the neuronal connectivity in bats' brain and understanding their neural activation patterns, analysed according to the IIT's principles will give a fairly educated and grounded answer, assuming that IIT is correct.

Still, such an answer may be too far from the certainty that we would like to achieve eventually. At the moment, however, the precision of any guess on quality of consciousness in other species is very bad and it has no possibility of generalization across species in a quantitative way. If IIT is validated enough to the extent that we can believe, for example, bats' echolocation should feel like vision, but not audition, that would be a tremendous breakthrough in consciousness research! Consider the age of the universe. 1000 years ago, we had no idea about the age of the universe. With current precision cosmology, however, the estimate of 13.7 billion years is believed to be with an estimated error of 1%. Actually, how many of us know the age of grandmothers or friends with 1% error?

Predicting the sensory experience based on a mathematical framework, be it IIT or anything else, might become possible and important soon in the future. Artificial neural circuits for repairing damaged brain areas are being developed. If we can attach/detach such circuits, we can test the prediction about how our sensory experience changes as we attach and detach the device. As restoring sensory deficits in patients due to disease or brain trauma is an important medical issue, there will be needs and potentials for such technology. Eventually, we may be able to develop an artificial "bat" circuit, which will allow us to directly experience "what it is like to be bats"!

Acknowledgement

The author's research was supported by Australian Research Council Future Fellowship (FT120100619) and Discovery Project (DP130100194). The author thanks Andrew Haun, Jennifer Windt, and Michael Dahlquist for their comments on the earlier version of this draft.

References

Awodey, S. (2010). *Category Theory*: Oxford University Press.

Bach-y-Rita, P., & S, W. K. (2003). Sensory substitution and the human-machine interface. *Trends Cogn Sci, 7*(12), 541-546.

Baez, J. C., & Stay, M. (2009). Physics, Topology, Logic and Computation: A Rosetta Stone. *Arxiv*.

Balduzzi, D., & Tononi, G. (2008). Integrated information in discrete dynamical systems: motivation and theoretical framework. *PLoS Comput Biol, 4*(6), e1000091. doi:10.1371/journal.pcbi.1000091

Balduzzi, D., & Tononi, G. (2009). Qualia: the geometry of integrated information. *PLoS Comput Biol, 5*(8), e1000462. doi:10.1371/journal.pcbi.1000462

Barrett, A. B., & Seth, A. K. (2011). Practical measures of integrated information for time-series data. *PLoS computational biology, 7*(1), e1001052. doi:10.1371/journal.pcbi.1001052

Bogen, J. E. (1995). On the neurophysiology of consciousness: I. An overview. *Conscious Cogn, 4*(1), 52-62.

Chalmers, D. J. (1996). *The conscious mind* (白楊社, Trans.). New York: Oxford University Press.

Chang, J. Y., Pigorini, A., Massimini, M., Tononi, G., Nobili, L., & Van Veen, B. D. (2012). Multivariate autoregressive models with exogenous inputs for intracerebral responses to direct electrical stimulation of the human brain. *Frontiers in human neuroscience, 6*, 317. doi:10.3389/fnhum.2012.00317

Crick, F., & Koch, C. (2003). A framework for consciousness. *Nat Neurosci, 6*(2), 119-126.

Dang-Vu, T. T., Schabus, M., Desseilles, M., Albouy, G., Boly, M., Darsaud, A., . . . Maquet, P. (2008). Spontaneous neural activity during human slow wave sleep. *Proc Natl Acad Sci U S A, 105*(39), 15160-15165.

doi:10.1073/pnas.0801819105

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

Engel, & Singer, W. (2001). Temporal binding and the neural correlates of sensory awareness. *Trends in Cognitive Sciences, 5*(1), 16-25.

Haun, A., Kawasaki, H., Kovach, C., Oya, H., Howard, M. A., Adolphs, R., & Tsuchiya, N. (2016). Contents of Consciousness Investigated as Integrated Information in Direct Human Brain Recordings. *bioRxiv*, http://dx.doi.org/10.1101/039032.

Herculano-Houzel, S. (2012). The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. *Proc Natl Acad Sci U S A, 109 Suppl 1*, 10661-10668. doi:10.1073/pnas.1201895109

Hoel, E. P., Albantakis, L., & Tononi, G. (2013). Quantifying causal emergence shows that macro can beat micro. *Proc Natl Acad Sci U S A, 110*(49), 19790-19795. doi:10.1073/pnas.1314922110

Hoffman, W. C. (1966). The Lie algebra of visual perception. *Journal of mathematical Psychology, 3*(1), 65-98.

Jones, G. (2005). Echolocation. *Curr Biol, 15*(13), R484-488. doi:10.1016/j.cub.2005.06.051

Kanai, R., & Tsuchiya, N. (2012). Qualia. *Current biology : CB, 22*(10), R392-396. doi:10.1016/j.cub.2012.03.033

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature, 455*(7210), 227-231. doi:10.1038/nature07200

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science, 324*(5928), 759-764. doi:10.1126/science.1169405

Koch, C. (2004). *The Quest for Consciousness: A Neurobiological Approach* (岩波書店, Trans.). CO: Roberts and Publishers.

Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: progress and problems. *Nat Rev Neurosci, 17*(5), 307-321. doi:10.1038/nrn.2016.22

Lamme. (2010). How neuroscience will change our view on consciousness. *Cognitive NeuroScience*, 1-57.

Laureys, S. (2005). The neural correlate of (un)awareness: lessons from the vegetative state. *Trends Cogn Sci, 9*(12), 556-559.

Lee, U., Mashour, G. A., Kim, S., Noh, G. J., & Choi, B. M. (2009). Propofol induction reduces the capacity for neural information integration: implications for the mechanism of consciousness and general anesthesia. *Consciousness and Cognition, 18*(1), 56-64. doi:10.1016/j.concog.2008.10.005

Lemon, R. N., & Edgley, S. A. (2010). Life without a cerebellum. *Brain, 133*(Pt 3), 652-654. doi:10.1093/brain/awq030

Leopold, D. A., Maier, A., & Logothetis, N. (2003). Measuring subjective visual perception in the nonhuman primate. *Journal of Consciousness Studies, 10*(9-10), 115-130.

Mac Lane, S. (1998). *Categories for the Working Mathematician*: Springer.

Massimini, M., & Tononi, G. (2015). *Nulla Di Piu Grande* (亜紀書房, Trans.).

Nagel, T. (1974). What Is It Like to Be a Bat? *The Philosophical Reveiw, 83*(4), 435-450.

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. *PLoS Comput Biol, 10*(5), e1003588. doi:10.1371/journal.pcbi.1003588

Oizumi, M., Amari, S., Yanagawa, T., Fujii, N., & Tsuchiya, N. (2016). Measuring Integrated Information from the Decoding Perspective. *PLoS Comput Biol, 12*(1), e1004654. doi:10.1371/journal.pcbi.1004654

Oizumi, M., Tsuchiya, N., & Amari, S. (2016). A unified framework for information integration based on information geometry. *PNAS*.

Parvizi, J., Jacques, C., Foster, B. L., Withoft, N., Rangarajan, V., Weiner, K. S., & Grill-Spector, K. (2012). Electrical stimulation of human fusiform face-selective regions distorts face perception. *The Journal of neuroscience : the official journal of the Society for Neuroscience, 32*(43), 14915-14920. doi:10.1523/JNEUROSCI.2609-12.2012

Ramachandran, V., & Hubbard, E. (2001). Synaesthesia--a window into perception, thought and language. *Journal of Consciousness Studies, 8*(12), 3-34.

Ramachandran, V. S., & Blakeslee, S. (1998). *Phantoms in the Brain: Probing the Mysteries of the Human Mind*. New York: William Morrow.

Schabus, M., Dang-Vu, T. T., Albouy, G., Balteau, E., Boly, M., Carrier, J., . . . Maquet, P. (2007). Hemodynamic cerebral correlates of sleep spindles during human non-rapid eye movement sleep. *Proc Natl Acad Sci U S A, 104*(32), 13164-13169. doi:10.1073/pnas.0703084104

Skrbina, D. (2003). Panpsychism as an underlying theme in western philosophy A survey paper. *Journal of Consciousness Studies, 10*(3), 4-46.

Tegmark, M. (2016). Improved Measures of Integrated Information. *Arxiv*, http://arxiv.org/abs/1601.02626v02621.

Tong, F., Nakayama, K., Vaughan, J. T., & Kanwisher, N. (1998). Binocular rivalry and visual awareness in human extrastriate cortex. *Neuron, 21*(4), 753-759.

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neurosci, 5*, 42.

Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. *Biol Bull, 215*(3), 216-242. doi:215/3/216 [pii]

Tononi, G. (2012). Integrated information theory of consciousness: an updated account. *Archives italiennes de biologie, 150*(2-3), 56-90. doi:10.4449/aib.v149i5.1388

Tononi, G. (2015). Integrated information theory. *Scholarpedia, 10*(1), 4164.

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nat Rev Neurosci, 17*(7), 450-461. doi:10.1038/nrn.2016.44

Tononi, G., & Koch, C. (2015). Consciousness: here, there and everywhere? *Philos Trans R Soc Lond B Biol Sci, 370*(1668). doi:10.1098/rstb.2014.0167

Tsuchiya, N., Haun, A., Cohen, D., & Oizumi, M. (2017). Empirical tests of integrated information theory of consciousness. In A. Haag (Ed.), *Return of Consciousness*. Sweden: Axon Foundation.

Tsuchiya, N., Taguchi, S., & Saigo, H. (2016). Using category theory to assess the relationship between consciousness and integrated information theory. *Neuroscience research*. doi:http://dx.doi.org/10.1016/j.neures.2015.12.007

Tsuchiya, N., Wilke, M., Frässle, S., & Lamme, V. A. (2015). No-Report Paradigms: Extracting the True Neural Correlates of Consciousness. *Trends in Cognitive Sciences, 19*(12), 757-770.

Wilke, M., Logothetis, N. K., & Leopold, D. A. (2006). Local field potential reflects perceptual suppression in monkey visual cortex. *Proc Natl Acad Sci U S A, 103*(46), 17507-17512. doi:0604673103 [pii]

10.1073/pnas.0604673103

Yu, F., Jiang, Q. J., Sun, X. Y., & Zhang, R. W. (2015). A new case of complete primary cerebellar agenesis: clinical and imaging findings in a living patient. *Brain, 138*(Pt 6), e353. doi:10.1093/brain/awu239

Figure Legend

Figure 1. IIT in a nutshell (part 1). a) A system composed of 2 neurons. A state of each neuron is either on (white) or off (black). Each neuron copies the state of the other with a time delay of τ (tau). As a system, it has 4 possible states, 1) both on, 2) and 3) one on and the other off, and 4) both off. Uncertainty of the state is called entropy. We can quantify entropy (H) as log_2(the number of possible states) = $log_2(4)$ = 2, in this case. b) If the present state (at time=t) of the two-neuron is "on-off" (i.e., X(t)="state 3 in a"), then, its past state at t-τ had to be "off-on" (i.e., X(t-tau)="state 2 in a"), because each neuron copies the state of the other. Thus, knowing the present state removes all uncertainty about the past state. The reduction of uncertainty is quantified as mutual information: $I=H-H^*$=2-0=2. (H* quantifies the reduced uncertainty; $H^*=log_2$(the number of possible states given the present state)=$log_2(1)$=0 in this case.) c) If each neuron is considered separately, each cannot specify its own past state. A sum of the mutual information among separated parts is I*. In this case, I^*=0. How much information is lost when the system is cut is integrated information: $\phi=I-I^*$=2-0=2. Note that all ingredients for integrated information, from H, H*, I, I* and ϕ is a function of both the connectivity and the state of the system. d) Integrated information, ϕ, can be considered for any subset of the system. For example, ϕ_{AB}, ϕ_{BC}, ϕ_{CA}, and ϕ_{ABC}, represent integrated information between A and B, B and C, and C and A, as well as among A, B and C.

Figure 2. IIT in a nutshell (part 2). Non-conscious perception and uniqueness of each sensory modality. a-c) Key concepts to understand how IIT treats non-conscious processing are the Minimum Information Partition (MIP) and the exclusion principle. a) Two independent systems, as the case of two sets of two neurons depicted here, should have no integrated information because there is no interaction between the two sets. Each subset is identical to the example in Figure 1, with ϕ=2. When two non-interacting subsets are considered with the appropriate cut

(MIP), there is no loss of information across the cuts (I = I*, and ϕ=0). b) With an inappropriate cut, there is loss of information (I* decreases), and integrated information is overestimated (ϕ>0). Thus, it is critical to estimate the MIP accurately. c) Another example of a 4-neuron system. If AB and AC are strongly interacting, ϕ_{ABC} will be above 0. If D just provides weak input to C, the MIP among ABCD is identified as ABC vs. D, correctly identifying ϕ_{ABCD} to be nearly 0. IIT claims that the subset within a system that achieves the maximum ϕ only matters for conscious experience of the system, and everything else is non-conscious (the exclusion principle). The local maximum subset is called a "complex" in IIT. In this example, the complex is ABC. A pattern of integrated information within ABC (i.e., ϕ_{AB}, ϕ_{BC}, ϕ_{CA}, and ϕ_{ABC}) determines quality of experience of ABC. Any integrated information outside of the complex (e.g., ϕ_{CD}) corresponds to non-conscious processing. d) Interactions among neurons in the complex determine the quality of experience in each modality. Peculiar quality of experience in each modality (e.g., visualness) is determined by patterns of integrated information within the neurons for that modality as well as those across modalities in a holistic manner. In other words, vision cannot feel like vision unless it is related with other senses.

Figure 3. An example of patterns of approximated integrated information, ϕ* (Oizumi, Amari, et al., 2016), from the actual neural recordings (Haun et al., 2016). a) We recorded intracranial neural activity in the fusiform gyrus, which is strongly suspected to generate conscious percept of a face (Parvizi et al., 2012; Tong, Nakayama, Vaughan, & Kanwisher, 1998). Recording was performed in awake subjects under epilepsy monitoring. Subjects performed several tasks and saw various stimuli under conscious and non-conscious conditions. 4 traces show the evoked neural activity when the patient consciously saw a face from channel A, B, C, and D. b) From the recordings, we computed necessary ingredient of integrated information (ϕ*), which is entropy (H) and mutual information (I), for all subsets of A,

B, C and D, over time. Based on these, we computed ϕ* for each subset and each time. c) A magnitude of ϕ* over 11 subsets is represented as a shape at 400 ms after the stimulus onset. The height of each dot is the magnitude of ϕ* for each subsystem. For details see (Haun et al., 2016). d) Based on patterns of integrated information, we were able to infer what subjects consciously saw in each trial at a high precision, when the electrodes were implanted in the object sensitive area. Here, the dendrogram demonstrates that the pattern of integrated information was closely related to the image that the subject saw on each trial.

Figure 4. Schematic of how we address the question of bat's consciousness, combining IIT with category theory. Category theory allows us to compare distinctive domains of knowledge, such as structures of conscious experience and patterns of integrated information. If any change in experience changes integrated information pattern and vice versa, a strong relationship of "isomorphism" can be established between them, as IIT proposes (Oizumi et al., 2014). Other than isomorphic relation, category theory offers varying degrees of "similarity" (Tsuchiya et al., 2016). IIT needs to be validated to establish isomorphism between consciousness and mathematical structure across various animals. When that is achieved, similarity of integrated information patterns in bat's echolocation with those for vision, audition or non-conscious processing will be decisive as to the nature of bat consciousness.

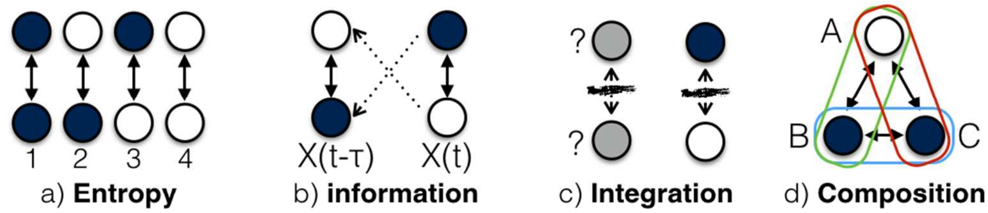

Figure 1

287x110mm (96 x 96 DPI)

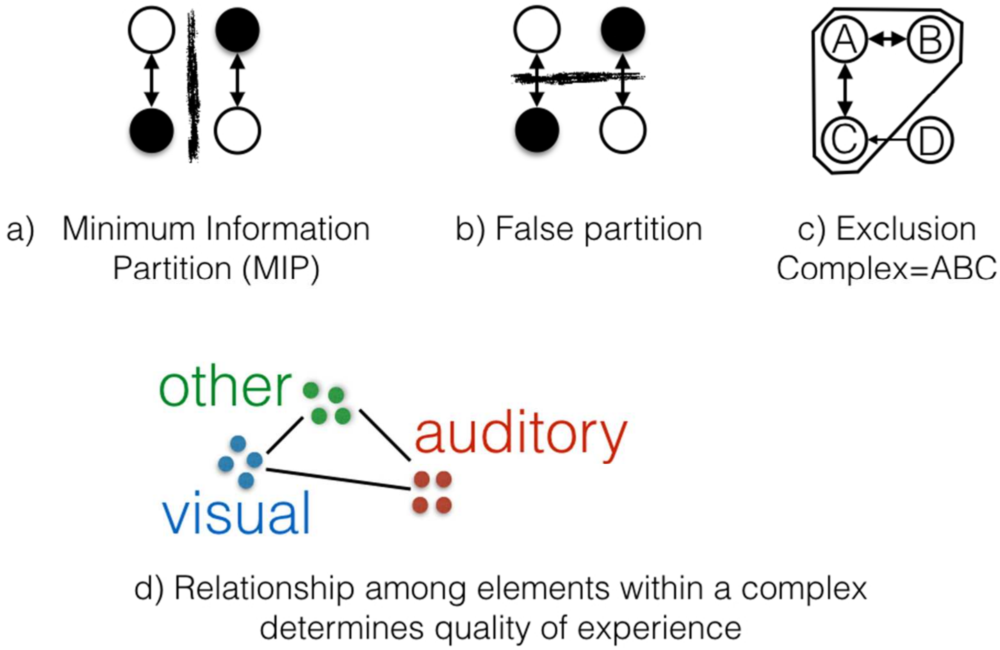

Figure 2

240x155mm (96 x 96 DPI)

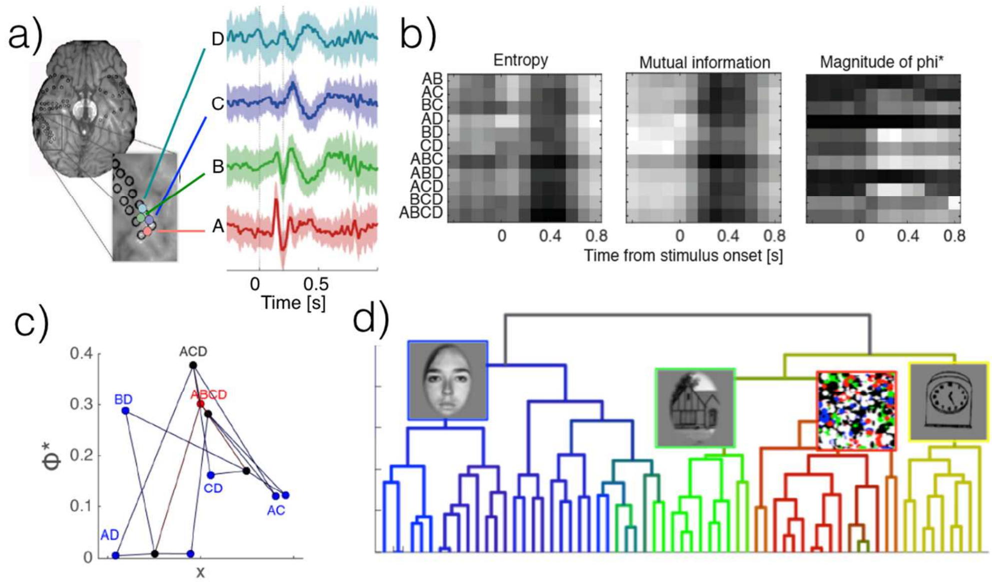

Figure 3

342x200mm (96 x 96 DPI)

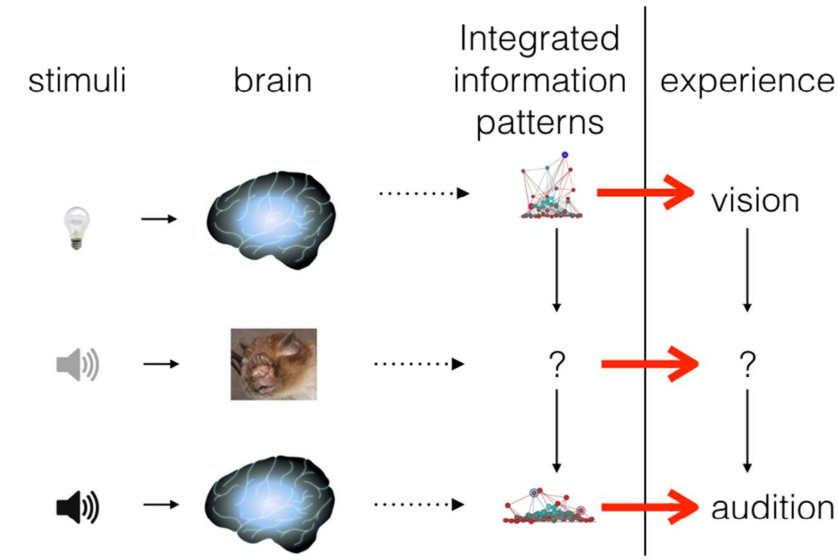

Figure 4. Schematic of how we address the question of bat's consciousness, combining IIT with category theory. Category theory allows us to compare distinctive domains of knowledge, such as structures of conscious experience and patterns of integrated information. If any change in experience changes integrated information pattern and vice versa, a strong relationship of "isomorphism" can be established between them, as IIT proposes (Oizumi et al., 2014). Other than isomorphic relation, category theory offers varying degrees of "similarity" (Tsuchiya et al., 2016). IIT needs to be validated to establish isomorphism between consciousness and mathematical structure across various animals. When that is achieved, similarity of integrated information patterns in bat's echolocation with those for vision, audition or non-conscious processing will be decisive as to the nature of bat's consciousness.

361x270mm (72 x 72 DPI)