A Critique of Some Assumptions Underlying Scientific Theories of Consciousness, Exemplified Through a Discussion of the Integrated Information Theory of Consciousness

Martin Larsson

Master Thesis in Psychology at the Department of Psychology

University of Oslo

November, 2011

A Critique of Some Assumptions Underlying Scientific Theories of Consciousness, Exemplified Through a Discussion of the Integrated Information Theory of Consciousness

Martin Larsson

2011
A Critique of Some Assumptions Underlying Scientific Theories of Consciousness,
Exemplified Through a Discussion of the Integrated Information Theory of Consciousness
Martin Larsson
http://www.duo.uio.no/
Print: Reprosentralen, University of Oslo

Abstract

The Integrated Information Theory of Consciousness (IITC) aims to give a set of formalized and exact mathematical formulas to compute the level of consciousness that a certain system state generates, with the level of consciousness corresponding to the richness of the information contained in the conscious experience. In this thesis, it is argued that some of the underlying assumptions that the IITC extrapolates from to infer the exact nature of these formulas are unjustified. These assumptions asserts that most system states generates either no or an insignificant amount of consciousness. For example, it is assumed that the human brain, during dreamless sleep, generates no significant amount of consciousness. In this thesis, it is argued that there are no empirical, theoretical or probabilistic reasons for actually believing that these assumptions are valid. Rather, a more agnostic view regarding the amount of consciousness that is generated by these system states is championed. It is also argued that *if* there is some significant amount of consciousness present during some, or all, of these system states, it should be primitive in nature. The amount of consciousness itself for these system states is, however, impossible to approximate. It is then argued that the critique presented is applicable to any scientific theory of consciousness that works with roughly the same basic assumptions as the IITC, something which it is argued that most of these theories do. Finally, some merits of current scientific theories of consciousness, along with a plea aimed at the researchers within the field, as well as suggestions for future research, is presented.

Acknowledgements

This thesis has been independently chosen and developed by me (this is the author speaking, by the way). It has been typeseted with LaTeX (http://www.latex-project.org/) and the illustrations have been generated with the TikZ, PGF and gnuplot packages.

Thanks to Tobias Malm and Tom Everitt for long and interesting discussions about the subject matter over the years. Without their brains (iff no other brains of equal intelligence and with the same enthusiasm for the subject would have communicated with me), my brain would not have been able to write this thesis. Also thanks for their constant support during the writing process itself. Thanks to Frode Kristian Hansen at the Institute of Theoretical Astrophysics at the University of Oslo who, very willingly, discussed the intersection between consciousness studies and cosmology with me (even though not much of it ended up in the final thesis). Thanks to Victoria Forsberg for proofreading the whole shabang at a time when I probably would not even have been able to spot profanities interspersed between the mathematical formulas. Thanks to the people at english.stackexchange.com who collaborated with me to come up with a good title. Finally, thanks to everybody at tex.stackexchange.com who, with their answers to my mountain of questions, made it possible for me to actually write the thesis using LaTeX. It was a pain in the behind to get everything to work as it should, but in the end, it was worth it.

Contents

Consciousness has been discussed within philosophy for several millennia. However, it is only relatively recently that it also has become an object of study within science. After many years of empirical research on this topic, scientific theories of consciousness, coherent frameworks that sets out to tie all the knots together and present systemized explanations to how consciousness arises and what it does, started to appear. Two such theories are, for example, Baars' (2002) Global Workspace Theory and Lamme's (2003) theory about recurrent processing and its role in the generation of consciousness.

The goal of this thesis is to evaluate if these types of theories, in their current form, have anything to contribute towards coming up with an answer to the age old question of consciousness. However, to have a focused discussion, only one instance of such a scientific theory of consciousness will be looked at, namely the Integrated Information Theory of Consciousness (IITC) (Balduzzi & Tononi, 2008, 2009).

The IITC was chosen because of several reasons. First, papers about the IITC, or less developed proto-versions of it, has been published in several different journals with highly scientific, rather than philosophic, profiles such as *PLoS Computational Biology* (Balduzzi & Tononi, 2008, 2009), *BMC Neuroscience* (Tononi, 2004; Tononi & Sporns, 2003) and *Science* (Tononi & Edelman, 1998). This means that one should be able to define the IITC as a *scientific* theory of consciousness, rather than just one among many philosophical. Second, it was deemed to be a good representative for most current, popular, scientific theories of consciousness with regard to the underlying philosophical framework it is built upon. This means that an evaluation of the IITC to a large degree also could be applicable to other scientific theories of consciousness, thereby saying something general about them all. Third, the IITC is highly mathematically explicit and is applicable to very low-level systems. This has the consequence that the IITC can, much easier than other theories, be evaluated without it being able to escape into elusive and vague territories. It can, for example, deal with counter-arguments like the small network argument (Herzog, Esfeld, & Gerstner, 2007) in a proficient way, where other theories would have no clear answer.

This thesis is divided into three main sections. In the first section, called "Review of the IITC", the IITC will be presented. This will be a description of the theory, with the only original contribution comming from some of the analogies and the extended explanations of the inherent concepts.[1] In the second section, called "Examination of the IITC", the IITC will be examined and subsequently critiqued. This critique will center around the, for the theory, foundational assumptions of which system states that lacks consciousness, as well as how the IITC manages to defend

[1] It should be noted that this first part of this thesis, compared to the other parts, is quite heavy on the mathematics. This might, at first glance, seem unnecessary, given that the following sections does not address the mathematical formulas directly, rather attacking the fundament the mathematical formulas in turn are built upon. However, it is important to give the IITC a fair treatment, describing the theory in its full form to make sure that no hidden solutions to the critique against it presented in this thesis lurks around somewhere within the mathematical formulas.

itself against critique of these. In this section, it will be argued that these underlying assumptions of the IITC does indeed *not* hold up, something which have important consequences for the theory's ability to explain consciousness. In the last section, called, "Scientific Theories of Consciousness", the examination and critique of the IITC will be related back to scientific theories of consciousness in general. In this section, it will be argued that most scientific theories of consciousness are vulnerable to the same type of critique which is given in the preceding examination section of the IITC.

In this thesis, no specific philosophical position of consciousness will be either favored or discounted from the get-go. Consequently, any affirmation or refutation of any such position will be explicitly argued for in the text.

Review of the IITC

The IITC aims to explains two aspects of consciousness: (a) what level of consciousness a certain system in a certain state has (Balduzzi & Tononi, 2008) and (b) what kind of experiences a certain system in a certain state has (Balduzzi & Tononi, 2009). The first aspect, that is, the quantity of consciousness, is captured by the value of the variable ϕ, representing the amount of integrated information in a system. The second aspect, that is, the quality of consciousness, is captured by the shape of a polytope that is formed within an abstact multi-dimensional qualia space, yielded by the configuration and current state of a specific system.

Background and motivation

The IITC has sprung out of a certain tradition within the field of scientific theories of consciousness where consciousness is seen as something that arises when the processing of some piece of information is processed on a global scale, recruiting many parts of the brain, rather than only being localized to certain specific areas. One of the most prominent of these theories is Baars' (1983, 2002) global workspace theory. In essence, this theory can really be seen as a less formalized framework that the IITC in turn builds upon to arrive at more specific predictions. Other people who have adopted some kind global processing view on consciousness are Dennett (2001) and Damasio (1989).

Two key aspects of consciousness that is highlighted within the scope of the IITC is that it is highly *integrated* at the same time that it is highly *differentiated* (Tononi & Edelman, 1998). That it is highly integrated means that conscious experience forms a coherent whole and that it cannot be broken down into individual constituents without the loss of meaning. As an example of this, if the number 1 and the number 7 is briefly presented adjacent to each other, what is seen is the number 17, which conceptually is not decomposable to 1 and 7 (Edelman & Tononi, 2000, p. 24). A result of this unity is for example that it is impossible to experience two incongruent scenes at

the same time, which is demonstrated when it comes to ambiguous figures (Sengpiel, 1997) and perceptual rivalry (Tononi, McIntosh, Russell, & Edelman, 1998; Srinivasan, Russell, Edelman, & Tononi, 1999).

That consciousness is highly differentiated means that for every conscious state, there is an almost infinite amount of alternative experiences that are not instantiated. In this sense, when a conscious state realizes one certain possible outcome it also differentiates itself from all the other possibilities. This makes the state highly informative in the respect that it reduces uncertainty (see the discussion about entropy on the following page).

States of no consciousness, such as for example dreamless sleep, has been associated with depression of neural activity (Silber et al., 2007) and reduced blood-flow (Braun et al., 1997), compared to highly conscious states such as waking or REM-sleep. However, these measures do not say much about the more specific cooperation between different parts of the brain as stimuli becomes globally accessible. One example of how this is investigated is when it comes to experiments of binocular rivalry. In these experiments, two different pictures are used where each one then is fed to a separate eye of the subject. This results in that the subject only reports seeing one of the pictures at any given time, although which picture is dominating spontaneously switches back and forth between the two. In a series of such experiment (Srinivasan et al., 1999; Tononi et al., 1998), a method called "frequency tagging" was used. Here, each stimuli flickered with a unique frequency between 7-12 Hz making it possible to track its subsequent ramifications in the brain. With the help of MEG measurements, it was then showed that when a stimuli was reported, compared to when it was not, the activity of that stimuli in the brain was stronger, more distributed and created a stronger coherence between distant brain regions.

In a large scale simulation of a part of the thalamocortical system used for processing visual stimuli (Tononi, Sporns, & Edelman, 1992), it could be shown that discrimination and selection between multi-feature objects (built up by color, shape and movement) was successful when there also was a high amount of integration in the system as a whole. From this and other simulations (Lumer, Edelman, & Tononi, 1997a, 1997b), Tononi and Edelman (2000) made the following conclusion regarding the pattern of activation that they suggested supported consciousness:

> re-entrant signaling within the cortex and between the cortex and the thalamus, bolstered by fast changes in synaptic efficacy and spontaneous activity within the network, can serve to rapidly establish a transient, globally coherent process which is distinguished by strong and rapid interactions among the participating neuronal groups.

Any theory that aims to describe consciousness, as it is described above, would then have to capture all these aspects. That is, it would have to mimic the important role of integration, differentiation, activation and synchrony. This is what the IITC sets out to do.

Quantity of consciousness

Entropy. In 1948, Shannon published his highly influential paper "A Mathematical Theory of Communication" in which he defined entropy within an informational context. Entropy is here specified as the average information that one is missing when one only has access to the probabilities of all the different outcomes of some probabilistic variable but not the actual outcomes themselves. It can also be seen as the average amount of surprise a rational observer will experience when faced with the actual outcomes when only knowing about the probability distribution. For example, if one would flip an unfair coin which showed heads 95% of the time, an outcome showing heads would not be very surprising. Granted, every tail showing up would be highly surprising but since this event would not occur very often, the average amount of surprise yielded from a flip of this coin would be quite low. On the other hand, if the coin was fair, the uncertainty as to what side would come up would be as great as it could be before each individual flip. That is, no guess, before flipping the coin, regarding what side would come up, would be more rational than the other and therefore, each flip would yield quite a bit of surprise.

More formalized, the entropy, H, is defined as

$$H(p) = \sum_{i=1}^{n} p_i \log_2 \left(\frac{1}{p_i} \right) \tag{1}$$

where p_i is the probability of a certain outcome, i, contained within the probability distribution, p, which consists of all the probability values, $\{p_1, \ldots, p_n\}$, for all the possible outcomes, $\{1, \ldots, n\}$. So, applied to the examples given above, the entropy formula says that any outcomes of the unfair coin would on average yield $0,95 \cdot \log_2(\frac{1}{0,95}) + 0,05 \cdot \log_2(\frac{1}{0,05}) \approx 0,29$ bits of information while the fair coin would on average yield $0,50 \cdot \log_2(\frac{1}{0,50}) + 0,50 \cdot \log_2(\frac{1}{0,50}) = 1$ bit of information.

Information. This concept of informational entropy can be applied to an information handling system, something that is done within the IITC to account for the "information" (in contrast to the "integrated") part of the theory (Balduzzi & Tononi, 2008). The model that the IITC mainly is applied to for explanation and exemplification is a neural network where the communication between nodes takes place in a stepwise fashion, that is, with discrete time steps, and where each node can be either on or off (see Figure 1)[2]. The activity of the network depends only on the activity of the preceding time point meaning the nodes are memoryless. A rudimentary account of how the theory could be applied to a neural network working within continuous time has also been developed but it is not necessary to take this model into consideration to understand and discuss the basic idea of the IITC.

[2] All the figures in this thesis has been created by the author himself, with Figure 4 and Figure 5 being heavily inspired by the depictions in Balduzzi and Tononi (2009).

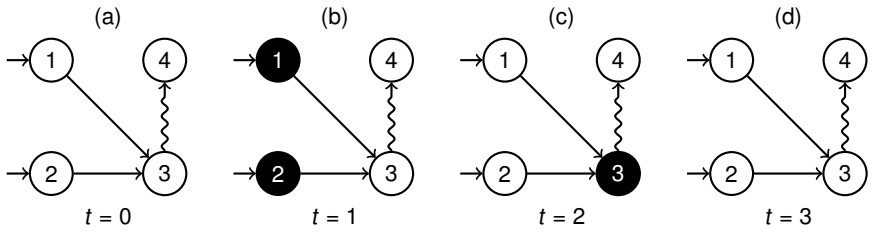

Figure 1. Example of a neural network with activity spreading through it in time. The arrows from node 1 and 2 to node 3 represent an AND-gate. That is, for node 3 to be activated, both node 1 and 2 have to be activated at the preceding time point. The arrow from node 3 to node 4 represent a noisy copy mechanism with a fidelity of 50%. That is, if node 3 was activated at the preceding time point, the probability that node 4 will be activated at the current time point is 0.5. If node 3 was not activated at the preceding time point, node 4 will never be activated in the current one. Node 1 and 2 are activated by means of some unknown input mechanism from outside while node 3 and 4 only are activated in response to activity from other nodes in the network. When a node gets activated, it stays activated for the current time point after which the activity dies out. In the figure, activated nodes are depicted by filled circles. (a) The network at time t=0. Every node is silent. (b) The network at time t=1. Node 1 and 2 has been activated by some unknown mechanism. (c) The network at time t=2. Since node 1 and 2 was activated at t=1, node 3 now gets activated through the AND-gate. The activation in node 1 and 2 has died out. (d) The network at time t=3. Since node 3 was activated at t=2, node 4 had an 50% chance of getting activated in the current time point. However, this copying mechanism obviously failed as evident from the lack of activation of node 4. The whole network is once again, as it was at t=0, silent.

The crucial point of the "information" part of the theory is to what extent the system reduces uncertainty regarding which activation pattern/s could have led to the current state of the system, this in comparison to all the possible preceding states it could have been in, given that nothing is known about how the system internally communicates with itself. The more uncertainty that is reduced in this respect, the more information is contained in the system.

As an analogy, not described in the literature about the IITC, imagine a person, let us call her Samantha, who is located somewhere in central Stockholm. Based on this information alone, one can infer that she, one hour ago, could have been in Uppsala, a city located only 71 km north of Stockholm (a distance one easily can travel under an hour), guess that she probably wasn't in London, a city located 1431 km away (she would have had to fly a fighter jet and parachute out over downtown Stockholm to transfer between the two cities in an hour, a not very common event), and know that she definitely wasn't in Wellington, a city located about 17405 km away on the very opposite side of the earth (unless Samantha is in possession of a teleporter, she is out of luck here). Further, if one have more information, for example about Samantha's personal interests, her friends whereabouts and current events in the local area, one could start to assign different probability values to her whereabouts one hour ago. In the same way, one can, through inspection

of the neural network and its connections, infer which activation patterns could have preceded the current one and how likely each one is (see Figure 2). However, different probability values are only possible if any of the communication mechanisms has some kind of random element to it or if the mechanisms behind the input to one or more of the nodes is not known and therefore is considered as unpredictable extrinsic noise.[3]

This ratio measure of actual possible former states of a system compared to possible states when nothing is known about how it internally works is called the effective information, *ei*, of a certain system, X_0, in a certain state, x_1 (also known as a *system state*), and is defined as

$$ei(X_0 \to x_1) = H[p(X_0 \to x_1) \| p^{max}(X_0))] \tag{2}$$

where $p^{max}(X_0)$ is called the *a priori* repertoire and $p(X_0 \to x_1)$ is called the *a posteriori* repertoire (both these concepts will be explained below). Further, $H[p\|q]$ denotes the relative entropy, or Kullback-Leibler divergence, which is defined as

$$H[p\|q] = \sum_{i=1}^{N} p_i \log_2\left(\frac{p_i}{q_i}\right) \tag{3}$$

where p and q are two probability distributions with the same number of possible outcomes.

Relative entropy is best understood in the context of coding theory (cf. Ling and Xing, 2004). If one wants to send a message, reporting about a certain set of outcomes, the best strategy, given that one wants to send a message that is as short as possible, is to assign the shortest code lengths to the statistically most common outcomes. In this way, the total number of symbols being transferred will, on average, be as few as possible. The relative entropy formula states how many extra bits each sample of an outcome from a certain probability distribution, p, is expected to be when the coding itself is composed with another probability distribution, q, in mind. So, for example, for the coin tossing example described on page 4, the relative entropy of the fair coin compared to the unfair one would be: $0.50 \cdot \log_2(\frac{0.50}{0.95}) + 0.50 \cdot \log_2(\frac{0.50}{0.05}) \approx 1.20$ bits. However, note that the relative entropy operation is not symmetrical, meaning that the reversed case, the relative entropy of the unfair coin compared to the fair coin, will yield another result: $0.95 \cdot \log_2(\frac{0.95}{0.50}) + 0.95 \cdot \log_2(\frac{0.95}{0.50}) \approx 1.76$ bits.

The a priori repertoire, $p^{max}(X_0)$, denotes the probability distribution consisting of the probability values for all possible states preceding the current one when the communication mechanism between the nodes is *not* considered (Balduzzi & Tononi, 2008). That is, each preceding activation pattern gets the same probability value and no activation pattern is seen as impossible, even if the actual communication mechanism renders it so. Using the geographical location analogy above, this would come about when we assign the same probability values for each candidate city as to

[3] When it comes to real world applications, unless the system consists of the whole universe it is in, it *will* interact with the world outside of its scope, thereby receiving such, subjectively perceived, random input.

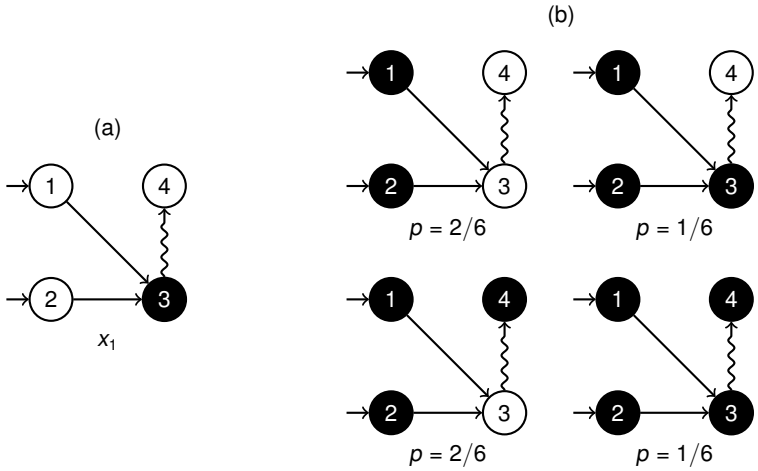

Figure 2. Example of deduction of former network states. The network has the same inner workings as the one in Figure 1. (a) The state x_1 is observed. (b) Given the state x_1, there are four different states that possibly could have preceded it. However, they are not equiprobable as can be seen from the probability values given for each possible activation pattern. This is because of the unknown activation mechanism for node 1 and 2 and the noisy copy mechanism from node 3 to node 4. First of all, both node 1 and 2 must have been activated at the preceding time point since node 3, which is set up to an AND-gate from node 1 and 2, is active at the current time point. Further, the fact that node 4 is inactive at the current time point could have been preceded by both activity and inactivity in node 3 at the preceding time point. If node 3 was inactive, it would have lead to inactivity in node 4 in all possible cases. However, because of the noisy copy mechanism, if node 3 was active, it would have lead to inactivity in node 4 in 50% of the cases. This means that in $2/3$ of the cases when node 4 is inactive, it was preceded by activity in node 3, and in $1/3$ of the cases it was preceded by activity in node 3. Lastly, the activation of node 4 could have, with equal probability, been either on or off at the preceding time point since it would not have affected the current state in any way. These deductions then gives us the necessary means to compute the probability for the different possible preceding states (here, the first number in the name of the state represents the activation of the first node, the second number the second node, and so on): $p(x_{1100}|x_1) = \frac{2/3}{2} = 2/6$; $p(x_{1110}|x_1) = \frac{1/3}{2} = 1/6$; $p(x_{1101}|x_1) = \frac{2/3}{2} = 2/6$ and $p(x_{1111}|x_1) = \frac{1/3}{2} = 1/6$.

where Samantha might have been an hour ago. This would for example mean that the probability that she was in Wellington an hour ago, or any other city for that matter, would be as high as the probability that she was in Uppsala at that time. Since each outcome is as likely as all the others, that is, no guess regarding the preceding state is more rational than the other, the entropy of the a priori repertoire coincides with the maximum possible entropy for the given system (cf. the coin tossing example regarding the fair coin on page 4).

The a posteriori repertoire, $p(X_0 \rightarrow x_1)$, denotes the probability distribution consisting of the probability values for all possible states preceding the current one when the communication mechanism between the nodes *is* considered. That is, contrary to the *a priori* repertoire, here the communication mechanisms between the nodes exclude impossible states and assigns probability values to the others. Using the geographical location analogy, we could now exclude Wellington as a possibility of Samantha's prior whereabouts and instead assign Uppsala with a relatively high probability value.

As can be seen from the definition of effective information (see Figure 3 for a visualization of it) on page 6, it takes on its maximum value for a system when the entropy of the a posteriori repertoire is as small as possible. The most extreme example of this, which is possible for some systems, is when the a posteriori repertoire is zero. This comes about when only one state is inferred with probability 1 since that means that no uncertainty is resolved when one observes the actual outcome (which was the only probable alternative and therefore also necessary). Using the geographical location analogy, if we somehow could be absolutely sure that Samantha was situated in Uppsala an hour ago (we overheard a telephone call where she stated that this was the case), we could assign that outcome a probability value of 1 and the effective information would be at its maximum.

Integration. The "integrated" part of the IITC comes about when one takes under consideration how much information of a system state that cannot be reduced to the sum of information contained within the separate parts of any permutation of the system. That is, if information is lost when the system state is divided up into separate parts and the information given from each separate part is just summed up, compared to if all the information of the system state is computed in one big sweep, the system state is integrated to some extent.

Integrated information, ϕ, in a certain system, X_0, in a certain state, x_1, is defined as

$$\phi(x_1) = H\left[p(X_0 \rightarrow x_1) \| \prod_{M^k \in P^{MIP}} p(M_0^k \rightarrow \mu_1^k)\right] \tag{4}$$

where M_0^k stands for the k^{th} part of the original system, X_0, under some partition, P; μ_1^k stands for the state of this k^{th} part of the original system and P^{MIP} denotes the so called *minimum information partition* . The minimum information partition is basically whatever partition of the system, with

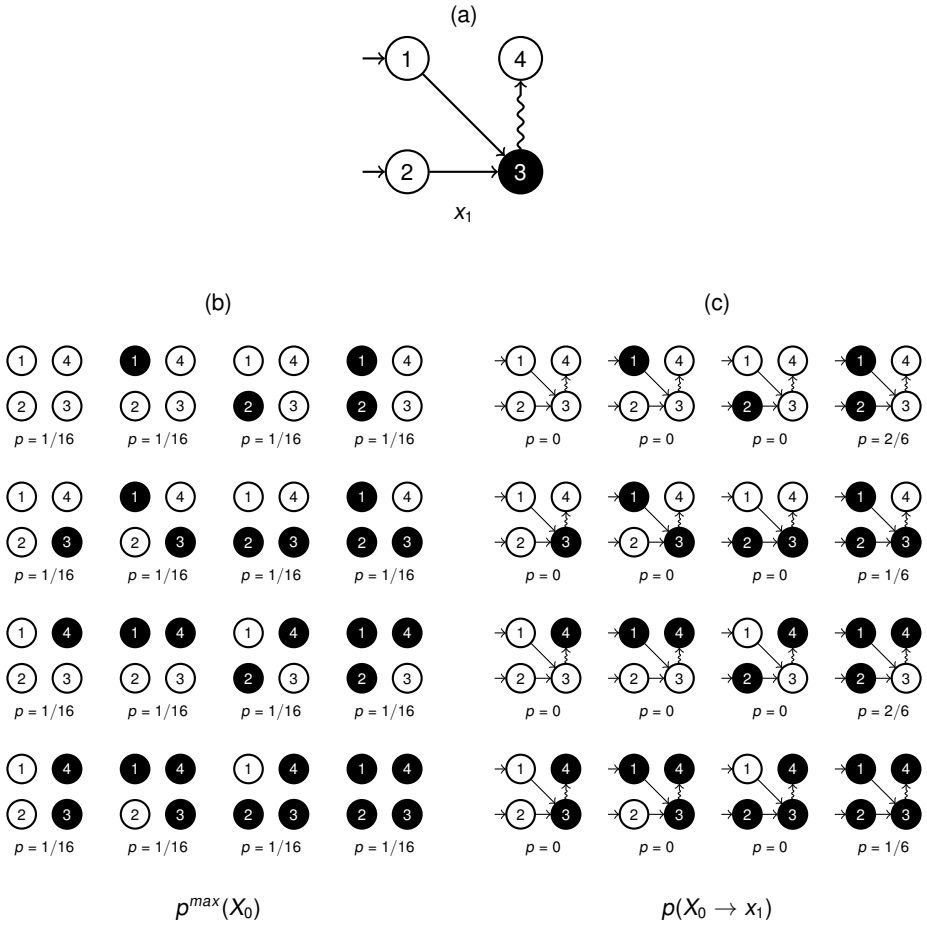

Figure 3. Visualization of the (*b*) *a priori* and (*c*) *a posteriori* distributions of (*a*) a certain system with the same internal workings as the one in Figure 1, in a certain state, x_1. As can be seen, in the (*b*) *a priori* distribution, when the communication mechanisms between the nodes is *not* considered, all alternatives are equally possible. However, in the (*c*) *a posteriori* distribution, when the communication mechanisms between the nodes *is* considered, only the states in the rightmost column are assigned p-values over zero. These values corresponds to the actual possible alternatives outlined in Figure 2. The effective information of the system then becomes: $ei(X_0 \to x_1) = H[p(X_0 \to x_1)||p^{max}(X_0)] = \sum_{i=1}^{N} p_i \log_2(\frac{p(X_0 \to x_1)}{p^{max}(X_0)}) = 2 \cdot \frac{2}{6} \cdot \log_2 \frac{2/6}{1/16} + 2 \cdot \frac{2}{6} \cdot \log_2 \frac{2/6}{1/16} + 12(0 \cdot \log_2 \frac{0}{1/16}) \approx 2.08$ bits.

the addition of a normalization process, that makes the ϕ-value as small as possible for the current state, x_1:

$$P^{MIP} = \arg\min_{p} \left\{ \frac{ei(X_0 \rightarrow (x_1/P))}{N_p} \right\} \tag{5}$$

where the normalization of a certain partition, N_p, is defined as

$$N_p = (m-1) \cdot \min_{k} \{H^{max}(M_0^k)\} \tag{6}$$

in which m denotes the number of parts that the partition consists of. This normalization process is introduced to adjust for two things: (a) The fact that partitions into many parts tend to yield higher effective information in a system than partitions into fewer parts, and (b) the fact that bipartitions, where one of the parts contains just one element, tends to yield less effective information than bipartitions with equal sized parts.

If the geographical localization analogy presented on page 5 is expanded a little bit, it can be applied to the concept of information integration too. Instead of considering just one person whose whereabouts at an earlier time point we are trying to discern, we can introduce several persons at the same time. For example, imagine two persons, Eddie and Rose. Eddie is situated in Stockholm and Rose is situated in Wellington. Further, assume that they act completely independent of each other, that is, the behavior of one of them does not, in any way, affect what happens to the other (this would for example be broken if they were friends and could call each other). Now, if one gathers information regarding, for example, Eddies' local circumstances, saying that the train between Stockholm and Uppsala has been canceled for the last hour, this does not in any way help in figuring out where Rose might have been an hour ago.[4] However, if one acquires information that all the flights to and from Stockholm have been canceled for the last hour, this might be useful for finding out where Rose has been as well. A cancelation of that size might be due to some international situation, making it slightly more likely that there are some problems with the flights in and out of Wellington too. In the latter situation, one would loose relevant information and the result would contain more uncertainty if the problem was divided up into two, one focusing on the Stockholm area and one focusing on the Wellington area, giving each problem to different persons and not allowing them to exchange any information between each other. This means that the information in this latter example, where several flights were canceled, is integrated while it is not integrated in the earlier example where only a single train was canceled. A more elaborate example of integrated information and its logical implications will be given on pages 14–15 when discussing the quality of consciousness aspect of the IITC.

[4]This is not entirely true since everything that lies within the light cone of Rose's possible locations one hour back in time will also affect her, an aspect that will be ignored right now for the sake of the argument. This is after all only an illustrative analogy.

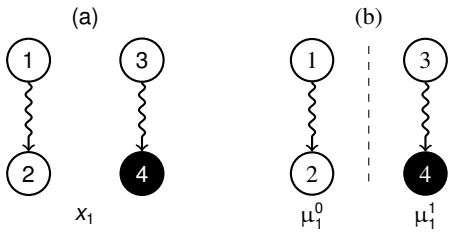

Figure 4. (a) Example of an idealized camera instantiated in the form of two separate photodiodes. Each arrow represents a copy mechanism with 100% fidelity. The nodes at the top layer, that is, node 1 and 3, are activated in an unknown fashion and are therefore seen as random variables. The effective information for the whole system, X_0, in the current state, x_1, is: $ei(X_0 \rightarrow x_1) = 4 \cdot (\frac{1}{4} \cdot log_2(\frac{1/4}{1/16})) + 12 \cdot (0 \cdot log_2(\frac{0}{1/16})) = 2$ bits. (b) However, the integrated information for the whole system is zero. For the minimum information partition, when the system is divided up into two separate parts containing one photodiode each, the integrated information becomes: $\phi(x_1) = H[p(X_0 \rightarrow x_1) \| p(M_0^0 \rightarrow \mu_1^0) \cdot (M_0^1 \rightarrow \mu_1^1)] = 4 \cdot (\frac{1}{4} \cdot log_2 \frac{1/4}{1/4}) = 0$. Here, the two a posteriori repertoires contains two states each but when they are multiplied with each other they yield a probability distribution consisting of four states, the same four states that are contained within the a posteriori repertoire for the whole system.

One example that is brought up on several places in the IITC literature is the example of an idealized digital camera. This camera consists of a number of completely separated photosensors, each one a small neural network built up by two nodes and a simple copy mechanism from one of them to the other, meaning that whatever state the first node is in gets copied to the other one. In this way, one of the nodes, the one which has an efferent connection from it, represent the light that is captured by the camera, while the other node, the one which has an afferent connection to it, represent the sensor of the camera, capturing the image. Since each photodiode acts completely independent, that is, have no connections to any other photodiode, no matter how many megapixels we imagine this camera to have, the integrated information, ϕ, taken over the whole camera, will always be zero, while indeed the effective information for the whole system will continue to grow with the size of the system (see Figure 4).

Complexes. The last concepts important to consider for the quantity of consciousness part of the IITC are the concepts of *complexes* and *main complexes*. For a certain system, X, a subset of this system, S, forms a complex when it enters a state, s_1, such that

$$\phi(s_1) > 0 \tag{7}$$

and

$$\phi(t_1) \leq \phi(s_1) \tag{8}$$

where t_1 is the state of any system T, where $S \subset T \subset X$. In other words, S is a complex when it is in such a state, s_1, that it to some extent integrates information and when S is not part of any bigger system, T, that currently generates an even higher ϕ-value. Further, S forms a main complex if S is a complex and

$$\phi(r_1) < \phi(s_1) \tag{9}$$

where r_1 is the state of any system R, where $R \subset S$. In other words, if S is a complex and each of its parts are necessary to keep the ϕ-value at the same level, that is, no set of parts could be taken away from S without affecting the ϕ-value negatively, then S is a main complex.

Relation to consciousness. In the IITC framework, the ϕ-value of a system that is a main complex is said to correspond to the level of consciousness it experiences. That is, only systems that are main complexes are said to have any consciousness at all. This has the consequence that even if the ϕ-value for a given system, X, is nonzero, this does not necessarily mean that X has consciousness to any degree, even though some subset of X must.

If one considers the idealized camera (presented on pages 10–11 and in Figure 4) in the context of IITC, the camera itself, seen as a whole, has no conscious percept at all, since the integrated information for the whole system is zero. Instead, each photodiode is minimally conscious since each one of them generates a certain minuscule amount of integrated information (and in the process also forms main complexes). This example highlights the fact that the IITC is a panpsychic theory. That is, it states that everything that can be said to perform some kind of computation is conscious, although not necessarily to an especially high degree, and not necessarily as a unity rather than a number of separate parts with separate consciousness. The theory, however, assigns more consciousness to systems that generally are believed to have more of that, so that, for example, a brain is deemed as more conscious than a photodiode because of its higher amount of integrated information.

Now, to compute the actual ϕ-value of a system of any significant size is virtually impossible due to the enormous computational resources needed to perform this. Even for relatively small systems, it is a good idea to only consider bipartitions of it, a procedure that specifies the lower bound of what value ϕ possibly could take when considering the total partition.

When a system is both functionally specialized and functionally integrated, that is, when elements of the system both have unique types of connection patterns to the other nodes and each element have connections to all the other nodes, the ϕ-value of the system is at its greatest (Tononi & Sporns, 2003). Evidence suggests that the thalamocortical system seems to have a capacity for both high functional specialization (Bartels & Zeki, 2005) and high functional integration (Engel, Fries, & Singer, 2001). One way of testing this, through approximating the relative amount

of integrated information in the cortex during different states, is through utilizing a combination of transcranial magnetic stimulation (TMS) and high-density electroencephalography (Massimini, Boly, Casali, Rosanova, & Tononi, 2009). Here, TMS is applied to a certain area of the cortex, and the propagation effect of this is then measured with respect to neural activity (Komssi & Kähkönen, 2006).

In a series of experiments (Massimini et al., 2005, 2007), this technique was applied to subjects during wakefulness, slow wave sleep and REM sleep. This investigation showed that during wakefulness and REM sleep, states usually associated with a high level of consciousness, perturbations propagated across the whole cortex. In addition, disruptions of different brain areas generated different type of EEG patterns. On the other hand, during slow wave sleep, perturbations of the same magnitude did not propagate outside the specific area that was stimulated, and local EEG patterns were similar in response to stimulation in different places. Only when the perturbations were increased to a certain degree did the effect spread outside of the stimulated area, but only in a very nondiscriminatory way where a simple slow wave was produced as a response in the cortex.

In other words, during slow wave sleep, the cortex has less capacity for functional specialization and integration than during wakefulness and REM sleep. This means that the cortex during slow wave sleep either is divided up into separate functionally isolated modules or, when the activation in any local area is strong enough, produces a simple, homogenous and non-specific response to different inputs. In terms of the IITC, the ϕ-value for the cortex as a whole gets lower during slow wave sleep, something which is predicted by the theory if subjects in this state are assumed to be less conscious than in the other states mentioned. Further, if the bistable activation patterns of slow wave sleep are simulated in a small neural network (Balduzzi & Tononi, 2008), it can be shown that no significantly positive ϕ-value can be maintained for any longer time. Rather, the integrated information collapses to zero every time the system reaches an extreme hyperactive state, thereby giving more credence to the hypothesis that slow wave sleep does in fact not generate any high ϕ-value in a brain.

Quality of consciousness

Qualia space. The qualia space, Q, is an abstract space consisting of as many dimensions as there are possible states in the a priori repertoire of the system in question (Balduzzi & Tononi, 2009). For neural networks consisting of nodes with two possible states each, on or off, the number of dimensions in the associated qualia space will therefore be 2^n, where n stands for the total number of nodes in the system. Each axis in the qualia space is ranged from 0 to 1, representing probability values of the very state that the axis denotes. Any given system state will be represented inside the qualia space as a polytope with the edges of this multidimensional shape being set by the probability values for every state in the a posteriori repertoire (see Figure 5).

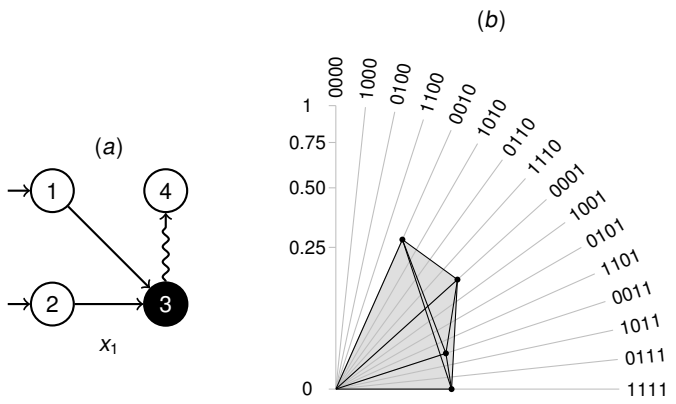

Figure 5. (a) The state x_1 of a system X_0, with the same internal workings as the one in Figure 1, is observed. (b) This generates a polytope within a 16-dimensional qualia space, here depicted in a two-dimensional fashion (note that the axes are plotted in a square root scale). Each axis represents a certain state and the length of the axis ranges from $p = 0$ to $p = 1$. The dots represents the probability for each state in the a posterior repertoire for all the states where $p > 0$ (for all states where $p = 0$, the dots, if they would be depicted, are in the origin of the graph). Together, these denotes the edges of a polytope within the qualia space.

Concepts. A q-arrow is a vector, drawn from one point within the qualia space onto another. When considering the whole communication mechanism in a system, that is, all the connections between all the different nodes, the q-arrow will be drawn from the origin to the point that represents the combination of all probability values for all states in the a posteriori repertoire. However, one can also consider subsets of the communication mechanism, either in the null context, that is, with the q-arrow starting from the origin (which effectively describes the same scenario as the example above), or in a context of a non-empty subset of the communication mechanism, in which case the q-arrow will start from wherever the q-arrow generated by the context subset ends.

A q-arrow is said to be *tangled* or constitute a *concept* when it cannot be decomposed into the sum of separate q-arrows, each one of these separate q-arrows denoting a certain part of the set of communication mechanisms that the original q-arrow is generated by. To illustrate this, in a way that is not described within the literature about the IITC, imagine a person, let us call him Frank, who is going to visit Oslo. We know that he, to get to Oslo, can take traveling route *A* or *B*. We also know that the individual likelihood for both these traveling routes to be closed down is 0.5. Now, suppose that we see Frank in Oslo. We can then infer that traveling route *A* and/or *B* was open, otherwise he would not have been able to get there.

Suppose that we split up our information about Frank's possible traveling routes and give the information to two different observers, *X* and *Y*. We tell *X* that one of Frank's two possible traveling routes is *A* and that it is closed down 50% of the time. We tell *Y* the same thing but substitute *A* with *B*. Now imagine that *X* sees Frank in Oslo. For all she knows, Frank could have taken route *A*

or this other route (which she does not know is B). If she is going to assign any probability value to this other, for her unknown, traveling route that Frank could have taken, she will have to treat it as noise and say that it, too, is closed down 50% of the time. Out of this information, she can infer that Frank had a probability of $1 - (0.50 \cdot 0.50) = 0.75$ chance of getting to Oslo.

Before X makes an observation regarding if Frank has arrived in Oslo or not, there are four different possible scenarios: (a) Both traveling routes could be closed down, (b) both traveling routes could be open, (c) traveling route A could be open while the unknown traveling route could be closed, and (d) traveling route A could be closed while the unknown traveling route could be open. Each one of these scenarios has a likelihood of $0.50 \cdot 0.50 = 0.25$ of happening. When X subsequently observes Frank in Oslo, there is three equiprobable different scenarios left that could have preceded his arrival, since both traveling routes being closed down now is out of the picture. This means that X can infer that A probably was open since there only is $1/3$ probability left that it would be closed. If we ask X specifically about traveling route B, she will not be able to say anything since she does not know whether that was one of Frank's possible traveling routes or not. The only thing that she can say is that A probably was open. This whole train of thought presented above will be exactly the same for Y, only that she can infer that B probably was open.

Now, notice how the combined information of X and Y, that A probably was open and that B probably was open, is not the same as our original statement that either A or B was open since the combined information of X and Y does not exclude the possibility that both A and B were closed down, but only states that it is unlikely. This means that the information we get when we consider Frank's possible traveling routes, without dividing up our prior information about the routes, is tangled or forms a concept.

Compare this with the scenario where Frank does not arrive in Oslo. X can infer that A had to be closed down since Frank otherwise could have taken that route. Y can infer the same thing with respect to B. The combined information of X and Y then says that both A and B were closed down, the same thing as a person knowing all the information about Frank's possible traveling routes would have inferred. The latter person's information is then *not tangled*.

This whole analogy with Frank can be related back to a neural network to illustrate how tangled information is represented there. This is explained in Figure 6.

More formalized, a q-arrow is a concept when the amount of *entanglement*, γ, of that very q-arrow is greater than zero. Entanglement of a q-arrow is defined as

$$\gamma(X_0(m, x_1) \rightarrow X_0(m \cup r, x_1)) = H\left[X_0(m \cup r, x_1) || \prod_{M^k \in MIP_\gamma} M_0^k(m \cup r^k, x_1)\right] \tag{10}$$

where MIP_γ stands for the minimum information partition with regard to the entanglement, that is, the partition of the system that makes the entanglement of the q-arrow in question as small as

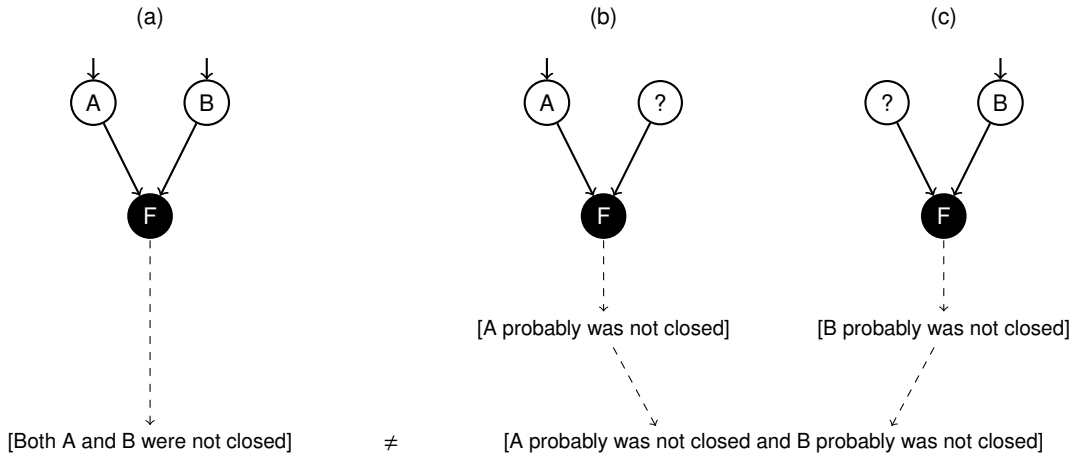

Figure 6. The traveling example, described on pages 14–15, depicted as a neural network with two nodes (A and B) instantiating an OR-gate with a third node (F). Here, node A and B can be seen as representations of the two different traveling routes, where an activation of a certain node means that that traveling route is open, and the output node can be seen as a representation of the whereabouts of Frank, where an activation means that he is in Oslo. (a) An observer knowing about both A and B and their connection to F sees Frank in Oslo. The observer can deduce that [Both A and B were not closed]. (b) An observer knowing only about A and its connection to F sees Frank in Oslo. The observer can deduce that [A probably was not closed]. (c) An observer knowing only about B and its connection to F sees Frank in Oslo. The observer can deduce that [B probably was not closed]. Now, notice that the statement yielded from a combination of (b) and (c), [A probably was not closed and B probably was not closed], is not the same as the statement from (a), [Both A and B were not closed]. This means that the latter information is tangled.

possible (the mathematical definition is given further down). Further, m and r are subsets of the set $\top$ which contains all the communication mechanisms between the nodes in the current system X_0.

As can be seen, entanglement is always measured within a certain context, m, where r is added to it. That is, the minimum information partition is only applied to r, and the resulting entanglement value is only about the q-arrow that starts at the point in the qualia space that is generated when only the communication mechanism context, m, is considered (this context can, however, be empty, and therefore put the starting point in the origin of the qualia space).

The minimum information partition for entanglement is defined similarly as the minimal information partition for information integration, namely

$$MIP_\gamma = \arg\min_{\wp} \left\{ \frac{\gamma(X_0(m, x_1) \to X_0(m \cup r, x_1)/\wp)}{N_\wp} \right\} \tag{11}$$

where the normalization, $N_\wp$, is defined similarly and with the same motivations (described on page 10) as in the quantity part of the theory, namely

$$N_{\varphi} = (l-1) \cdot \min_{k}\{H^{max}(R_0^k)\} \tag{12}$$

in which l stands for the number of parts where $S^k \neq \emptyset$, that is, the number of parts that are nonempty.

Modes. A concept that is drawn between a certain starting point, a, in Q, to the absolute top of the quale, $\top$, that is, the point where all the communication mechanisms between all the nodes has been considered, in the formula below denoted as *mech*, is defined as a *mode* if (a) the concept is tangled to any degree, that is, if

$$0 < \gamma(X_0(a, x_1) \rightarrow X_0(mech, x_1)) \tag{13}$$

and (b) there exists no b, where $a \subset b$, such that

$$\frac{\gamma(X_0(b, x_1) \rightarrow X_0(T, x_1))}{H^{max}(B_0)} < \frac{\gamma(X_0(a, x_1) \rightarrow X_0(T, x_1))}{H^{max}(A_0)} \tag{14}$$

where A_0 and B_0 respectively are the nodes directly affected by the communication mechanism contained in a and b. In other words, if a certain q-arrow, v, is tangled, and more densely tangled than any of v's constituting q-arrows, where the amount of denseness is decided by the entanglement value of the q-arrow divided by the maximum entropy of the nodes that directly contribute to the constitution of the q-arrow, then v also makes up a mode. Further, a certain mode, u, is a *sub-mode* if u is a proper part of a larger mode, and an *elementary mode* if u is not a *sub-mode* and all the q-arrows u can be broken down into have strictly lower entanglement values than u itself. Notice the parallel to complexes and main complexes presented on pages 11–12.

Relation to consciousness. In the context of the IITC, a concept, that is, a q-arrow in Q, specifies the content of what is consciously perceived, provided that the q-arrow is an elementary mode or a sub-mode. The elementary mode specifies the whole conscious percept, for example the whole experience of being at a concert. The sub-modes, on the other hands, specifie the sub-modalities of the experience such as for example auditive and visual modalities. Further, sub-modes that are contained within these modalities specifies sub-modalities, for example the experience of form or motion which are sub-modalities of vision. The sub-mode at the bottom of the hierarchy specifies a concept that cannot further be broken down to any constituting experiences, for example the experience of seeing pure redness, something which generally is seen as a basic experience that cannot be further divided.

More concretely, each possible elementary mode of a system specifies a unique conscious experience. The more tangled it is, the more content rich the experience is, since more information is generated when the specific discrimination of the q-arrow is made. As analogy, imagine that a photodiode (taken from the idealized camera presented on pages 10–11) and a human being is

watching a blank screen that either is on or off. Both will be able to make a discrimination between when the screen is on versus when it is off. However, while the photodiode does this by being in one of two different states, thereby only reducing uncertainty for two possible outcomes, the human being does not only discriminate the lit screen from the dark screen; she also discriminates it from every possible visual stimuli that could have popped up on the screen (given that the difference between the different stimulus is large enough so that a discrimination actually can be made; e.g not a single pixel on a super high resolution screen) which is a huge number. Thereby, when the human being watches the screen, uncertainty is not only reduced for the two possible outcomes of the screen being on and off, but at the same time also for every possible stimuli that could have appeared.

As can be seen from their respective definitions, a main complex and a main mode will always overlap fully. This means that the amount of consciousness, ϕ, really is a measure of how content rich a certain experience is. That is, ϕ does not refer to something extra over and above the amount of content of the experience. When talking about the amount of consciousness for a certain system then, the concept of amount of consciousness can be broken down to the content richness of the experience.

Examination of the IITC

Given the review of the IITC above, it is now possible to turn to examining whether the theory succeeds in what it sets out to do, namely to explain under what circumstances consciousness arises. In this section, empirical, theoretical and probabilistic reasons will be investigated in order to answer this question.

Consciousness space and territories

When discussing the IITC, it will be fruitful to introduce the concept of the *consciousness-space*, C. This is a two-dimensional space with the amount of consciousness, that is, ϕ, assigned to one axis and all the different possible states of all the different possible systems in the world assigned to the other axis. More formally, the latter axis depicts all the elements of the set $X = \{s \mid s \text{ is a system state in some system, } Z\}$.[5] To be clear, Z can here take the form of *any* conceivable system in the world, built up by *any* combination of separate parts one can imagine. Examples of such systems could be a specific brain, a specific toaster, a specific brain plus a specific toaster, *et cetera*. Further, s can take the form of any state of the system it belongs to. Examples of different system states would then be a specific brain processing a specific color, a specific toaster toasting two specific slices of bread, a specific brain processing a specific sound plus a specific toaster not toasting anything, *et cetera*.

[5] Note that both ϕ and X, depending on the ontological reality of the world, could be either finite or infinite.

Within *C*, one can subsequently depict different territories[6], where a territory is defined as a number of system states where each system state is assigned a ϕ-value. A territory can then be depicted as a two-dimensional continuous or non-continious function in *C*.

Now, for the IITC, there are basically four different territories that make sense to talk about: the observed, the postulated, the generated and the actual. *The observed territory* is made up of all the empirical observations of consciousness.[7] *The postulated territory* draws from the observed territory and lays out a suggestion what *the actual territory*, which depicts the actual ontological fact of the matter, looks like. *The generated territory* is, as the name suggests, generated by some kind of rule of inference; in the case of the IITC, the mathematical functions that describes how ϕ is computed. A mock-up graphical illustration of the different territories is given in Figure 7.

The goal of the IITC is for the generated territory, which in turn is dictated by the postulated territory, to as closely as possible trace the shape of the actual territory. This means that the success of the whole endeavor depends on how well the postulated territory approximates to the actual territory. That is, if the postulated territory is an inaccurate description of the actual territory, it does not matter how well the generated territory fits with the postulated territory; the theory still will not succeed in what it sets out to do, namely to give a description of the actual territory. In computer programming terms, this is called "garbage in, garbage out" and it will be argued in this thesis that this expression is applicable to the IITC.

Definition of consciousness within the IITC

Tononi (2007) has defined consciousness in the following way:

> The definition that I like to use, to avoid misunderstandings, is that consciousness is what fades when we fall into dreamless sleep, an experience that I guess everybody has. Early in the night, if I wake you up, very often, you have absolutely nothing to say. You look indeed like a zombie and if I ask you 'what was going through your mind?'; nothing! You weren't there, the world wasn't there. Everything is gone. That is, experience is - maybe gone is not the right word - but it is so diminished that: who cares?[8]

[6]The word "territory" is here not arbitrarily chosen, but rather follows a tradition where one speaks of the map's relation to the territory (e.g. Korzybski, 1933). In this thesis, however, the concept of the map will be dropped in favor of just specifying different kind of territories.

[7]As such, it is not theory independent but rather depends on the subjective assessment of the strength of different claims (see section "Measurement of consciousness" on pages 21–27).

[8]This exact quote is taken from a public lecture but similar definitions have also been stated in actual scientific articles about the IITC, for example in Tononi (2008), where it is stated that 'everybody knows what consciousness is: it is what vanishes every night when we fall into dreamless sleep and reappears when we wake up or when we dream." However, since the IITC predicts that consciousness gets diminished, rather than vanishes all together, when, for example, a human being goes to sleep, the first quote is deemed as more harmonious with the theory than the second and will therefore be the one that is being used in this thesis.

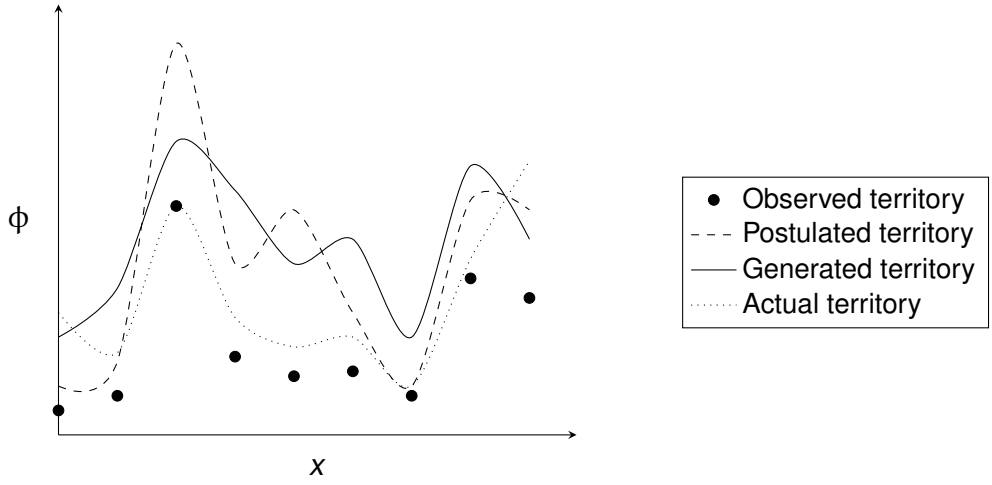

Figure 7. A mock-up graphical illustration of the observed-, the postulated-, the generated- and the actual territory, depicted within *C*. As can be seen, the observed territory is made up of interspersed points. This depicts the fact that the observed territory, from a pragmatic point of view, never will be fully exhaustive but rather will be made up of a fairly limited set of observations. The other three territories are depicted in a way where each possible state in *X* is given a ϕ-value. For the postulated territory, this might or might not be the case for a certain theory. If, for example, certain system states are not assigned any ϕ-values, the resulting shape would contain holes. The same thing goes for the generated territory, which might or might not cover the full set of *X*. However, it is assumed in this thesis that the actual territory is ontologically non-ambigious, that is, that for every system state, there is a fact of the matter which level (if any) of consciousness it generates. However, the critique in this thesis is not dependent on this fact and it would still apply even if consciousness, to some extent, *would* be ontologically ambiguous.

This definition, while indeed being an accurate description of how the level of consciousness varies according to the IITC, does a poor job when it comes to capturing all the possible aspects of consciousness. That is, it presupposes a certain set of territories (for example, territories in which consciousness fades away during sleep) and therefore, by definition, rules out alternative ones. As such, it is a non-starter for a thorough investigation of consciousness. Instead, in this examination section, a more basic and broad definition of consciousness, as will be defined in the section "Re-definition and division of consciousness" on pages 27–29, will be used (this definition does not rule out the possibility that the actual territory indeed looks the way it is defined within the IITC; it only leaves room for alternatives). It will then be argued that there are no justified reasons for supposing that the actual territory, either by empirical, theoretical or probabilistic reasons, looks the way that it is defined within the IITC, meaning that the theory also misses its target of giving a likely explanation and description of the actual territory.

More formally, let

$$PT := \text{territories sufficiently close to the postulated territory of the IITC} \tag{15}$$

and

$$PT^C := U \backslash PT \tag{16}$$

where U is the set of all naïvely possible territories. In other words, PT^C represents all the alternative territories to PT.[9] Now, the precise constitution of the postulated territory of the IITC and the *exact* range of what falls into PT has not been defined and will not be so either, partly because there is no exact definition of the constitution of the postulated territory of the IITC available, but mostly because that would miss the point entirely. The point of this investigation is not to nitpick in possible errors in the details of the postulated territory of the IITC, but instead to present an argument for why the actual territory of consciousness could look *dramatically* different from that postulated territory. Also, if the argument, in the end, is going to be applicable to other theories than only the IITC, theories that might not share the exact same (although very similar) postulated territory, PT will have to be a little more broad than it would be if it only encompassed what is relevant for the IITC. With that said, even though PT will not be *exactly* defined, throughout the discussion, it will become at least sufficiently clear, from a pragmatical point of view, what constitutes PT and what does not constitute PT (that is, PT^C).

Using the definitions of PT and PT^C above, the main discussion in this whole section will be divided up into three parts: First, possible *empirical* reasons for excluding PT^C will be investigated. Second, possible *theoretical* reasons for excluding PT^C will be investigated. Third, possible *probabilistic* reasons for excluding PT^C will be investigated. However, before getting into the main discussion, some basic concepts needs to be gone through. This is what will be done in the sections "Measurement of consciousness" and "Redefinition and division of consciousness" below.

Measurement of consciousness

Basically, there are three different types of claims one can make regarding the consciousness of a certain system state: (a) positive claims (which further can be divided up into ψ-, π-, α-, β-, δ- and λ-positive claims), (b) neutral claims and (c) negative claims. A positive claim is a statement that a certain system state is conscious. A negative claim is a statement that a certain system state *not* is conscious. A neutral claim is a statement that it is unclear whether a certain system state is conscious or not.

The positive and the negative claims can vary in perceived epistemological strength (cf. Lamme, 2006). For example, a certain positive claim could by some observers be seen as stronger,

[9] The "C" in "PT^C" stands for "complement".

or more certain to be true, than another positive claim. However, the strength of the claims will always be subjectively dictated, even though some claims will be inter-subjectively agreed upon by almost everybody, and another observer could therefore, for example, hold the opposite view, stating that the latter claim is stronger than the former.[10] Neutral claims cannot, however, vary in their strength. That is, if one were to go from making a neutral claim, regarding if a certain system state was conscious or not, to favor one of the hypothesis, one would, by definition, make a claim that was either positive or negative and therefore not neutral anymore.

Except for the division stated above, each individual claim can further be placed within a four-field matrix where the position of the claim is decided by whether it denotes a *first person* or *third person* observation and whether it refers to a conscious state in *present* or *passed* time, relative to the time point the last piece of information that lies as ground for the judgment is observed. Whether a claim is based on a first or third person observation is fairly straight-forward. However, what constitutes present versus passed time needs some kind of definition since there exists no precise common sense division between the two.

In a response to some ideas presented by Edelman (1990, Chapter 9), Pöppel (1994) has argued that the perceptual "now" can be set to about 3 seconds. That is, within this time window, information can be treated as a unit, whereas in longer periods of time the information will start to be divided up and processed into compressed chunks that have to be accessed in a secondary fashion. This time frame has also been stressed by Fraisse (1984) as the psychological present, referring to the time window in which an observer can perceive the duration of time directly rather than being restricted to retrospectively reason about it. This definition of the present will lay as ground for the categorization of present and passed time in this thesis. That is, claims that refer to time points longer ago than three seconds will be categorized as claims about passed time, whereas claims that refer to time points up to three seconds ago will be categorized as claims about present time. However, it is important to note that it does not matter if this division, in the end, is arbitrary from an ontological point of view and only makes sense in a pragmatical sense in that it, more or less, approximates human beings subjective experience of the temporal division between now and then.

Lastly, a claim can either be about any experience within the full experience context, that is, the set, E, of all possible experiences, $\{e_1, e_2 \dots e_n\}$, or it can be about a certain subset of E, for example that which encompasses every case of some kind of visual experience. In the former case, the claim will denote if there is any consciousness at all for the specific system state in question, while in the latter case, only the existence of a certain type of experience is mentioned.

[10] Note that the strengths of the different positive claims themselves will not be debated in this thesis. This is a discussion where much can be said, but to not derail from the main arguments to much, it will be left where it is. Instead, for the sake of the argument, a generous stance will be taken where positive claims that in general are accepted by the scientific community also will be accepted here.

First person observations. Following is a list of the possible first person observations one could make when it comes to claims about consciousness:

- ψ-*positive claims* are statements based on direct introspection, that is, claims that each individual can do only about herself, and that states that oneself is conscious of something at the present moment (cf. Descartes, 1641/2008, p. 16-33).

- π-*positive claims* are, just like ψ-positive claims, subjective, but they differ in the respect that they refer to an earlier time point rather than to the present. That is, they are introspective statements utilizing memories, stating that oneself was conscious of something at some earlier point in time.

- *Negative and neutral claims* can, when it comes to first person observations, be made for passed states, both in the partial and the full experience context. However, when it comes to present states, only negative and neutral claims can be made for the partial experience context. That is, it is a contradiction in terms to claim that one has an experience (which a first person claim is) that can support that one does not have any experiences or make it uncertain if one has any experiences or not.

Third person observations. Following is a list of the possible third person observations one could make when it comes to claims about consciousness:

- α-*positive claims* are statements about states of systems where the behavior of the system is taken as a direct proof that it also is conscious of something at that very instant. These claims ranges from statements that generally are seen as unproblematic in their assertion that a certain behavior is coupled with consciousness to statements that stir up more controversy. Direct observation of verbal or non-verbal overt advanced motor behavior is for example, when it comes to consciousness studies, the standard way of assessing if a system in a certain state is conscious of something or not (Lamme, 2006). These can, for example, include situations like when a person says that she notices a certain percept (i.e "I see a red dot right now") or when she responds to a certain stimuli by pressing a button. Examples of more indirect observations are when testing vegetative patients for meaningful responses through different neuroimaging methods. Owen et al. (2006) has for example demonstated how certain vegetative patients shows the same type of BOLD response as awake, healthy controls when asked to imagine playing tennis. Further, Monti et al. (2010) has extended this method to make it possible for certain patients to answer yes or no questions by indicating their response through engaging in either motor or spatial imagery tasks. However, as mentioned, claims based on empirical findings like these are often more contested and discussed

than claims based on more direct observations, exemplified by the debate (Greenberg, 2007; Nachev & Husain, 2007; Owen et al., 2007) following Owen et al.'s (2006) results.

- λ-*positive claims* are statements being made about systems that do not only explicitly state that they are conscious of something but that also show signs of reflecting over their own consciousness and sees it as a problematic thing to explain. These statements therefore constitute a proper subset of the set of all α-positive claims. They are also the only category of claims that are bounded to a specific type of consciousness content. That is, while the other claims can be about any kind of experiences from *E*, the λ-positive claims are by nature restricted to a certain subset of *E*, namely the one where this type of reflection and inquiry is included.

- β-*positive claims* are statements about the consciousness of systems at earlier time points, inferred based on current observations of the system in question. This could for example be the case of someone reporting on the content of a dream after waking up. Here, the verbal description from that person can be seen as evidence that the person was conscious of something earlier during sleep. Also, these claims can be recursive in the way that someone can be observed to state that she remembers having made a π-positive claim earlier even though the content of that particular claim has been forgotten by her. This would indirectly assert that there was some consciousness before the referred to π-positive claim.

- δ-*positive claims* are statements about system states where the consciousness is inferred in a sort of secondary statistical fashion, that is, not directly through some kind of first hand confirmation from the subjects in question, either while it is present, through an α-positive claim, or afterwards, through a β-positive claim.

  Block (2005, 2007) has argued (among others; cf. Lamme, 2010) that there is a surplus of consciousness overflowing the consciousness that people know about themselves. An example of this is Sperling's (1960) classic experiments. In these, arrays of letters were shown to the participants very brief. When asked to recall these letters right afterwards, the participants could only get a couple of them right. However, when they were asked to just recite a specific row of letters, even though this row was specified after the letters had disappeared, the recall rate was much higher. This suggest that the participants had access to quite a few letters right after the presentation but that this access quickly faded away from memory before they had time to report them all. Block (2007) argues that this information is not merely *accessible* for consciousness, that is, that it has the potentiality to become conscious (in this case, if the corresponding row is asked to be recalled), but that it in fact represent conscious information, even when it is not accessed and presented verbally. This is an example of a δ-positive claim.

Another example, not formulated by Block himself, of when one could make δ-positive claims is when it comes to dreams. When someone awakens from sleep and reports that she has been dreaming, this is an example of a β-positive claim. However, often, people wake up without having any recollection of any dreams during the night. This could suggest that they did not dream anything. However, if people are woken up during REM-sleep, they often report dreaming something (Hobson, Pace-Schott, & Stickgold, 2000). This suggests that a subset of those who report not dreaming anything after waking up spontaneously might in fact have dreamt but forgotten it. This is another example of when one could make a δ-positive claim.

It is important to realize that when making δ-positive claims, one is most often only making a probabilistic and not an absolute claim. That is, for any given state where an α-positive claim cannot be made, and where a β-positive claim does not follow, one can only say that there is a certain probability that the subject was in fact conscious of something. This probability could of course be 100% but this only happens if the measurement that lie as ground for the δ-positive claim always is positive, that is, subjects would *always* have to report dreams when woken up and to *always* report *all* the letters in a specific row when prompted to.[11] In every other case, there is no way to be sure whether the subject actually was conscious solely based on reports (or rather, lack of reports) from subjects themselves. One can only come with informed guesses coupled with a certain percentage estimation.

- *Neutral and negative claims* can, when it comes to third person observations, be made, both in the full experience context and with regard to some subset of it.

The landscape of possible claims. Together with the neutral and negative claims, this internal division of the positive claims makes up an exhaustive landscape of all the claims one can make about consciousness (see Figure 8). Note that this landscape covers for example, among other things, panpsychic theories. In this case, one can define the neutral and negative claims in the full context as empty sets and come with α- and β-positive claims regarding systems simply based on the fact that they exist, and can be inferred to have existed earlier, which sets the behavior demand for these positive claims at their absolute minimum. This effectively makes δ-positive claims redundant and therefore also an empty set.[12]

Further, from the definitions given for the different claims, some logical consequences ensue:

[11] In such a scenario, one could even argue that the claim should be categorized as an α-positive claim rather than a δ-positive claim.

[12] However, even though the IITC in essence is a panpsychic theory, for pragmatic reasons, this is not how it will be handled, as will be explained on pages 26–27.

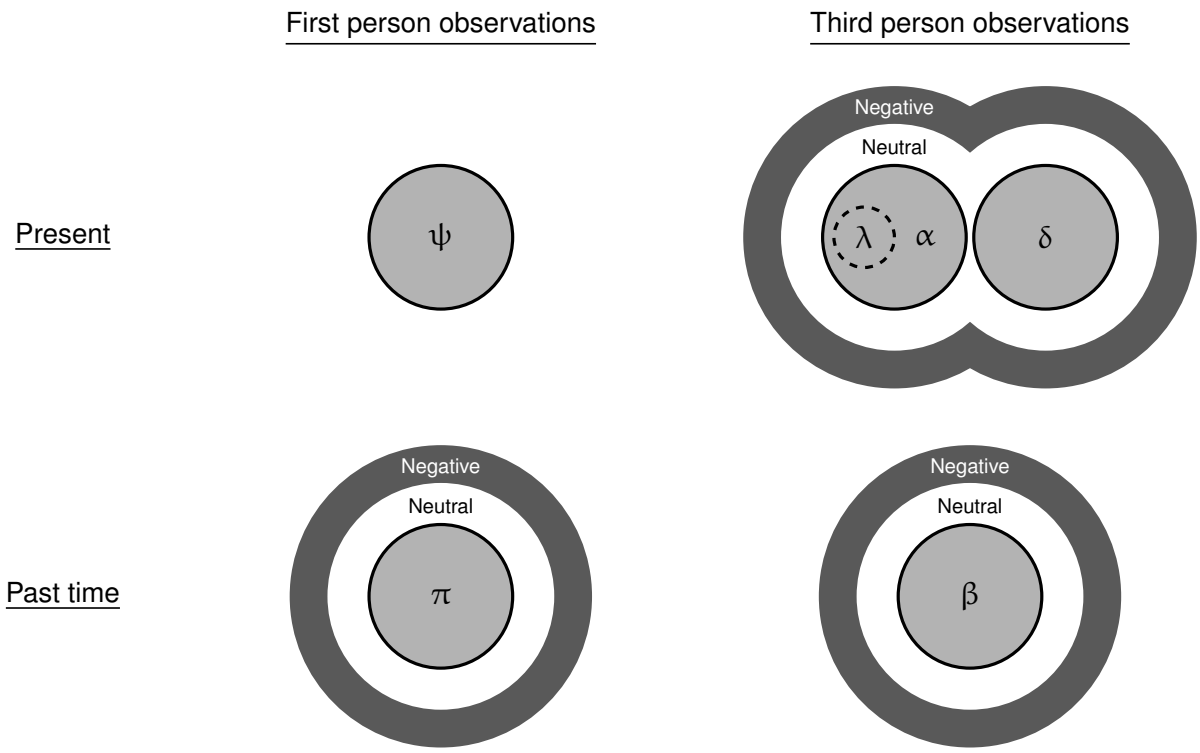

Figure 8. Schematic overview of possible claims about consciousness. The definitions of the different labels are given in the main text on pages 22–25.

1. Both ψ- and π-positive claims are by nature subjective and will only stay ψ- or π-positive to the persons making them. Other people will, if these claims are explicitly communicated, not see these statements as first person statements since they themselves will take part of them only in a third person fashion, hence stripping the claims of their original subjective nature.

2. By definition, ψ-positive claims are always true. That is, if the entity making the claim is not conscious, it would not be a ψ-positive claim but merely an α-positive claim observed by others. This means that an unconscious entity cannot make ψ-positive claims. A π-positive claim, however, can vary in its truth value, but the entity making it has to at least be conscious at the present time, otherwise it would not be a first person observation.

3. Neutral claims do not have any truth values since they do not say anything about the world but merely expresses an uncertainty to what the correct answer is. As soon as one answer is favored, the claim is not neutral anymore.

Categorization assignment of claims within the IITC. When it comes to the IITC, the decision of in which category to place a certain statement within the postulated territory will have to be pragmatic. Technically, since the IITC is a panpsychic theory, every system that has some kind of activity is coupled with a positive claim of consciousness. However, for practical concerns, one

could set an arbitrary low point of consciousness, let everything below it constitute a negative claim and let every case where one is unsure of whether to place it above or below constitute a neutral claim. For the rest of this thesis, when referring to the fact that a certain system state is asserted to have no consciousness within the IITC, this then also includes the case where the system has a minimal amount of consciousness. This goes well together with Tononi's (2007) own definition of consciousness, presented on pages 19–20, in which he states that even though consciousness never really vanishes all together, it can be so small that one shouldn't care about it.

When questioning the justification of the negative claims within the IITC, something that will be done in this thesis, one then really questions how the minimal amount of consciousness that is postulated can be justified, rather than questioning any complete lack of consciousness. The motivation to even so use the term "negative" is to harmonize the discussion with other scientific theories of consciousness where negative claims actually are made (that is, theories that are not panpsychic as the IITC). Also, as will be seen, negative claims within the IITC are postulated because of a lack of any reportable instances or memories of consciousness, and not because of the detection of a minimal amount of it, which rather is theoretically extrapolated *ad hoc* when the theory has been laid out. This makes the term "negative" even more suitable, also for claims within the IITC.

Redefinition and division of consciousness

The reason that an agent comes to believe that there is some kind of problem regarding consciousness that also needs an explanation can either be internally or externally motivated. If it is internally motivated, it is because the agent herself comes with ψ-positive claims. If it is externally motivated, it is because the agent notices that other agents are making λ-positive claims. In this way, an agent without consciousness herself could still come to pose questions about consciousness.

The type of consciousness that will be investigated in this thesis is the internally motivated one. As such, it cannot be given a definition unless the agent already is in possession of consciousness. That is, to explain to an unconscious agent what consciousness is, is about as fruitless as trying to explain what sight feels like to a person who has been blind all her life. However, if the person *is* conscious, consciousness is easy to define for her. It is simply, but not limited to, everything she thinks, feels and comes in to contact with. That is, from her point of view, she has never seen, felt or experienced something that has not also been consciousness. It might be the case that, for example, other people actually exist (rather than just being mental concepts in her own head) and also are conscious, but this scenario is empirically indistinguishable from a scenario where the only thing that exists is her own consciousness. Consciousness is in this sense contrasted with the extrapolated existence of the physical, which normally is thought to be something non-sentient.[13]

[13] Something that, for example, panpsychists does not agree with.

Block (2005) has made a distinction between phenomenal consciousness and access consciousness, and subsequently between the phenomenal neural correlate of consciousness (NCC) and the access NCC, a distinction that has been used a lot in the scientific literature about consciousness. Phenomenal consciousness is defined as that which differs between, for example, experiencing something red and experiencing something green. Access consciousness, on the other hand, is defined as the conscious content information that is broadcasted, sort to speak, to large parts of the brain and is cognitively processed on a high, reportable level. A similar distinction, that between primary consciousness and higher-order consciousness, has been made by Edelman (2003). Here, primary consciousness is defined as the most rudimentary form of consciousness, something that arises even in the lower animals when perception and motor events gets integrated with memory. Higher-order consciousness, on the other hand, ostensibly arises when an organism is able to remember its history, form plans and be conscious of being conscious.

The concepts presented above will not be used in this thesis out of two reasons: (a) they are deemed to not be exhaustive enough, and (b) the way they already have been used in the literature has the consequence that they come with some inherent metaphysical and epistemological connotations that might be hard to shake off. At this point, to keep the names of the concepts but further divide them and redefine their meaning would be more confusing than clarifying. Therefore, a new set of concepts will be used in this thesis.

The main distinction will be that between qualia-consciousness (*q*-consciousness) and knowledge-consciousness (*k*-consciousness). The latter concept can further be divided up into reported knowledge-consciousness (*rek*-consciousness), potentially reported knowledge-consciousness (*prek*-consciousness) and non-reportable knowledge consciousness (*nrk*-consciousness).

Here, *qualia-consciousness* is referring to any mental experience, no matter what the content of that experience is. This could, for example, be something so minimal as a mental experience of seeing redness and nothing more, or something so content rich as a mental experience of reasoning about the nature of consciousness and reflecting on the experience of having experiences. If there is a positive answer to the question if there is something like to be a certain system state (cf. Nagel, 1974), then there also is *q*-consciousness.

To have *knowledge-consciousness*, on the other hand, requires a form of meta-experience, namely an experience of having knowledge of having an experience. This would mean that it would not only be enough to have an experience of seeing red, but one would also have to know this to actually have *k*-consciousness. In this way, *k*-consciousness is a subset of *q*-consciousness.

As stated above, *k*-consciousness can be further divided up into three subcategories: *Reported k-consciousness* refers to *k*-consciousness that actually is reported, either in a internal or an external fashion. This can be exemplified by a trivial case when somebody sees something red and states:

"I see something red". However, even if the person does not verbally report this, she can still internally detect this fact, something that also would count as something *rek*-conscious. *Potentially reported k-consciousness* refers to something either *k*-conscious or just *q*-conscious that can be reported given that sufficient attention is aimed towards it. This also includes states that not only *can* be reported, but actually have been reported, that is, *rek*-consciousness. *Non-reportable k-consciousness* refers to *k*-consciousness that is not reportable and therefore also never is reported. An example of this is total locked-in syndrome (Bauer, Gerstenbrand, & Rumpl, 1979), that is, the condition where a patient is totally unable to communicate with the outside world (even with their eyes as in normal locked-in syndrome) although their cognitive processes seem to work as they should, *if it is the case* that these patients actually are conscious.

More formalized then, the relationship between the different type of consciousness (see Figure 9) can be captured with the set relationships

$$\textit{rek}\text{-consciousness} \subseteq k\text{-consciousness} \subseteq q\text{-consciousness} \tag{17}$$

$$\textit{nrk}\text{-consciousness} \subseteq k\text{-consciousness} \subseteq q\text{-consciousness} \tag{18}$$

$$\textit{prek}\text{-consciousness} \subseteq q\text{-consciousness} \tag{19}$$

$$\textit{nrk}\text{-consciousness} \cap \textit{prek}\text{-consciousness} = \emptyset \tag{20}$$

Also, *raw q-consciousness* is defined as everything that is *q*-consciouss but not *nrk*-conscious or *prek*-conscious, that is

$$\text{raw } q\text{-consciousness} \coloneqq q\text{-consciousness} \backslash (\textit{nrk}\text{-consciousness} \wedge \textit{prek}\text{-consciousness}) \tag{21}$$

Note that all these divisions only are meant to show the theoretical possibilities and offer a useful nomenclature to speak about different types of consciousness. By acknowledging this logical partition, nothing is said about the actual size and constitution of the different subsets (as is indicated by the fact that the $\subseteq$ symbol is used rather than the $\subset$ symbol). For example, the nomenclature allows that the set of *nrk*-consciousness is empty and/or that *k*-consciousness fully overlaps *q*-consciousness and therefore not is a *proper* subset of it. For the IITC, all these different subsets are possible and indeed also real.

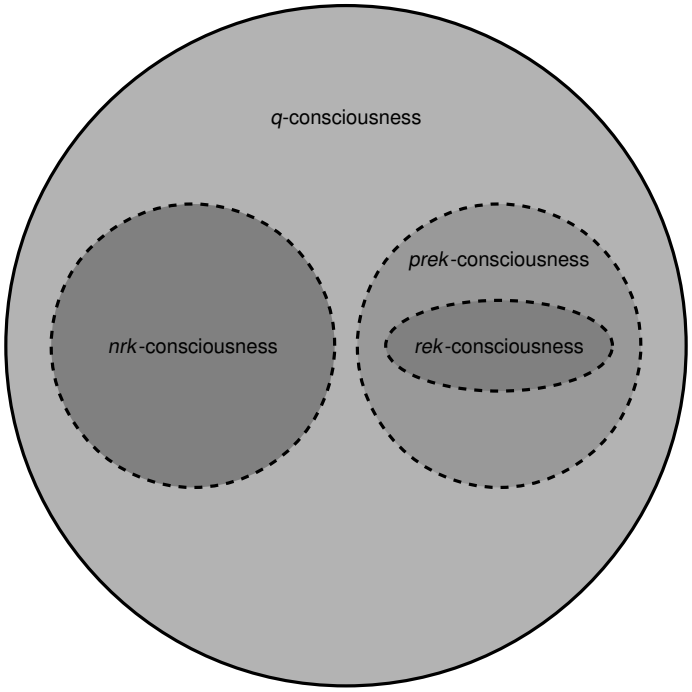

Figure 9. Relationship between *q*-consciousness and the three different categories of *k*-consciousness: *rek*-consciousness, *prek*-consciousness and *nrk*-consciousness. Illustrated is the fact that (a) *k*-consciousness is a subset of *q*-consciousness, that (b) *rek*-consciousness is a subset of *prek*-consciousness and that (c) the intersection of *nrk*-consciousness and *prek*-consciousness is empty. Note that in the figure, the size of each set is expanded so that they never fully overlap with their supersets or are empty. This, however, as is described in the text on the previous page, is not a *necessary* consequence of the nomenclature itself.

Possible empirical reasons for excluding PT^C

Foundational negative claims for the IITC. Some third person negative claims of consciousness lie as a foundation for the IITC in the way that they, together with a set of third person positive claims, strongly specify the topography of the postulated territory that the generated territory of the IITC is supposed to fit. These claims take the form of examples of states of the human brain that do not generate any (or rather, as discussed on pages 26–27, only an insignificant amount of) conscious experiences along with examples of systems that do not generate any conscious experiences regardless of what state they are in. Some of these statements are listed below, all taken from Balduzzi and Tononi (2008):

1. "We know that consciousness fades during sleep early in the night, although neurons in the thalamocortical system remain just as active as during quiet wakefulness."

2. "Consciousness is typically reduced when neuronal activity in the human brain is severely depressed, as under deep anesthesia or in certain comatose states."

3. "Consciousness [. . .] lapses when neuronal activity is excessive and hypersynchronous, as is the case in generalized seizures."

4. "We know that certain regions of the brain, for example the thalamocortical system, are essential for consciousness, whereas other regions, such as the cerebellum, are not, though the cerebellum has even more neurons and is seemingly just as complicated."

5. ". . . yet few would argue that the camera [described on pages 10–11 and in Figure 4] is conscious."

Statements 1, 2 and 3 above are examples of systems, which during many of their possible states are ascribed consciousness, are deprived of it during certain circumstances. Statements 4 and 5, on the other hand, are examples of systems that under no circumstances are ascribed consciousness; one of these systems, the cerebellum, having a great amount of causal influence on a system that is ascribed consciousness, the thalamocortical system, and one of these systems, the camera, being an isolated instance of something lacking consciousness.

In the postulated territory that lie as ground for the IITC, statements like 1-5 are associated with negative claims regarding consciousness because of their lack of overt behavior reporting consciousness. This is, for example, evident from the following quote by Massimini et al. (2005) in an article tightly connected to the IITC:

> The fading of consciousness during NREM sleep episodes early in the night, *evidenced* [emphasis added] by short or blank reports of cognitive activity upon awakening (Stickgold, Malia, Fosse, Propper, & Hobson, 2001), would then be associated with an impairment of cortical effective connectivity.

Here, the proof that there is no, or very little, consciousness during NREM-sleep comes from the fact that the person in question does not exert any kind of behavior indicating consciousness during this state; nor does she exert any significant amount of (if any) behavior reporting some kind of memory of a conscious episode after she has woken up, which could lie as ground for an β-positive claim (and in turn a δ-positive claim).

Epistemic correlationism. Normally, when there is, from a system, a lack of overt behavior signaling consciousness (a) during or (b) after it has been in a certain state, and if there are no cases of similar systems being in similar states exerting any of these kinds of behaviors, there are no grounds for making either α-, β- or δ-positive claims about that particular system state. However, one could question if this also warrants making negative, and not just neutral, claims about it. In Block's (2007) words, this view is called *epistemic correlationism*. In sum, this is the view that the question regarding if a certain state generates any consciousness or not is scientifically

unaddressable in the cases where a clear α-, δ- or retrospectively β-positive claim cannot be made. That is, global cognitive access, something that is necessary for making α- and β-positive claims (which in turn are necessary for inferring δ-positive claims) regarding a state, limits our knowledge of our conscious states and we cannot know which states, not accessed by this process, that are conscious.[14]

One could pose the question what observations one would be making in a world where, for example, all the system states described in statements 1-5 above generated consciousness to a large extent. In the case of statements 1-3; just because these states would be consciously experienced would not automatically mean that the person having these experiences also would be able to cognitively access and, subsequently, form high-level concepts about them and be able to report about them, either at the time of having the experiences or afterwards. In this scenario, where there was consciousness but no possibility of cognitive access to it, the experiences would be constituted of either (a) raw *q*-conscious states, (b) *nrk*-conscious states or (c) a combination of both. The same thing would be the case for systems with no ability for overt reporting behavior at all, like the ones in statements 4 and 5. In these scenarios, even if the cerebellum or the camera would be conscious, they would have no means of expressing this fact to others.

This discussion can be related to what Bostrom (2002, p. 1) labels observation bias in connection to his discussion on what he calls the *anthropic bias*:

> How big is the smallest fish in the pond? You catch one hundred fishes, all of which are greater than six inches. Does this evidence support the hypothesis that no fish in the pond is much less than six inches long? Not if your net can't catch smaller fish.

This very scenario can be transferred to the situation discussed above. If a system state takes place and no consciousness is reported from it, does this evidence support the hypothesis that there in fact was no consciousness coupled with that particular system state? Not if it would not be able to report about the consciousness, or in some other way signal its existence, even if it were there.

As another analogy, imagine that the SETI-project (http://www.seti.org/), a project with the goal of finding extra terrestrial life by analyzing electromagnetic signals from outer space for some kind of intelligible patterns, all of sudden would come into contact with a plethora of alien civilization. Further, imagine that the researchers of the project very soon would come to the general conclusion that on all planets in the universe where there existed life (not only the ones they had detected), radio technology had most likely also been discovered. This would clearly be a logical fallacy since a necessary condition for an alien civilization to be discovered in the first place would be that it had acquired radio broadcasting technology of some sort. The radio technology would

[14] As will be seen later, Block does not subscribe to this view himself.

then be a necessary condition for the *detection* of extra terrestrial life but not necessarily for the *existence* of it.

Consciousness during dreamless sleep. As a concrete example of the discussion above, the reasons for believing that no consciousness is present during so called dreamless sleep (represented by statement number one in the negative claims list for the IITC on pages 30–31) will be investigated. Similar analyses could be done for other conditions, for example comatose (statement number two in the list) or hypersynchronous (statement number three in the list) states, or other systems, for example the cerebellum (statement number four in the list) or an idealized digital camera (statement number five in the list). However, there is no need to go through each and every possible example since the argument would take on a more or less equal approach for each one of them, thereby making one concrete example sufficient.

Imagine having a certain machine, m, which consists of a couple of separate modules communicating with each other (see Figure 10 for a complete description of these modules). At time t, m measures the existence of one of its modules, the intelligence module, the very same module that later can look at the register of the machine and make inferences about the data. If the machine can find the intelligence module, it places the current time, t, in a register. If it cannot find it, it does nothing. This operation is then reiterated with a certain time interval, k.

Now, imagine that this machine is up and running and that we, at some point in time, a, disconnects the cord between the detector and the register module, leaving everything else in place. We then connect the two modules again at some time point, b, and make sure that $b - a > k$. Now, three questions, along with three answers, arises:

(Question 1): Will the register of m have registered the existence of the intelligence module between a and b?
(Answer 1): No.

(Question 2): Is [Question 1 $\wedge$ Answer 1] evidence for an observer looking at the register of m that the intelligence module of m did not exist between a and b?
(Answer 2): No, since the observer cannot discern between the case where the intelligence module of m actually did not exist and the case where m just could not register it due to some malfunctioning.

(Question 3): Is [Question 1 $\wedge$ Answer 1] evidence for the intelligence module of m that it did not exist between a and b?
(Answer 3): No, since the intelligence module of m cannot discern between the case where it actually did not exist and the case where m just could not detect and record it due to some malfunctioning.

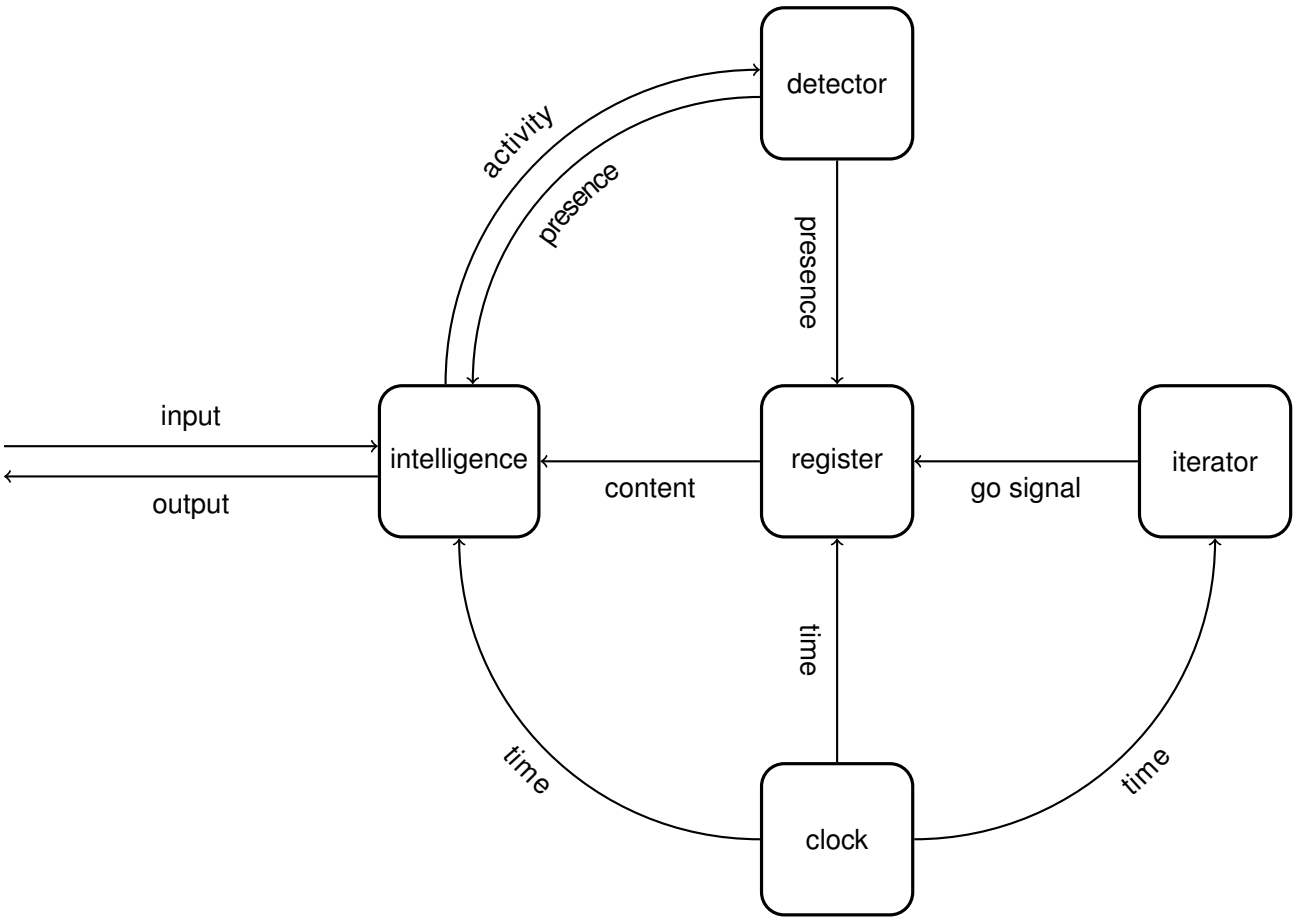

Figure 10. Individual components and flow chart of data transfers within a machine, m, which measures the existence of itself. *The clock module* continuously outputs the result of an internal counter. *The iterator module* is set by a certain internal parameter, k, to send out a go signal every k^{th} time step based on time data it receives. *The register module* is capable of writing data to an array. If the register simultaneously receives a go signal and a presence signal (see below), it writes the current time to an entry in the array. *The intelligence module* can look at the content of an array from the register, or from activity from the detector, and come with inferences about it. It can then output these interferences to the outside world. It can also be fed instructions from the outside world what to do. *The detector module* can decide whether activity data comes from an intelligence module or not and then output the answer. What is determined as activity from an intelligence module, and therefore what kind of activity a module has to produce to be defined as an intelligence module, is decided by a specific algorithm operating within the detector module.

Note that if the time measuring mechanism would stop between *a* and *b* and then resume its activity afterwards again, there would be nothing in the register of *m* that would indicate that anything unusual had taken place at all. As of now, the intelligence module can deduce that something has happened but not exactly what.

Now, if a question were to be put to the intelligence module sometime between *a* and *b* regarding its current existence status, the answer from it would be positive. That is, at every time point between *a* and *b*, the intelligence module, in conjunction with the detector module, would recognize its own existence even though it would not be able to confirm the existence at that point afterwards. Therefore, the question regarding the existence of the intelligence module between *a* and *b* would be trivial for an observer who took part of existence statements from *m* during that time period. The conclusion would be that the intelligence module had existed all the time but that the internal registration of this fact within the register module was malfunctioning. The intelligence module would have had the equivalence of *k*-consciousness at every instance but never any memory of it.

If this scenario is going to be likened to a scenario with a human being, it would be a scenario where the brain functioned as normal with the exception that the memory system was malfunctioning in its encoding. That is, at every time point, the person would insist that she was conscious, but she would not be able to remember her own conscious history up to that point in time. A less extreme form of this, where memory of past states can be retained for some limited time, can be seen in patients with anterograd amnesia (Scoville & Milner, 1957; Corkin, 2002).

To shift the example closer to that of sleep, imagine that in addition to the temporary disconnection of the connection between the detector and the register module, the intelligence module itself would be altered in some specific way that *at least* hindered it from outputting possible interferences about itself between *a* and *b*. That is, even if interferences could be made by it during the altered state, these would not make it to outside observers. In other words, the intelligence module *could* have the equivalence of *k*-consciousness but would not, in that case, be able to communicate this fact to others (in this case it would have *nrk*-consciousness).

For the intelligence module itself, the problem of its own existence between *a* and *b* would, in the situation above, still persist. This would also be true for an outside observer, only fed with information from the output of the intelligence module between the time points *a* and *b*, since this output would not contain any existence statement. However, an observer looking at *m* itself during the whole time would still be able to have full epistemological control of what happened at every stage. This is where the analogy between the *m*'s detection of itself and consciousness in humans breaks down. What counts as an intelligence module will have to be arbitrarily set and the definition will be completely transparent for someone looking at the detection module. This is not the case with consciousness that is, as Edelman and Tononi (2000) states in the basic assumptions

of the IITC (listed on page 40), subjective and private. For the analogy to hold, we would have to see *m* itself as a black box, where one only was being allowed to observe the outputs from the system, in essence making the inferences coming from the intelligence module the bottleneck of our epistemological access (as will be discussed on page 50, this does not hold if one subscribes to the radical eliminitavistic position).

Now, one could pose the question in what way this scenario is different from the scenario with a certain human, *h*, instead of *m*, failing to make an α-positive claim during, or β-positive claim after, sleep regarding some time point when she, for example, was in so called dreamless sleep.

Humans roughly have three different systems for keeping track of the passage of time: millisecond-, interval- and circadian timing, each one involving different neural mechanisms (Buhusi & Meck, 2005). At least a subset, made up of some combination of parts of these systems, are active during sleep, evident by the fact that humans are able to, reasonably well and without external cues, guess the amount of time they have been sleeping (Aritake-Okada et al., 2009) as well as the ability of some people to control, through a conscious decision before going to sleep, when to wake up (Moorcroft, Kayser, & J, 1997). Further evidence of this comes from people suffering from paradoxical insomnia, a condition in which these time telling systems malfunctions during sleep (Mercer, Bootzin, & Lack, 2002; Means, 2003; Pinto et al., 2009; Vanable, Aikens, Tadimeti, Caruana-Montaldo, & Mendelson, 2000; Edinger & Fins, 1995) but not while awake (Rioux, Tremblay, & Bastien, 2006; Tang & Harvey, 2005). This makes the person believe that she took longer time to go to sleep and spent shorter time actually sleeping, even though she has the same sleep patterns as normal, healthy controls, and even though she has no problems with the time telling systems while awake.

While asleep, the brain goes in and out of dramatically different, compared to while being awake, functional activation patterns, which characterizes the different sleep stages, as evident from recording with imaging techniques such as EEG (Silber et al., 2007), PET (Braun et al., 1997) and fMRI (Lovblad et al., 1999; Kaufmann et al., 2007). Among many things, the frontal lobes are affected so that they work in an abnormal way, especially during slow wave sleep. This area of the brain is important for high level inferences of the type that are reflected in the content of *k*-consciousness, for example through its salient role in working memory, something that is necessary for these cognitive processes to function in the first place (Smith & Jonides, 1999).

Further, an important part of the function of sleep is its role in memory consolidation (Diekelmann & Born, 2010) where the neurons of the brain, dramatically altered in the way they work by the change of the neurochemical enviroment, modify different neuronal patterns that codes for memories acquired earlier (Tononi & Cirelli, 2006). Also, as already described on page 13, during slow wave sleep, local perturbations of the cortex produces local and stereotypic, or global but

nondiscriminatory effects, compared to the global and discriminatory effects seen in REM-sleep and during wakefulness (Massimini et al., 2005, 2007).

Now, without making an explicit argument about exactly how this would come about, it would not seem especially far fetched to suggest that the dramatically different way that the brain operates during sleep could have consequences for its ability to cognitively access, register and report possible conscious states. That is, during so-called dreamless sleep, there *could* be some form of *nrk*-consciousness and/or raw *q*-consciousness without this fact being cognitively detected.

Conflation between consciousness and cognitive functions. The argument given above then states that if someone does not report any form of consciousness, either in the full or in the partial context, at or after a certain point in time, this does not also automatically mean that there was not consciousness present. It could simply have been the case that there was no cognitive function being able to capture and subsequently report that consciousness.

Lamme (2006) has argued that this conflation between consciousness and cognitive functions can be churned out through combining a number of neuroimaging studies. First, he argues that there are clear cases of when people are conscious and not conscious with accompanying neural data. In these cases, where the person in question is not conscious of, for example, a stimuli being presented, there is no confusion between consciousness and cognitive functions. Lamme states:

> There is no need to doubt the presence of conscious experience when a subject reports a clearly visible event, *nor its absence in cases such as blindsight or deeply masked stimuli* [emphasis added] [...]. In those cases, arguments about the conflation of conscious experience with other cognitive functions do not hold up. It is the middle ground where additional, neural arguments become of value [...].

Lamme goes on and further describes the phenomena of blindsight (Azzopardi, Fallah, Gross, & Rodman, 2003; Stoerig & Cowey, 1997; Goebel, Muckli, Zanella, Singer, & Stoerig, 2001). This is a condition where patients are unable to report stimulus in a certain part of their visual field. However, when forced to guess the stimuli in question, or carry out some kind of behavioral task related to what they were shown, they manage to perform above chance, which is evidence that their brain, at least in some rudimentary way, picks up and processes certain information of the stimulus to the point where behavior can be guided by it.

Lamme (2006) gives two arguments to why it should be believed that these patients do not have any visual experience although they are able to utilize information based on ostensibly, unseen visual cues: First, the stimulus never spontaneously inflict on behavior, only when forced in laboratory settings. Second, there is no way to manipulate the situation so that the patients ever report seeing these stimulus. This is not the case when it comes to such phenomena as, for example, inattentional (Rock & Mack, 2000) and change blindness (Wolfe, 1999) in which the subject

fails to see the introduction or change of a certain stimuli because attention is occupied elsewhere. Here, if the stimuli is salient enough, for example if the subject's name is pronounced, it can still get through and be reported showing that it is not inaccessible, only currently not accessed. This is, however, not the case for blindsight patients who, no matter what, cannot report seeing the stimuli.

It is hard to see how these arguments ultimately should be relevant regarding the question if the processing of the visual input in these patients generates any consciousness or not. To be successful in the endevour of refuting this possibility, Lamme would have to suggest that the human brain is capable of detecting all possible conscious states and then transform the information within these states into *rek*-consciousness. That is, if there is consciousness somewhere, the cognitive apparatus will have *some way* (even if this way is very hard to find and is not utilized most of the time) of accessing and reporting it. If the cognitive machinery cannot report a certain state, the state subsequently is not generating any consciousness. However, as has been argued for, this cannot automatically be assumed to be the case.

Block (2007) has expanded this argument. He refers to Fodor (1983) who has argued that early parts of the perceptual system is modular in the way that they constitute isolated entities, which are not accessible in a way that would allow for the person having them to report what was going on inside of them. For example, it can be shown that certain features of the visual input are computed in the brain, but the results of these computations are not part of what eventually is reported by the person herself (Nakayama, He, & Shimojo, 1995). Instead, they are part of lower-level deductions that eventually end up in the higher-level concepts that constitute the accessible material. Block poses the question:

> Are the unreportable representations inside these modules phenomenally conscious? Presumably there is a fact of the matter. But since these representations are cognitively inaccessible and therefore utterly unreportable, how could we know whether they are conscious or not? It may seem that the appropriate methodology is clear in principle even if very difficult in practice: Determine the natural kind (Putnam, 1975; Quine, 1969) that constitutes the neural basis of phenomenal consciousness in completely clear cases—cases in which subjects are completely confident about their phenomenally conscious states and there is no reason to doubt their authority—and then determine whether those neural natural kinds exist inside Fodorian modules. If they do, there are conscious within-module representations; if they don't, there are not.

Sure enough, one could find clear cases of system states where consciousness could be highly inter-subjectively acknowledged, but the problem of separating consciousness from the registering mechanism, when deciding upon the natural kind, still persists if (a) there is an actual possibility that there exists non-accessible conscious states (something that will be discussed in the section

titled "Possible theoretical reasons for excluding PT^C"), and (b) it also is highly unclear whether these non-accessible conscious states also exists (something that will be discussed in the section titled "Possible probabilistic reasons for excluding PT^C"). That is, it does not suffice to have clear cases where consciousness does exist. One also has to have cases where consciousness can be shown *not* to exist.

If making an analogy to the philosophical debate of meaning, which Block (2007) drags in when he refers to Putnam (1975); to be able to decide what constitutes the natural kind of a concept, the scope of the concept needs to be limited. That is, one needs examples of things that are not included in the concept to be able to get the investigation started.

Continuous decrease of observation ability. Now, in the preceding discussion, emphasis has been on system states that, for the postulated territory of the IITC, are coupled with negative claims in the full experience context, that is, where the assertion is that there is no consciousness present at all. However, the postulated, and subsequently the generated, territory of the IITC does not operate with a binary conception of ϕ. Rather, between system states that are claimed to be highly conscious and system states that are claimed to generate no consciousness at all, the IITC predicts somewhat of a continuous decrease of ϕ, so that while almost asleep, one only experiences a fraction of the consciousness one has while fully awake. However, this scenario can be contested using the same arguments that were used to refute the negative claims in the full experience context. That is, when a human being is in a state between wake and sleep, there *will* be, by definition, some kind of alteration in her brain state, an alteration that makes some internal states harder or even impossible for her to detect, report and remember. That is, if the person, while almost asleep, would be able to report on having clear, multifaceted thoughts, just as when being awake (something that would have to correspond to some kind of awake like state in the brain for these reports to be made), we would by definition say that she was just that; awake. Therefore, from an empirical standpoint, the assertion that consciousness gradually decreases between the state of being awake and that of being asleep, runs into the same type of problems as claims that a certain state does not have any consciousness at all.

Summary. In this section, it has been argued that there exist at least some territories that are members of PT^C such that, if they would be true, they would not be empirically discernible from scenarios where a territory being a member of PT would be true. From this, it follows that PT^C cannot be excluded on purely empirical grounds. In the next section, possible reasons for excluding PT^C on theoretical grounds will instead be investigated.

Possible theoretical reasons for excluding PT^C

After having investigated if there are any purely empirical reasons to exclude PT^C, the next logical step is to see if there are any theoretical reasons to exclude PT^C. If it can be shown that the philosophical position that the IITC is built upon

1. Necessitates a territory that looks like some element of *PT*.

2. Is a coherent and working theory of consciousness.

3. Is the simplest explanation of consciousness that satisfies condition number two.

then *PT* should also be accepted, which in turn means that PT^C should be excluded.

The reason for the first demand should be straightforward. The reason for the second and third demand can be summarized by Einstein's (1934) often paraphrased statement saying "It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience." That is, the simplest theory should be accepted as long as it does not leave out any crucial part of the phenomena at hand.[15]

The first task of this section will then be to specify exactly what philosophical position that lies as ground for the IITC, something that is not clearly spelled out in the literature. After that, an investigation of how well demands 1-3 above confirm to this position will be undertaken. After this, a brief investigation of other possible philosophical positions, along with whether they fit with the IITC and whether they confirm to demands 1-3, will be gone through.

The philosophical position of the IITC. In their book "A Universe of Consciousness" (2000), a work released before the development of the IITC into its more elaborate mathematical form, Edelman and Tononi lists three basic assumptions (pp. 14-15) that lie as ground for their subsequent theory building. Since no refutation or revision of these assumptions have been made since, they are assumed to still hold for the IITC in its current form. The three assumptions are:

1. Only conventional physical processes are necessary for a satisfactory explanation of consciousness.

2. Consciousness has evolved through natural selection.

[15]The validity of this claim will briefly be contested on pages 50–52. Although it will be argued that from an ultra-skeptic standpoint, it might not be sound, it will even so be accepted in order to show that the general argument in this thesis also holds when using the assumptions of everyday scientific practice. Also, notice that, if anything, a refutation of demand number three for the current theoretical investigation only would open up the flood gates for other, more complicated theories with no criteria of deciding between them, making it even harder for PT^C to be excluded..

3. Qualia is by nature subjective and private, and no scientific *description* of it will be identical with the *experience* of it.

The physic assumption rules out alternatives that assume something more than the physical, for example dualism. The evolutionary assumption forces consciousness to, in some way, influence the behavior of the organism that possesses it because that is the only way it could be selected for in the first place. If consciousness is seen as a non-effcecious byproduct of computation, like for example according to epiphenomenalism, evolutionary processes will not be able to account for why it also arose since it would have no effect on the survival of the organism. The qualia assumption emphasizes that a theory of consciousness is not the same thing as consciousness itself, and that knowledge of the first does not give knowledge about the second. As an analogy, Edelman and Tononi (p. 12) points to the fact that a scientific description of a hurricane, how it forms and behaves, is not the same thing as the hurricane itself (cf. "The map is not the territory", Korzybski, 1933, and "Ceci n'est pas une pipe", Magritte, 1928).

As a clarification to this last point: Jackson (1982, 1986) has proposed a thought experiment involving Mary, a scientist who specializes in the field of color perception but who is forced to observe the world on a black and white monitor from within a black and white room. Mary learns everything there is to know about wavelengths of light, neuronal responses in the brain and behaviors of people seeing different colors. That is, she knows everything there is to know about color vision from a physical standpoint. Jackson then poses the question whether Mary would acquire some new knowledge about color experience if she was let out of the room to actually see color for herself. The qualia assumption above states that Mary indeed will gain some new knowledge during these circumstances since the description of color vision that she holds is not the same thing as the color experience itself.

The first assumption above makes it clear that the IITC is a theory built upon a physicalistic philosophical position. More specifically, it is built upon a nonreductive physicalisitic position (a position that will be further explained in the very next section), as is evident from the following passage from Balduzzi and Tononi (2009):

> according to the [IITC], the "redness" of red, and similarly any qualitative aspect of experience, is not specified by the firing of particular neurons, nor by particular patterns of activity or correlations, nor is it a property of certain anatomical circuits, but it exists only at the level of the set of information relationships generated by a complex of elements in a certain state. Specifically, the "redness" of red, and similarly any qualitative aspect of experience, corresponds to a specific q-fold within a quale, generated by the activation of a set of specialized mechanisms. As such, it exists only in the context of the quale, just like a particular convexity in a complex solid only exists in the context of

the solid. This perspective also implies that specific qualities of consciousness, while generated by a local mechanism, cannot be reduced to it.

Nonreductive physicalism. Nonreductive physicalism is a philosophical theory of mind that states that while the mental is something real, it is fully instantiated by purely physical events, although it cannot be reduced to these. Hofstadter (2007, p. 37-41) has made an illustrative point of how nonreductionism is thought to work: Imagine a huge domino lineup where each brick is spring loaded so that it can flip up again after a certain refractory period after it has been knocked down. If the dominos are placed in a certain pattern, and the refractory period for them are adjusted to a certain value, the whole set-up can instantiate, for example, a computational process that answers the question whether a certain number is a prime or not. For example, a certain row of dominos could fall if a devisor was found to the input number, indicating that it was not a prime. Once this happened, someone observing the process could point to one of the dominos within this row and ask why that specific domino fell. Hofstadter claims that although the answer that that specific domino fell because the one preceding it fell is correct, it is far to myopic and does not capture the whole truth. Instead, he suggests, the answer to why that specific domino fell is because the input number was not a prime. He goes on to say that what is interesting with this explanation is that it is not physical in its nature, but that it instead is an abstract, global, math-level answer. If this indeed would be the actual answer to the question, it would be an example of something nonphysical causing something physical, although this cause, at the same time, was totally instantiated by the physical and therefore not demanded any extra, nonphysical substances.

As already has been noted, the IITC is based on such a nonreductive physicalistic framework, in the way that consciousness is seen as a high-level, abstract notion of a multi-dimensional polytope in the qualia space.

To satisfy demand number one above, the IITC has to generate a territory that looks like some element of *PT*. In this thesis, this will be assumed to be correct. Further, since nonreductive physicalism does not postulate any more substances over and above the physical to explain consciousness, the form of this position that is proposed for the IITC will naïvely have to be seen as a fairly simple theory, with good chances of at least tying with other theories when it comes to demand number three, *if* it first can satisfy demand number two. Demand number two is therefore what will be investigated from hereon.

Kim's critique Kim (1989) has argued that nonreductive physicalism is a non-tenable position. According to him, this is because it assumes two things that are incompatible with each other. First, nonreductive physicalism assumes what Kim calls the (a) *causal closure of the physical domain*. This is the hypothesis that any physical event has a fully sufficient physical cause, rather than nonphysical (notice the similarity with the first basic assumption of the IITC, stated on page 40).

Second, nonreductive physicalism assumes (b) *psychophysical causation*, that is, that physical behavior takes place because of underlying mental states; it is *because* of the subjective feeling, for example of being scared or hungry, that at least some of the organism's repertoire of behaviors will be carried out (notice the similarity with the second basic assumption of the IITC, stated on page 40).

Now, according to Kim, there is a fundamental contradiction going on here. If a certain behavior is caused *both* by a physical and a mental state, it also means that that very behavior is causally overdetermined. That is, given that one grants the causal closure thesis, one can also completely account for the behavior in purely physicalistic terms. In this context, it does not seem like one has to, or indeed even could, add an explanation that it is the mental that causes the behavior since the behavior already has been accounted for. It is like two writers, both claiming to have written a certain book. If the authorship is granted solely to one of them, analogous to the step that is taken with the assumption of the causal closure of the physical domain, it seems like the other one automatically is out of the picture.

Kim (2005, p. 41) states this principle in the following way: "No single event can have more than one sufficient cause occurring at any given time - unless it is a genuine case of causal overdetermination". More formalized, Kim's principle can be stated as (2005, p. 17): "If an event e has a sufficient cause c at t, no event at t distinct from c can be a cause of e (unless this is a genuine case of causal over-determination)." The concept of "overdetermination" here denotes the concept of when a certain event has several different causes at the same time. The standard example of this is a firing squad that, on a certain cue, simultaneously fire at a prisoner who is to be executed. Each bullet that hits the prisoner is in itself enough to kill him. However, since he is hit by several bullets, none of them can individually be ascribed as the bullet that killed him, therefore making his death overdetermined by all the bullets combined.

Kim (1989) enumerates a number of possible solutions that could be proposed to solve this problem, along with his answers to them:

1. One could assert that the physical and the mental are both partial causal explanations and that both are needed to account for the actual behavior. However, this seems to go against the causal closure hypothesis since it maintains that the physical cause is *sufficient* for the current event. To say that something else is needed, in this case the mental, is in direct conflict with this position.

2. One could claim that while the physical cause indeed is sufficient and fully explains why the behavior comes about, the mental could be added on top of this, saying that if the physical cause would not take place, the mental would take care of carrying out the behavior. This would mean that the behavior indeed would be overdetermined but that the mental would

step aside, so to speak, in most cases. However, this would also be in violation of the causal closure hypothesis. By definition, if the mental in some cases can affect the physical, that is, without any physical cause for the event being present, the causal closure hypothesis would not be true.

3. One could simply say that the physical and the mental are the very same thing. That is, some part of the causal explanation of the behavior constitute the actual mental experience, for example activation in a certain way in some part of the brain. However, as Kim points out, this is a step that the nonreductionistic physicalist, by definition, cannot take since what nonreductionism asserts is precisely that the mental cannot be reduced to something purely physical.

Now, according to Kim, there are two different routes one could take to come to a logically consistent position:

1. One could keep faith in (a), the causal closure of the physical domain, but reject (b), psychophysical causations. This could in turn take two different forms: One could either denial that there are any mental events at all, adopting a kind of radical eliminitavism, or one could accept the existence of mental events but claim that they do not affect the physical domain; a form of epiphenomenalism.

2. One could denial (a), the causal closure of the physical domain, but maintain that there are (b), psycophysical causations. This would lead to a form of dualistic interactionism.

Now, if Kim's critique is correct, this would mean that the nonreductive physicalism of the IITC cannot necessitate a territory that looks like some element of *PT*, simply because it is a non-tenable philosophical position to begin with. Therefore, two answers to Kim's critique (stating that overdetermination is a mundane phenomena, and switching the meaning of causation), which are deemed to be the best answers around, will now be brought up along with an evaluation of these.

Overdetermination Strand (2010) has argued that overdetermination not only is more usual than what Kim suggest but that it is ubiquitous in the world and that it therefore could be present also when it comes to mental causation. He gives two concrete and mundane examples of this:

1. Imagine a ball hitting a window that breaks upon contact. Also, assume that the ball was a sufficient cause of the breaking of the window. One could imagine another object that consisted of all the parts of this ball except for one atom. If this other ball hits the window, it would probably break it too, subsequently making this alliterative ball a sufficient cause for the breaking of the window. Therefore, the first ball hitting the window seems to be a

sufficient cause for the breaking of the window, because it consists of some proper part that in and of itself is sufficient for breaking the window, namely the alternative ball. The causal sufficiency is then conserved even though one changes the cause itself, that is, adding and removing one atom from the object.

2. Imagine a table consisting of four legs and a tabletop, located 92 cm above the floor, where the legs are a sufficient cause for keeping the tabletop from falling to the ground. However, any combination of three legs would also do the job, that is, being causally sufficient of keeping the tabletop up. Therefore, it seems to exist four different minimal causes, that is, any combination of three legs, which would be causally sufficient for the effect. These are all instantiated through the table's four legs that means that the effect of keeping the tabletop from falling is overdetermined.

If the examples given above are correct, it would mean that overdetermination could be shown to be an often occurring phenomena in the world and that it therefore could be a likely candidate to solve the problem of joining the causal closure and psychophysical causation hypotheses.

Causation as difference making List and Menzies (2009) approache the problem from another angle and argues that what is wrong with Kim's account is that he is working with a flawed concept of causation. What they themselves suggest is that causation should be understood in terms of difference making. To illustrate this concept, they refer to an example, first stated by Woodward (2008), in which he draws from research by Musallam, Corneil, Greger, Scherberger, and Andersen (2004).

In this research, activity from the parietal reach region in the macaque monkey was recorded, a region that is hypothesized to, within the monkey, encode for higher order plans to reach for a specific object rather than to encode for the specific, low-level, motor commands associated with this intention (Murata, Gallese, Kaseda, & Sakata, 1996; Snyder, Batista, & Andersen, 1997; Batista, 1999). What was recorded by Musallam et al. (2004) was the aggregate firing rates of a number of neurons. These firing rates could subsequently be grouped in a number of categories, each one representing a specific intention to reach for a certain object.

Woodward (2008) points to the fact that trials within these different categories could differ quite a lot with respect to individual neuronal behavior and still be grouped together, that is, there are a lot of different ways for a group of neurons to instantiate a certain aggregate firing rate profile. Formalized, this would mean that a certain intention, i_x, to carry out a certain reaching behavior, r_x, could be instantiated by any element in a certain set, N_{a_x}, consisting of all the different neural firing behaviors (drawn from the population of all the possible spike trains, $\{n_1 \ldots n_n\}$) that would realize a certain aggregate firing profile, a_x, which in turn would correspond to or realize that very

intention i_x.

Woodward emphasizes that there is nothing in the argument depending on the specifically mental character of i_x when it comes to its causal power to realize r_x. From the inference he is making, a_x is just as likely a candidate for the realization of r_x as i_x is. However, he also states that as long as the two statements (a) "i_x causes r_x" and (b) "a_x causes r_x" do not predict different outcomes under any conditions, they are not competing causal claims and therefore a choice between them is not necessary.

Now, List and Menzies (2009) sets up a thought experiment where a monkey has a specific intention, i_1, in this case realized by the specific neural spike train, n_{11}, which leads to a certain reaching behavior, r_1. They then list the following claims (with the $\Box$ symbol being the ordinary modal logic symbol for necessity):

1. monkey has intention i_1 $\Box\!\!\rightarrow$ monkey performs r_1.

2. monkey does not have intention i_1 $\Box\!\!\rightarrow$ monkey does not perform r_1.

Here, both statements 1 and 2 are true. However, the next two more lower level statements are not both true:

3. monkey has neural property n_{11} $\Box\!\!\rightarrow$ monkey performs r_1.

4. monkey does not have neural property n_{11} $\Box\!\!\rightarrow$ monkey does not perform r_1.

Since i_1 could be realized by some other neural firing behavior within the set N_{a_1}, that corresponds to or realizes i_1, statement 4 is false.

List and Menzies (2009) go on to state that a correct view of causation is one where a change of the cause also should change the effect. For example, when dealing with scenarios where the cause and effect can be represented as binary variables, changing the cause from being present to being absent, or the other way around, should also change the effect from being present to being absent, or the other way around. This is clearly not the case when comparing scenarios 3 and 4. The variation in presence of the cause n_{11} does not necessarily change the presence of r_1. However, it *is* the case when comparing scenarios 1 and 2. Here, when the cause i_1 is taken away, the effect r_1 also disappears.

In List and Menzies' model of causation, then, the cause cannot be (a) to specific, thereby missing relevant cases where the cause is another but the effect is the same, nor can it be (b) to broad, thereby encompassing cases where the cause is the same but the effect changes. According to this model then, what is wrong with Kim's critique of nonreductive physicalism is that he grants causation status to events that are not actual causes for the outcomes at hand. Rather, what one should be seeing as the actual causes are high-level, abstract concepts that cannot be found in any one specific physical state.

Ontological status of macro events and macro objects Counter-arguments to Kim's principle, like the ones presented above, hinges on one important assumption about the world, namely that macro events or macro objects, that is, events and object that are aggregated by smaller parts or can be made up of different parts, are ontologically real and not *only* concepts that in some sense *feels* real for humans. In the monkey example given above, a_x, which is instantiated by n_x, is said to be a cause of r_x. However, this can only be true as long as a_x, r_x and indeed even n_x are seen as actual existing events. In the overdetermination examples the same thing goes for the macro-events described in the examples. The breaking of the glass has to be an actual event. The same thing goes for the tabletop being held up.

As already stated, there are a lot of different neural spike trains, realizing a_x, that instantiates a certain reaching behavior, r_x. However, there seems to be no ontological reason for saying that these different spike trains in the end leads to the same thing, namely the reaching behavior r_x, other than that it makes sense from an anthropomorphic and pragmatical point of view. That is, for humans, it is practical to lump these different events together into one single grand event, and to later refer to that one by a single name, since we generally do not care *exactly* how the effect came about on the micro-level, only about the consequences on the macro-level. If we offer a monkey a banana, we will group all instances of the monkey reaching for the banana and grabbing it together since they all are identical in one important aspect; the outcome that we no longer have a banana left. That is, at the macro-level, the level we generally are concerned with, it makes no (apparent) difference whether the monkey utilizes a certain spike train over another to grab the banana; the outcome will, from a pragmatic point of view, still be the same. However, if looking at the system at a micro-level, different neural spike trains will result in different final outcomes, even if the differences between them are extremely subtle. If one then could argue that r_1 does not exist in any ontological sense, the monkey example, presented on pages 45–46, fails. The contrafactuals 1, 2, 3 and 4 will no longer be relevant since they involve at least one ontological unreal variable, that is, r_1.[16]

The same argument is applicable to the breaking of the window. It is pragmatically relevant to group the events where the window shows a certain macro-behavior together since the consequences for the humans involved, for example the need to replace it, will be the same at the macro-level no matter exactly how the window broke. The same thing goes for the tabletop. If it were to fall down, it can no longer be used as a table, meaning that humans will tend to group the instances where it is held up together. Once again, however, from an objective point of view, there will be nothing special about all the instances where the table top is held up other than that they are very similar to each other; more so than between events where the table top is held up and where it

[16]One could go on to question whether things such as for example aggregate firing profiles were real in any ontological sense too, but this would not add anything to the argument since it will be enough to question the ontological status of one of the included variables, in this case r_1, to force a complete revision of claims 1, 2, 3 and 4.

falls down.[17]

This discussion about the ontological status of semantic concepts goes back a long time. One of the earliest references to it can be found in Plutarch (75/n.d) when he talks about a ship where all the parts of it has been replaced:

> The ship wherein Theseus and the youth of Athens returned had thirty oars, and was preserved by the Athenians down even to the time of Demetrius Phalereus, for they took away the old planks as they decayed, putting in new and stronger timber in their place, insomuch that this ship became a standing example among the philosophers, for the logical question of things that grow; one side holding that the ship remained the same, and the other contending that it was not the same.

The question of whether a certain object or event remains the same when all, or a significant amount of, its parts have been exchanged or taken away, presents a problem only if such a thing as a true definition of that object or event, grounded in the ontological reality, exists. As long as no robust, generally accepted, argument is given for why this also should be the case, a position that is dependent upon this fact, something that the nonreductive physicalism is, is not suitable to lie as a foundation for a scientific argument regarding consciousness.

Alternative philosophical positions that would entail PT. Even if the critique above is correct, and the explicitly stated philosophical position of the IITC does not hold up, this does not automatically mean that one can refute the whole theory. That is, if any other philosophical position could replace the nonreductive physicalism of the IITC, the theory could still be valid. For this to happen, the alternative philosophical position would have to be consistent with the IITC as well as to all the demands listed on page 40.

The first step in a search for such a replacement position would be to look for philosophical theories that *necessitate* a territory looking like some element contained in *PT*. That is, the entailment of *PT* has to be a logical consequence of the theory in question, otherwise there can be no *theoretical* reason to also accept *PT* based on it.

In this respect, there seems to exist no alternatives to nonreductive physicalism. Reductive physicalism (e.g Churchland, 1994), dualism—both the epiphenomenalistic (e.g Chalmers, 1997) and the interactionistic (e.g Popper and Eccles, 1984) version—and monistic mental panpsychism[18] (e.g Strawson, 2006) all come in different flavors that makes the actual extent of consciousness unclear. That is, even if one for example would argue for interaction dualism in a successful manner,

[17] It could be that there is no fundamental discrete level of reality, meaning that every conceivable object or event that one could possibly propose would be subject to the same kind of critique presented above. However, so much worse for the non-reductive physicalism position.

[18] Note that "monistic mental panpsychism" is not a generally accepted term that, however, here is being used to in a clear way distinguish it from, for example, the nonreductive panpsychic physicalism that is championed in the IITC.

this would not also mean that the territory of consciousness would be set. It could theoretically range from only being coupled with everything that was *rek*-conscious to encompassing every piece of matter in the universe. Same thing goes for, for example, monistic mental panpsychism of which there are several different territory interpretations (Seager & Allen-Hermanson, 2010). To be clear, one *could* arrive at a specific territory through a theoretical discussion but this would require a much more specific set of arguments than simply arguing for a certain broad category, for example reductive physicalism. As long as no such theoretical argument exists that is generally accepted, a specific territory cannot be deduced either.

One possible exception to this is the position of radical eliminitavism[19] (Rey, 1983, 1988; Dennett, 1991). This is the position that the existence of qualia is only an illusion; something that does not really exists but that people only think exists. In the context of this position, the whole problem of consciousness goes away and the only thing left of it is for science to explain how the brain produces the states that we normally associate with qualia. That is, what needs to be explained is under what circumstances the brain draws and reports these incorrect conclusions about its own state. Once a mechanistic explanation for this is produced, there is nothing left of the consciousness problem at all. In this case, *q*-consciousness fully overlaps with *rek*-consciousness, subsequently meaning that *PT*, albeit in a form not fully in agreement with the IITC, should be accepted.[20]

Hawkins, the main proponent of the Hierarchical Temporary Memory model for how the brain operates (George & Hawkins, 2009), seems to subscribe to this view but he has also problematized his own consciousness, as evident from a recapitulation of a statement he made in a discussion about consciousness (Hawkins & Blakeslee, 2004, p. 131): "Given the way you are talking about consciousness, I have to conclude I am different from you. I do not feel what you are feeling, so maybe I am not a conscious being. I must be a zombie." What he refers to here are philosophical zombies, which denote thought up creatures that functions like ordinary human beings in every shape, way and form, except for the fact that they have no qualia (Kirk, 2011). That is, there is nothing there is like to be a zombie.

Now, it should be said that Hawkins makes this statement in a somewhat provocative manner,

[19]Note that "radical eliminitavism" is not a generally accepted term for the denoted concept that, however, here is being used to denote the most extreme form of eliminitavism among its more moderate versions. The exact phrasing is inspired from Savitt's (1975) distinction between an ontologically conservative and an ontologically radical conceptual change, when it comes to materialistic explanations of consciousness. An *ontologically conservative* conceptual change takes place when one fully can explain a phenomena in terms of concepts from another domain without denying the existence of the phenomena itself. In the case of consciousness, this would be the reductive physicalism position. An *ontological radical* conceptual change, on the other hand, takes place when a phenomena is denied to have any ontological existence all together. In the case of consciousness, this is the type of eliminitavism that is addressed in this thesis. Also note that what is referred to is the form of eliminitavism where the notion of qualia itself is questioned; not the form where only certain, higher-level, mental concepts are addressed (William, 2011).

[20]Although not directly relevant for the IITC in its current form, this discussion about radical eliminitavism is brought up as a base for the more general discussion about scientific theories of consciousness that will follow in the last part of this thesis.

making it unclear to what extent he really believes what he is saying. However, this statement still highlights what the radical eliminitavistic position really entails. That is, the position is coherent and specifies a territory that approximates to *PT* (since only *rek*-consciousness has to be explained), but at the same time, consciousness itself, interpreted in the Nagelean way (Nagel, 1974), really has to be abandoned, instead being replaced with a semantic concept that in the end defines q-consciousness as an empty set. This view is still highly controversial, and most people do not accept the full consequences that it entails. For example, given radical eliminitavism, one would have to accept that there would be no reason for a human being to, for example, prefer pleasure over pain (only an evolutionary predisposition to seek out the former and avoid the latter) since none of these states would generate any actual experience.

Going back to the discussion on pages 35–36 regarding the machine, *m*, it can now be seen why, for radical eliminitavism, one does not have to see *m* as a black box in order for the consciousness analogy to hold. Given that consciousness only denotes a certain verbal concept, not really referring to anything real, there no longer is anything hidden from view when it comes to consciousness. Nothing is subject and private since subjectivity and privacy itself does not exist. This also illuminates why, when qualia *is* accepted, one *does* have to see *m* as a black box in order for the analogy to hold. If qualia is private, and there are states where the qualia is not reported to an outside observer, that also means that if a function in nature is going to assign consciousness to different system states, this function cannot be objectively observable.

Summary. In this section, it has been argued that there only are two philosophical positions that, in and of themselves, might necessitate a territory that looks like some element from *PT*; the nonreductive physicalism that is championed in the IITC and radical eliminitavism. Further, it has been argued that nonreductive physicalism is a non-tenable position and that radical eliminitavism is deeply problematic due to the fact that it denies that there is a problem of consciousness that needs to be explained in the first place. Lastly then, in the next section, possible reasons for excluding PT^C on probabilistic grounds will be investigated.

Possible probabilistic reasons for excluding PT^C

Up to this point, it has been argued that there is a logical possibility that there could exist states that cannot, under any circumstances, be reported (that is, *nrk*-conscious and raw *q*-conscious states) and that does not confirm to *PT*. However, nothing really has been said about the actual likelihood of these states also being real and not just theoretical possibilities. In this section, it will be argued that whether raw *q*-consciousness exists or not is not obvious in any way (whether or not *nrk*-consciousness exists to some larger extent will be gone through in the section "quality of consciousness" on pages 55–57 of this examination section).

It has earlier, in the demands (listed on page 40) for connecting a certain philosophical position of consciousness to *PT*, been assumed that a more simple theory generally should be favored over a more complex. This assumption will now temporally be abandoned, to then later be accepted again, in order to flesh out some important points when it comes to probabilistic reasons for excluding PT^C.

An ultra-skeptic point of view. First, consider Russel's (1952/1997) famous teapot analogy:

> If I were to suggest that between the Earth and Mars there is a china teapot revolving about the sun in an elliptical orbit, nobody would be able to disprove my assertion provided I were careful to add that the teapot is too small to be revealed even by our most powerful telescopes. But if I were to go on to say that, since my assertion cannot be disproved, it is an intolerable presumption on the part of human reason to doubt it, I should rightly be thought to be talking nonsense.

What Russel states here seems valid. To assert that one should believe, or get *carte blanche* to believe something, because it cannot be disproved, is an invalid principle. However, to assert that one has to disbelieve something because it cannot be disproved, is equally invalid. This is fleshed out in another famous analogy, made by Sagan (1997, p. 172). He asks his reader to suppose that he claims to have a fire breathing dragon in his garage. The only catch is that this dragon systematically avoids detection by every conceivable physical test. It is invisible, so it cannot be seen; its fire is heatless so it will not show up on an infrared sensor; its body is incorporeal so spray paint will not stick; and so on.

Sagan then states, when talking about open-minded skepticism, that

> Imagine that, despite none of the tests being successful, you wish to be scrupulously open-minded. So you do not outright reject the notion that there is a fire-breathing dragon in my garage. You merely put it on hold. Present evidence is *strongly against it* [emphasis added], but if a new body of data emerge you're prepared to examine it and see if it convinces you.

Now, when it comes to undetectable dragons, one cannot either support or deny them on a strictly empirical, scientific basis. There is simply no empirical test that could resolve the issue. Therefore, in the absence of evidence for an undetectable dragon, a scenario that resonates perfectly with what one should expect to observe if the hypothesis that it existed was true (and indeed also untrue), one cannot, with the support of the scientific method, proclaim that it does not exist (just as little as one can proclaim that it does exist).

Goodman (1983) has made a similar argument when it comes to the problem of induction. Given that all emeralds that have been observed up until a certain time point, t, have been identified

as green, this generally is seen as inductive proof that all emeralds in fact will continue to be green. However, one could also, given the observations, assert that all emeralds are grue, a concept Goodman defines as something that is green up till some time point, t, and then turns blue. Observers in both these scenarios would make the very same observations, which makes any metaphysical conclusion merely an assumption.

In contrast, theories with clear empirical predictions are subject to scientific scrutiny and therefore also scientific negative claims. For example, the theory that Atlantis once has existed has clear empirical consequences in that, for example, geological traces on the seabed should be available for detection. If one then would set out on the tedious task of mapping out the whole seabed of the earth, and not find any traces of a former island from the era described in the Atlantis myths, one could claim to have fairly strong evidence against the hypothesis.

Now, as noted above, the preceding discussion does not favor simple explanations over more complicated ones. That is, no certain probability distribution regarding the existence of different entities is assumed. In such a climate, it indeed becomes impossible to say something about the existence of Russel's teapot or Sagan's dragon. However, in scientific practice, some assumptions regarding these matters are made.

The exact nature of these assumptions is hard to pinpoint and it is not within the scope of this thesis to come with an encompassing analysis of its constitution. Instead, call whatever principles that approximates to the assumptions that generally are used within scientific practice *SA*. However, just to make *SA* slightly more concrete, two principles that at least should approximate fairly well to principles that most of the time are adhered to are (a) Ockham' razor and (b) the fact that if a random existence claiming statement about the world is made (e.g there exists an invisible slime monster living on the far side of the moon), that statement is most probably false.

In order to harmonize the argument in this thesis with the scientific practice in general, from hereon, *SA* will be accepted again. With that said, the argument below should be fairly robust regardless of *SA*'s exact constitution.

A pragmatic point of view. In light of *SA*, both Russel's teapot and Sagan's dragon can, with extremely high probability, be discounted. In the absence of any specific evidence for believing in any of these claims, the only chance left for them for being true is that they just happen to be true out of pure luck. This is a real but negligible chance and the hypotheses can therefore, for all practical purposes, be seen as false. The important point, however, is that the argument in this thesis, about the uncertainty regarding the status of consciousness when no α-, β- or δ-positive claims can be made, are *not* instances of cases similar to Russel's teapot or Sagan's dragon. That is, even if *SA* is accepted, this does not automatically makes it possible to infer that, most likely, no consciousness is present in these cases.

First of all, consciousness is not an arbitrary made-up entity such as orbital teapots and in-

visible dragons. Rather, consciousness is an entity that already exists, and the question is about its extent and exact constitution rather than about its existence. As already has been discussed, it can be shown that the mechanisms that normally detects and reports the existence of consciousness are not working as usual during for example sleep, when consciousness, according to *PT*, vanishes. This then is a case where one knows why the boarder of the observed territory is drawn where it is; to explain the constitution of the observed territory, one only has to refer to the constitution of the measurement mechanism. When such an explanation is given, it becomes redundant to also invoke change in the phenomena itself to explain the observations.

As an analogy, imagine two people looking at a bird flying through the sky.[21] After a while, the bird disappears behind a house. One of the persons then claims that the bird does not exist anymore because she cannot see it. The other person claims that since he can explain the lack of observational evidence, namely that the house blocks his view, he does not have to postulate that the bird does not exist anymore. Since the bird *did* exist earlier and the lack of current observational evidence can be accounted for, he claims that applying ockham's razor would suggest that the bird still exists and that anyone wanting to arbitrarily change the birds existence status would have the burden of proof.

As another analogy, the edge of the observable universe, that is, the distance to the furthest away lying objects that one *theoretically* could detect, is about 46 billion light-years away from the earth (Lineweaver & Davis, 2005). This is because of two things: (a) the light emitted from possible objects lying further away have not made it to earth yet and therefore cannot be observed, and (b) since the light first was emitted, the universe has expanded, meaning that the objects today lies further away than the 13.7 billion light-years one would expect if only accounting for the age of the universe. In other words, it is known why the boarder of the observed territory regarding the universe is drawn where it is, and this seems to have nothing to do with the *actual* size and constitution of the universe, which rather, based on extrapolations from what can be seen, is assumed to be extremely large, maybe even infinite in size. In addition, the further one peeks out into the observable universe, the less advanced patterns of matter are seen, basically ending with just hot gas at the observable edge. This is because light emitted from objects lying further away from the earth have traveled for a longer time than objects lying closer, and therefore to a larger extent represent what the objects used to look like, rather than what they look like today, which should not significantly differ from other objects in the universe. This then is an example of where the actual territory most probably differs *dramatically* from the observed territory, both regarding the objects that can and the objects that cannot be seen.

[21] I owe this example to Tobias Malm.

Complexity as the answer to consciousness. To make it believable that the shape of the observed territory also approximates to the shape of the actual territory, a conceptual connection between the ability to register and report internal states and consciousness itself, would have to be established. In terms of the IITC, this means that a conceptual connection between integrated information and consciousness would have to be established.

As such, the IITC is not the first theory of consciousness to invoke complexity, as the amount of integrated information basically is[22], to explain when and how consciousness arises. However, there seems to be something deeply unsatisfactory with referring to the fact that brains are complex in order to explain why they also have consciousness. Tononi has briefly touched on this subject in an article together with Koch (Tononi & Koch, 2008):

> Consciousness is usually evaluated by verbal reports, and questions about consciousness ("Did you see any-thing on the screen?") are answered by "looking inside" retrospectively and reporting what one has just experienced. So it is perhaps natural to suggest that consciousness may arise through the ability to reflect on our own perceptions: our brain would form a scene of what it sees, but we would become conscious of it—experience it subjectively—only when we, as a subject of experience, watch that scene from the inside.

As already has been extensively discussed, this type of argument is problematic for a number of reasons, one of them being that it does not account for possible observation bias.

Moody (2003) has questioned the notion that the fact that brains realize complex patterns should make it any less surprising that they also are conscious. He states that:

> There is nothing at all to connect complexity conceptually with consciousness. Even though we observe the correlation of consciousness and complex brains, there is no known conceptual connection between the two. Consciousness is the property of having experiences. What is there about complexity that would link it to *that*?

He then goes on to compare this to the connection between complexity and intelligence. Here, the connection is much more clear. It can logically be inferred, or at least strongly argued for, how a system needs complexity to instantiate intelligence, and how the fact that a system is complex makes it less surprising that it also have intelligence.

Moody suggests that one here, to solve the problem, could be tempted to simply, "in a Humean spirit", accept that "anything can cause anything, if the correlations are hooked up right,

[22] If one were to contest this point, and proclaim that complexity and integrated information were two totally different concepts, one could in the following discussion exchange the concept of "complexity" with the concept of "integrated information" and see that the argument applies as well after the switch as before.

and that's all there is to it". Strawson (2006, p. 65) has called this a "law-like miracle"; that it is a "contradiction in terms, given the standard assumption that the emergence of Y from X entails the 'supervenience' of Y on X". However, as Moody (2003) points out, it is still *possible*. The problem with this move, Moody asserts, is that there does not seem to be any other emergent base properties in nature. Sure enough, a naturalist will have to accept that there are base facts about the world that are merely given, but currently, all such detected brute facts are at the lowest physical level (e.g mass of baryons and speed of light). Moody therefore concludes that

> True brute emergence occurs when there is *nothing* about the base phenomena that explains the emergent phenomenon, beyond mere correlation. This, I believe, we do not see in science, and to introduce it solely to "explain" consciousness would not be a step forward. It is one thing to say that consciousness must be added to the list of brute facts about nature. It is quite another thing to say that it must be added as the first and *only* emergent brute fact. Such a move is, in my view, unacceptably *ad hoc*.

Summary. In this section, it has been argued that there seems to exist no convincing probabilistic arguments for excluding PT^C, even if one operates from a set of basic assumptions of scientific practice. Since one can explain why the observed territory looks the way as it looks, and since there seems to be no conceptual connection between the shape of the observed territory and the shape of the actual territory, the only way left for the two territories to overlap would be due to a remarkable coincidence.

Definition of consciousness revisited

In light of the above investigation and clarification of inherent concepts, the definition of consciousness by Tononi (2007), brought up on pages 19–20, can now be revisited. The essence of that definition is that "consciousness is what fades when we fall into dreamless sleep". It has been argued in this whole critique section that what one really can insist fades when we fall into dreamless sleep is the amount of *prek*-consciousness, not q-consciousness itself, which could or could not also be fading. Therefore, if one only looks at this definition, the IITC does in fact seem to describe the presence of consciousness in a system. However, if the problem of consciousness is seen as the problem of *prek*-consciousness, a phenomena that presupposes q-consciousness, then it has little to do with the traditional problem of consciousness that the IITC obviously is created to solve. If q-consciousness is presupposed, it is indeed an important task to find out when it also takes the form of *prek*-consciousness, but the solution to that problem is automatically not the same thing as the solution to how or when q-consciousness appears.

Quality of consciousness

Up to this point, it has been argued that there is a real, maybe even probable, possibility that some element from PT^C is true. If this also is the case, one could pose the question what type of content could be instantiated in the unobserved territory. Here, one should be able to at least make some extrapolations from the observed to the unobserved territory in a manner that is not possible in the preceding investigation. This is because one here can focus on the positive claims of consciousness and ignore the negative.

When a person reports consciousness, she is really reporting something *rek*-conscious. A plethora of modern neuro-imaging studies has revealed that there exists a stable correlation between activity in the brain and the constitution of this *rek*-consciousness (cf. Purves, 2007). For example, when it comes to basic visual perception, each distinct *rek*-conscious percept yields specific activity patterns in the primary visual cortex to the point that if one is given fMRI-measures of the brain activity, it is, for a fairly simple stimulus input, possible to reconstruct the visual percept of the person (Thirion et al., 2006; Miyawaki et al., 2008; Naselaris, Prenger, Kay, Oliver, & Gallant, 2009; van Gerven, Cseke, de Lange, & Heskes, 2010), to some degree even for dynamic video stimuli (Nishimoto et al., 2011). Further, the more advanced and structured the conscious percept is, the more extensive and internally complex the correlating brain activity has to be. If somebody, for example, reports seeing and recognizing a face, increased brain activity is found within a hierarchy of brain modules, each one handling the computation of new higher-level features based on information received from preceding modules (Lamme, 2006).

Now, there seems to be no reason for believing that this correlation should not also hold for possible non-*prek*-conscious *q*-consciousness, so a constraint on how advanced and structured content a certain system can yield should be possible to set. For example, if the cerebellum is going to yield *nrk*-consciousness, it is also going to have to instantiate patterns of a certain complexity and extent that simply is not found there. One should therefore be able to conclude that the system that is the cerebellum is not yielding, for example, any mental experiences of self-reflection.

If one would assert that consciousness might indeed behave in a different way when not being *prek*-conscious, this would put the claim dangerously close to being an instance of Russel's teapot or Sagan's dragon. That is, if one posits properties to consciousness that are not seen within the observed territory, one basically posits a new entity all together that will, under *SA*, not hold up. Note that this is not the case for asserting that there can be raw *q*-consciousness. Here, something already existent, that is, the meta experience component of *k*-consciousness, is taken away, a move similar to, for example, asking whether there exists atoms without electrons.

So, for a certain system, where a clear α-, β- or δ-postive claim cannot be made, whether the system yields any *q*-consciousness and to what level, as has been argued above, is a non-answerable question. However, one should be able to come with extrapolations regarding how advanced and

structured the possible conscious content of it *could* be.

Summary

In this examination section it has been argued that there seems to be no apparent reasons to accept that *PT* necessarily is true or that it is wise to at least originate from such an assumption, something that is done within the IITC. Further, it has been argued that if PT^C is true, what can be said is that the possible existing territory outside of the observed territory probably will not, if it exists, be constituted by high-level content such as self-consciousness, but rather be of a more primitive nature, although possibly very content rich.

Scientific Theories of Consciousness

The critique in this thesis has so far been all about the IITC. However, with the exception of some specific parts of it, most of the discussion has really been about the possibility of excluding PT^C. As such, the argument is fairly general and should therefore be applicable to other theories of consciousness as well, as long as they, in some way, are built upon an assumption of some element of *PT*.

In this section, first, a conclusion regarding the general state of scientific theories about consciousness, along with what the researchers within the field actually do and do not investigate, will be made. Second, a pragmatic definition of consciousness will be given, one that is practical and that harmonize better with what the current scientific theories of consciousness actually are about. Third, a plea for more explicitness regarding philosophical assumptions within theories of consciousness will be formulated. Fourth, suggestions on future research within the *actual* field of consciousness research will be presented.

What actually is being investigated

As has been argued regarding the IITC, the postulated territory that is supposed to be explained is made up of some element from *PT*, which in turn approximates to a territory made up solely by *prek*-consciousness and not necessarily the full set of *q*-consciousness (see Figure 11). However, as already has been suggested, this is not unique for the IITC. Since any investigation, where the presence or absence of consciousness in different system states is set out to be explained at first needs a delineated territory to work with, all scientific theories of consciousness that are based on roughly the same territory as the IITC (that is, some element from *PT*) will be vulnerable to this same type of critique.

One example of such a theory is Baars' (2002) global workspace theory, a theory that briefly was mentioned on page 2 as one of the predecessors to the IITC. Within this theory, there really is

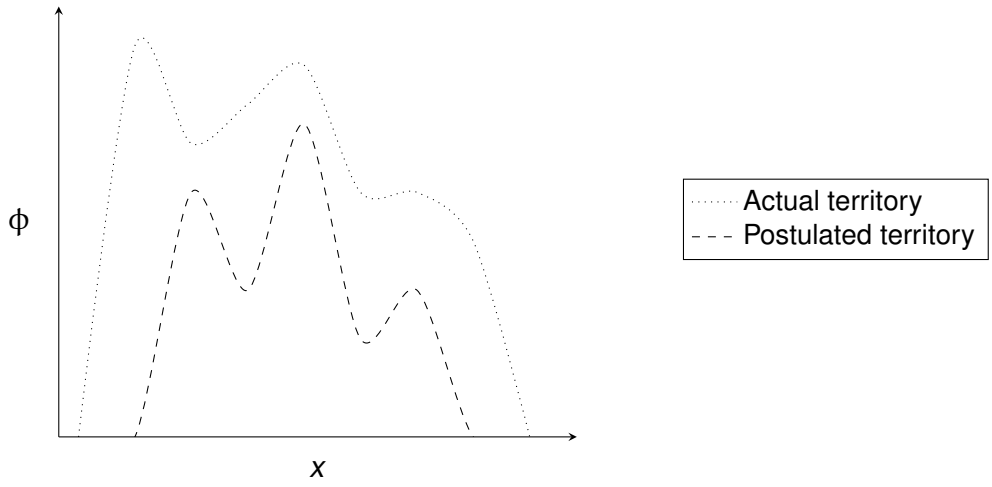

Figure 11. A mock-up depiction (adapted from Figure 7) of a *possible* scenario where the postulated territory of consciousness underestimates the shape and scope of the actual territory.

no question that a territory resembling some element from *PT* is assumed. As just one example of this, take the following quote:

> Today, conscious functions are studied experimentally by comparison with closely matched unconscious processes, an approach I have called 'contrastive analysis' [...]. Research traditions on subliminal priming, automaticity, and implicit cognition have made it clear unconscious comparison conditions for conscious processes are often available.

Another example of a theory that is vulnerable to the critique presented in this thesis is Lamme's (2003) theory of consciousness as recurrent processing. This theory states that what makes consciousness come about is recurrent interactions between groups of neurons in the brain. It is also here abundantly clear that some territory from *PT* is favored. One example of this is the following quote:

> Fundamental to the study of conscious experiences is the assumption that they are selective; we are not aware of everything we lay our eyes on. From the neuroscience perspective this implies that there is neural activity that produces conscious experience and neural activity that does not.

This does not mean that Lamme argues that we are not aware of everything we consciously see because the information is not processed in some way in the brain. Rather, he states that "obviously, there are also properties of stimuli that might never reach consciousness, not even when attended. Many invisible stimuli or attributes activate neurons."

Now, Baars' (2002) global workspace theory and Lamme's (2003) recurrent processing theory are only two representative examples of what the whole field of scientific theories of consciousness in essence looks like with regards to underlying assumptions. Therefore, if the critique against excluding PT^C is true, this would have great consequences regarding the possible interpretations of the empirical work within the field.

Merits of scientific theories of consciousness

Just because the IITC and other, so called, scientific theories of consciousness can be criticized when it comes to what they actually investigate, this does not mean that they are without merit or practical applications. One just needs to be wary of exactly in what way the theories can be useful and which philosophical implications that can be drawn from them.

It seems that even if the results from any kind of objective measurement of the brain cannot be said to correlate with the amount, or existence, of all available *q*-consciousness of that system, it seems that it, as already has been stated, *can* be said to correlate with *prek*-consciousness. Compared to raw *q*-consciousness and even *nrk*-consciousness, *prek*-consciousness is something that lies much closer to the states we generally want ourselves and others to be in. That is, as human beings, some states appeals more to us than others. For example: wakefulness is generally preferred to coma or death. What these desirable states have in common is that they all seem to be *prek*-conscious. That is, it is only the states that we actually are capable of detecting and reporting ourselves, either explicitly or just introspectively, that we can come to wish anything about.

Here, if, for example, the IITC model could continue to produce results where the value of ϕ could be shown to correlate with *prek*-consciousness, and measurements of ϕ-values in real brains could improve and get more reliable and accurate, the IITC could be used as a tool to monitor and guide us in our wish for ourselves and others to remain or come into the states that we want. An example of this is the case of comatose patients whose development of estimated ϕ-values over time, measured with the TMS/EEG-method described on pages 12–13, could give us a hint regarding their recovery to a more, for us, desirable state of healthy normality (Massimini et al., 2009, p. 211).

Another possible application for the IITC could be when it comes to pathologies like schizophrenia. When Bleuler first coined the term schizofrenia in the early 19$^{\text{th}}$ century (Kuhn, 2004), he did so to emphasize, first and foremost, the disintegration of different psychological functions that hindered a unity of the patients personality (Stotz-Ingenlath, 2000), a view that is still championed today (Friston, 1999). It has also, since then, been suggested that this disintegration of psychological functions can be neurologically explained in terms of functional parts of the brain that do not communicate with each other in a proficient enough way. This is evident from, for example, studies of oscillatory synchronicity in EEG readings (Uhlhaas & Singer, 2010), ion

flow through NMDA-channels responsible for long-range connections between distant cortical areas (Phillips & Silverstein, 2003), and the structure of human induced pluripotent stem cells derived from schizophrenic patients (Brennand et al., 2011).

Tononi and Edelman (2000) have argued, in an article where a form of proto-version of the IITC is being used, that a formalization of the integrated information of a system can be helpful to analyze data from patients with, for example, information integration pathologies like schizophrenia. When the visual cortex simulation (Tononi et al., 1992) presented on page 3, was manipulated to emulate some of the physiological underpinnings of schizophrenia, the global integration of the untouched model was lost without a lot of disruption of the specific function of the individual brain areas (Tononi & Edelman, 2000). This gave a strong rationale that a formalized information integration model could be able to detect schizophrenic symptoms in clinical studies.

A form of such a measure was then utilized in a PET study (Tononi et al., 1998) where schizophrenic and normal subjects performed some simple cognitive tasks. While any difference between the two groups could not be discerned with statistical parametric mapping (Friston, 1996), it *could* be discerned with the help of such an information integration measure, thereby showing the utility of it.

The need for more explicitness in the scientific field of consciousness studies

When Baars (2002) lists a number of positive and negative claims from different empirical research, he labels the two different categories as "Results of non-conscious conditions (not reportable)" and "Results of conscious conditions (accurately reportable)". This suggests a tentative understanding of what the two different categories really breaks down into, but there is no further mention of this in the rest of the article. If "conscious" and "unconscious" in this case only really means "reportable" and "non-reportable", and nothing else about actual q-consciousness is implied, this is poorly highlighted and the fact of the matter is that a reader in search of a clear answer to questions like these will have a hard time finding it.

In general, this illuminates a problem that encompasses pretty much all scientific work of consciousness, namely that the philosophical side of the arguments are almost always to sparse. It is important for the authors within the scientific field of consciousness studies to verbalize what actually is being investigated, what definition of consciousness they are working with, which territory in C they assume (and on what grounds), and which philosophical theory of mind they subscribe to. As of now, some, but not enough, attention is given to these questions, and it is often presented in a highly implicit manner rather than explicit. For example, when it comes to the IITC, the philosophical position that the authors subscribe to can only be derived after extensive reading of several papers and books, without especially clear directions where to look. This is unacceptable since the philosophical framework is the foundation that all subsequent theory building is built upon.

Here, scientists have to, if not join the philosophical debate, at least acknowledge the existence of these inquires and, if not willing to admit that they pose real problems, at least state their own position on them succinctly before moving on, instead of, as common today, leaving the readers hanging, subsequently forcing them to guess where the author stand with respect to these questions (if the authors have an explicit position at all, that is). This can be formulated as an explicit plea:

> *Plea*: When doing consciousness research, explicitly state your philosophical assumptions and your philosophical position/s (if not in the main text then in a footnote or in the supplementary material), so that the readers can understand, evaluate and relate your theory to other theories, without having to revert to guesswork or extensive detective work. This type of statement should be included in every article, alternatively a very clear reference should be given to work including such a statement.

Of course, one could circumvent the demands contained in this plea if one reverts to doing empirical research without making any kind of metaphysical interpretations of the results, for example, refrain from making statements about which states are associated with consciousness and which are not, but only report the actual data, letting others evaluate it from a philosophical point of view.

Future research

Given that the critique presented in this thesis is correct and PT^C cannot be excluded, how *does* one go about studying consciousness scientifically? The answer has already been hinted in the preceding discussion but will, for clarity, be recapitulated here.

As have been noted, current research, with its emphasis on neural correlates distinguishing between conscious and (allegedly) unconscious states, is a dead end regarding consciousness itself (in contrast to the investigation of the extent of *prek*-consciousness given that q-consciousness itself is presupposed). However, if one abolishes all negative claims of consciousness and replaces them with neutral claims instead, one can still do research regarding the *possible* quality of consciousness, even though the quantity of consciousness is a lost cause. That is, given that one assumes that there exists a strong correlation between the constitution of mental states and the constitution of physical states, one should be able to extrapolate what type of content could be contained in a possible existing unobservable mental state. In other words, if a feature is not physically computed within a system, there are no reasons for supposing that the corresponding mental state containing that feature could exist as a result of that system state.

Summary

In this thesis, it has been argued that current, scientific consciousness research is deeply problematic at best and totally misguided at worst. A lot of interesting empirical work has been done, research

that in its own right tells us things about how the human brain works that we could not even begin to understand before, but when the results, and subsequent theories based on these, are presented as illuminating to the question of what consciousness is—when it exists, how it arises and why we have it—the whole field has gone in way over its head.

Consider, for example, the following passage by Koch (2010), from the popular scientific journal *Scientific American*, when he talks about formalized scientific theories of consciousness such as the IITC:

> These investigations are already yielding new theories of consciousness, based on information science and mathematics, that can describe what characteristics a physical system (such as a network of neurons) would have to have to be considered conscious. Such theories will provide quantitative answers to questions that have long stumped us: Can a severely compromised patient be aware? When does a newborn baby become conscious? Is a fetus ever conscious? Is a dog aware of itself as a thinking being? What about the Internet with its billions of interconnected computers? Our society will have answers soon.

However, these answers will only be as good as their underlying assumptions, something that in this case does not work in these answers favor.

Scientists who are eager to solve the age old question of consciousness should bear Wittgenstein's (1921/2005, p. 189) famous closing line from the *Tractatus* in mind: "Whereof one cannot speak, thereof one must be silent." This holds true both when the "cannot" refers to the theoretical impossibility to actually draw conclusions from the data, as well as when it merely refers to the current researchers inability to understand what actually is supposed to be explained.

References

Aritake-Okada, S., Uchiyama, M., Suzuki, H., Tagaya, H., Kuriyama, K., Matsuura, M., ... Mishima, K. (2009). Time estimation during sleep relates to the amount of slow wave sleep in humans. *Neuroscience research*, *63*(2), 115–21. doi:10.1016/j.neures.2008.11.001

Azzopardi, P., Fallah, M., Gross, C. G., & Rodman, H. R. (2003). Response latencies of neurons in visual areas MT and MST of monkeys with striate cortex lesions. *Neuropsychologia*, *41*(13), 1738–1756. doi:10.1016/S0028-3932(03)00176-3

Baars, B. J. (1983). Conscious contents provide the nervous system with coherent, global information. In R. Davidson, G. Schwartz & D. Shaprio (Eds.), *Consciousness & self-regulation*. New York: Plenum Press.

Baars, B. J. (2002). The conscious access hypothesis: origins and recent evidence. *Trends in Cognitive Sciences*, *6*(1), 47–52. doi:10.1016/S1364-6613(00)01819-2

Balduzzi, D., & Tononi, G. (2008). Integrated Information in Discrete Dynamical Systems: Motivation and Theoretical Framework. *PLoS Computational Biology*, *4*(6). doi:10.1371/journal.pcbi.1000091

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

Bartels, A., & Zeki, S. (2005). The chronoarchitecture of the cerebral cortex. *Philosophical Transactions of the Royal Society B: Biological sciences*, *360*(1456), 733–50. doi:10.1098/rstb.2005.1627

Batista, a. P. (1999). Reach Plans in Eye-Centered Coordinates. *Science*, *285*(5425), 257–260. doi:10.1126/science.285.5425.257

Bauer, G, Gerstenbrand, F, & Rumpl, E. (1979). Varieties of the locked-in syndrome. *Journal of Neurology*, *221*(2), 77–91. doi:10.1007/BF00313105

Block, N. (2005). Two neural correlates of consciousness. *Trends in Cognitive Sciences*, *9*(2), 46–52. doi:10.1016/j.tics.2004.12.006

Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *The Behavioral and Brain Sciences*, *30*(5-6), 481–548. doi:10.1017/S0140525X07002786

Bostrom, N. (2002). *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. New York: Routledge.

Braun, a. R., Balkin, T. J., Wesenten, N. J., Carson, R. E., Varga, M, Baldwin, P, ... Herscovitch, P. (1997). Regional cerebral blood flow throughout the sleep-wake cycle. An H2(15)O PET study. *Brain*, *120*, 1173–97. doi:10.1093/brain/120.7.1173

Brennand, K. J., Simone, A., Jou, J., Gelboin-Burkhart, C., Ngoc Tran, S. S., Li, Y., ... Gage, F. H. (2011). Modelling schizophrenia using human induced pluripotent stem cells. *Nature*, *473*, 221–225. doi:10.1038/nature09915

Buhusi, C. V., & Meck, W. H. (2005). What makes us tick? Functional and neural mechanisms of interval timing. *Nature Reviews Neuroscience*, *6*(10), 755–65. doi:10.1038/nrn1764

Chalmers, D. J. (1997). *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press, USA.

Churchland, P. (1994). Can neurobiology teach us anything about consiousness? In *Proceedings and addresses of the american philosophical association* (Vol. 67, *4*, pp. 23–40).

Corkin, S. (2002, February). What's new with the amnesic patient H.M.? *Nature Reviews Neuroscience*, *3*(2), 153–60. doi:10.1038/nrn726

Damasio, a. R. (1989, November). Time-locked multiregional retroactivation: a systems-level proposal for the neural substrates of recall and recognition. *Cognition*, *33*(1-2), 25–62. doi:10.1016/0010-0277(89)90005-X

Dennett, D. (2001, April). Are we explaining consciousness yet? *Cognition*, *79*(1-2), 221–37. doi:10.1016/S0010-0277(00)00130-X

Dennett, D. C. (1991). *Consciousness Explained*. New York: Back Bay Books.

Descartes, R. (2008). *Meditations on first philosophy* (E. S. Haldane, Trans.). Retrieved from http://books.google.com/books?id=P1roPQNFKDcC\&lpg=PP1\&pg=PA25\#v=onepage\&q\&f=false

Diekelmann, S., & Born, J. (2010). The memory function of sleep. *Nature Reviews Neuroscience*, *11*(2), 114–26. doi:10.1038/nrn2762

Edelman, G. (1990). *The Remembered Present: A Biological Theory of Consciousness*. New York: Basic Books.

Edelman, G., & Tononi, G. (2000). *A Universe Of Consciousness: How Matter Becomes Imagination*. New York: Basic Books.

Edelman, G. M. (2003). Naturalizing consciousness: a theoretical framework. *PNAS*, *100*(9), 5520–5524. doi:10.1073/pnas.0931349100

Edinger, J. D., & Fins, A. I. (1995). The distribution and clinical significance of sleep time misperceptions among insomniacs. *Sleep*, *18*(4), 232–9.

Einstein, A. (1934). On the method of theoretical physics. *Philosophy of science*, *1*(2), 163–169.

Engel, A. K., Fries, P, & Singer, W. (2001). Dynamic predictions: oscillations and synchrony in top-down processing. *Nature Reviews Neuroscience*, *2*(10), 704–16. doi:10.1038/35094565

Fodor, J. A. (1983). *The Modularity of Mind*. MIT press: The MIT Press.

Fraisse, P. (1984, January). Perception and estimation of time. *Annual review of psychology*, *35*, 1–36. doi:10.1146/annurev.ps.35.020184.000245

Friston, K. J. (1999). Schizophrenia and the disconnection hypothesis. *Acta Psychiatrica Scandinavica*, *99*(s395), 68–79. doi:10.1111/j.1600-0447.1999.tb05985.x

Friston, K. (1996). Statistical parametric mapping and other analyses of functional imaging data. In A. Toga & J. Mazziotta (Eds.), *Brain mapping: the methods* (pp. 363–386). Waltham, MA: Academic Press.

George, D., & Hawkins, J. (2009). Towards a mathematical theory of cortical micro-circuits. *PLoS Computational Biology*, *5*(10), e1000532. doi:10.1371/journal.pcbi.1000532

Goebel, R., Muckli, L., Zanella, F. E., Singer, W., & Stoerig, P. (2001). Sustained extrastriate cortical activation without visual awareness revealed by fMRI studies of hemianopic patients. *Vision Research*, *41*(10-11), 1459–1474. doi:10.1016/S0042-6989(01)00069-4

Goodman, N., & Putnam, H. (1983). *Fact, Fiction, and Forecast* (4th ed.). Cambridge, MA: Harvard University Press.

Greenberg, D. L. (2007). Comment on "Detecting awareness in the vegetative state". *Science (New York, N.Y.) 315*(5816), 1221; author reply 1221. doi:10.1126/science.1135284

Hawkins, J., & Blakeslee, S. (2004). *On Intelligence*. New York: Times Books.

Herzog, M. H., Esfeld, M., & Gerstner, W. (2007). Consciousness & the small network argument. *Neural Networks*, *20*(9), 1054–6. doi:10.1016/j.neunet.2007.09.001

Hobson, J. a., Pace-Schott, E. F., & Stickgold, R. (2000). Dreaming and the brain: toward a cognitive neuroscience of conscious states. *The Behavioral and Brain Sciences*, *23*(6), 793–842. doi:10.1017/S0140525X00003976

Hofstadter, D. R. (2007). *I Am a Strange Loop*. New York: Basic Books.

Jackson, F. (1982). Epiphenomenal qualia. *The Philosophical Quarterly*, *32*, 127–136.

Jackson, F. (1986). What Mary didn't know. *The Journal of Philosophy*, *83*(5), 291–295.

Kaufmann, C, Wehrle, R, Wetter, T. C., Holsboer, F, Auer, D. P., Pollmächer, T, & Czisch, M. (2007, March). Brain activation and hypothalamic functional connectivity during human non-rapid eye movement sleep: an EEG/fMRI study. *Brain*, *130*(7), e75. doi:10.1093/brain/awm084

Kim, J. (1989). The myth of nonreductive materialism. *Proceedings and Addresses of the American Philosophical Association*, *63*(3), 31–47.

Kim, J. (2005). *Physicalism, or something near enough*. Princeton, NJ: Princeton University Press.

Kirk, R. (2011). Zombies. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Spring 2011 ed.). Retrieved from http://plato.stanford.edu/archives/spr2011/entries/zombies/

Koch, C. (2010). An Answer to the Riddle of Consciousness. *Scientific American*, 76.

Komssi, S., & Kähkönen, S. (2006). The novelty value of the combined use of electroencephalography and transcranial magnetic stimulation for neuroscience research. *Brain Research Reviews*, *52*(1), 183–192. doi:10.1016/j.brainresrev.2006.01.008

Korzybski, A. (1933). *Science and sanity: an introduction to non-Aristotelian systems and general ...* Institute of General Semantics.

Kuhn, R. (2004, September). Eugen Bleuler's Concepts of Psychopathology. *History of Psychiatry*, *15*(3), 361–366. doi:10.1177/0957154X04044603

Lamme, V. (2010). How neuroscience will change our view on consciousness. *Cognitive Neuroscience*, *1*(3), 204–220. doi:10.1080/17588921003731586

Lamme, V. A. (2003, January). Why visual attention and awareness are different. *Trends in Cognitive Sciences*, *7*(1), 12–18.

Lamme, V. A. (2006, November). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*, *10*(11), 494–501. doi:10.1016/j.tics.2006.09.001

Lineweaver, C. H., & Davis, T. M. (2005). Misconceptions about the Big Bang. *Scientific American*, *21*(Febraury), 1–5.

Ling, S., & Xing, C. (2004). *Coding theory: a first course*. Cambridge, England: Cambridge University Press.

List, C., & Menzies, P. (2009). Non-reductive physicalism and the limits of the exclusion principle. *Journal of Philosophy*, *106*(9).

Lovblad, K.-O., Thomas, R., Jakob, P. M., Scammell, T., Bassetti, C., Griswold, M., ... Warach, S. (1999). Silent functional magnetic resonance imaging demonstrates focal activation in rapid eye movement sleep. *Neurology*, *53*(9), 2193–2195.

Lumer, E. D., Edelman, G. M., & Tononi, G. (1997a). Neural dynamics in a model of the thalamocortical system. I. Layers, loops and the emergence of fast synchronous rhythms. *Cerebral Cortex*, *7*(3), 207–227.

Lumer, E. D., Edelman, G. M., & Tononi, G. (1997b). Neural dynamics in a model of the thalamocortical system. II. The role of neural synchrony tested through perturbations of spike timing. *Cerebral Cortex*, *7*(3), 228–236.

Magritte, R. (1928). The treason of images [Painting].

Massimini, M., Ferrarelli, F., Huber, R., Esser, S. K., Singh, H., & Tononi, G. (2005). Breakdown of cortical effective connectivity during sleep. *Science*, *309*(5744), 2228–2232. doi:10.1126/science.1117256

Massimini, M., Ferrarelli, F., Esser, S. K., Riedner, B. A., Huber, R., Murphy, M., ... Tononi, G. (2007). Triggering sleep slow waves by transcranial magnetic stimulation. *PNAS*, *104*(20), 8496–8501. doi:10.1073/pnas.0702495104

Massimini, M., Boly, M., Casali, A., Rosanova, M., & Tononi, G. (2009). A perturbational approach for evaluating the brain's capacity for consciousness. *Progress in Brain Research*, *177*, 201–214. doi:10.1016/S0079-6123(09)17714-2

Means, M. (2003). Accuracy of sleep perceptions among insomnia sufferers and normal sleepers. *Sleep Medicine*, *4*(4), 285–296. doi:10.1016/S1389-9457(03)00057-1

Mercer, J. D., Bootzin, R. R., & Lack, L. C. (2002). Insomniacs' perception of wake instead of sleep. *Sleep*, *25*(5), 564–571.

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M.-a., Morito, Y., Tanabe, H. C., … Kamitani, Y. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, *60*(5), 915–29. doi:10.1016/j.neuron.2008.11.004

Monti, M., Vanhaudenhuyse, A., Coleman, M., Boly, M., Pickard, J., Tshibanda, L., … Laureys, S. (2010). Willful Modulation of Brain Activity in Disorders of Consciousness. *New England Journal of Medicine*, *362*(7), 579.

Moody, T. (2003). Consciousness and complexity. *PCID*, *2.3*.

Moorcroft, W. H., Kayser, K. H., & J, A. (1997). Subjective and Objective Confirmation of the Ability to Self-Awaken at a Self-Predetermined Time Without U sing External Means. *Sleep*, *20*(14), 40–45.

Murata, A, Gallese, V, Kaseda, M, & Sakata, H. (1996). Parietal neurons related to memory-guided hand manipulation. *Journal of Neurophysiology*, *75*(5), 2180–2186.

Musallam, S, Corneil, B. D., Greger, B, Scherberger, H, & Andersen, R. A. (2004). Cognitive control signals for neural prosthetics. *Science*, *305*(5681), 258–62. doi:10.1126/science.1097938

Nachev, P., & Husain, M. (2007). Comment on "Detecting awareness in the vegetative state". *Science*, *315*(5816), 1221. doi:10.1126/science.1135096

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, *83*(4), 435. doi:10.2307/2183914

Nakayama, K., He, Z. J., & Shimojo, S. (1995). Visual surface representation: a critical link between lower-level and higher-level vision. In S. M. Kosslyn & D. N. Osherson (Eds.), *An invitation to cognitive science: visual cognition* (pp. 1–70). Cambridge, MA: The MIT Press.

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron*, *63*(6), 902–915. doi:10.1016/j.neuron.2009.09.006

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies. *Current Biology*, *21*(19), 1641–1646. doi:10.1016/j.cub.2011.08.031

Owen, A. M., Coleman, M. R., Boly, M., Davis, M. H., Laureys, S., Jolles, D., & Pickard, J. D. (2007). Response to Comments on "Detecting Awareness in the Vegetative State". *Science*, *315*(5816), 1221c–1221c. doi:10.1126/science.1135583

Owen, A. M., Coleman, M. R., Boly, M., Davis, M. H., Laureys, S., & Pickar, J. D. (2006). Detecting Awareness in the Vegetative State. *Science*, *313*(5792), 1402.

Phillips, W. a., & Silverstein, S. M. (2003). Convergence of biological and psychological perspectives on cognitive coordination in schizophrenia. *Behavioral and Brain Sciences*, *26*(1), 65–137. doi:10.1017/S0140525X03000025

Pinto, L. R., Pinto, M. C. R., Goulart, L. I., Truksinas, E., Rossi, M. V., Morin, C. M., & Tufik, S. (2009). Sleep perception in insomniacs, sleep-disordered breathing patients, and healthy volunteers–an important biologic parameter of sleep. *Sleep Medicine*, *10*(8), 865–868. doi:10.1016/j.sleep.2008.06.016

Plutarch. (n.d). Theseus. Retrieved from http://classics.mit.edu/Plutarch/theseus.html

Pöppel, E. (1994). Temporal mechanisms in perception. *International Review of Neurobiology*, *37*, 185–202, 203–207.

Popper, K., & Eccles, J. C. (1984). *The Self and Its Brain: An Argument for Interactionism*. London: Routledge.

Purves, D. (2007). *Principles of Cognitive Neuroscience*. Sunderland, MA: Sinauer Associates Inc.

Putnam, H. (1975). The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science*, *7*, 131–193.

Quine, W. V. (1969). Natural kinds. In *Ontological reality and other essays* (pp. 114–139). New York: Columbia University Press.

Rey, G. (1983). A reason for doubting the existence of consciousness. In R. Davidson, G. Schwartz & D. Shapiro (Eds.), *Consciousness and self-regulation, volume 3* (pp. 1–39). New York: Plenum.

Rey, G. (1988). A Question About Consciousness. In H. Otto & J. Tuedio (Eds.), *Perspectives on mind* (pp. 5–24). Dorderecht: Reidel.

Rioux, I., Tremblay, S., & Bastien, C. H. (2006). Time estimation in chronic insomnia sufferers. *Sleep*, *29*(4), 486–493.

Rock, I., & Mack, A. (2000). *Inattentional Blindness*. Cambridge, MA: The MIT Press.

Russell, B. (1997). Is there a God? In J. G. Slater & P. Köllner (Eds.), *The collected papers of bertrand russell, volume 11: last philosophical testament, 1943-68* (pp. 543–548). London: Routledge.

Sagan, C., & Druyan, A. (1997). *The Demon-Haunted World: Science as a Candle in the Dark*. New York: Ballantine Books.

Savitt, S. (1975). Rorty's disappearance theory. *Philosophical Studies*, *28*(6), 433–436.

Scoville, W., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery & Psychiatry*, *20*(1), 11. doi:10.1136/jnnp.20.1.11

Seager, W., & Allen-Hermanson, S. (2010). Panpsychism. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Fall 2010 ed.). Retrieved from http://plato.stanford.edu/archives/fall2010/entries/panpsychism/

Sengpiel, F. (1997). Binocular rivalry: ambiguities resolved. *Current Biology*, *7*(7), R447–50. doi:10.1016/S0960-9822(06)00215-6

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423, 623–656.

Silber, M., Ancoli-Israel, S., Bonnet, M., Chokroverty, S., Grigg-Damberger, M., Hirshkowitz, M., ... Others. (2007). The visual scoring of sleep in adults. *Journal of Clinical Sleep Medicine*, *3*(2), 121–131.

Smith, E., & Jonides, J. (1999). Storage and executive processes in the frontal lobes. *Science*, *283*(5408), 1657–1661. doi:10.1126/science.283.5408.1657

Snyder, L., Batista, A., & Andersen, R. (1997). Coding of intention in the posterior parietal cortex. *Nature*, *386*, 167–170. doi:10.1038/386167a0

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied*, *74*(11), 1–29.

Srinivasan, R, Russell, D. P., Edelman, G. M., & Tononi, G. (1999). Increased synchronization of neuromagnetic responses during conscious perception. *The Journal of Neuroscience*, *19*(13), 5435–5448.

Stickgold, R, Malia, A, Fosse, R, Propper, R, & Hobson, J. A. (2001). Brain-mind states: I. Longitudinal field study of sleep/wake factors influencing mentation report length. *Sleep*, *24*(2), 171–179.

Stoerig, P, & Cowey, A. (1997). Blindsight in man and monkey. *Brain*, *120*, 535–359. doi:10.1093/brain/120.3.535

Stotz-Ingenlath, G. (2000). Epistemological aspects of Eugen Bleuler's conception of schizophrenia in 1911. *Medicine, Health Care and Philosophy*, *3*(2), 153–159. doi:10.1023/A:1009919309015

Strand, A. (2010). Causal exclusion and the preservation of causal sufficiency. *Sats*, *11*(2), 117–135. doi:10.1515/sats.2010.011

Strawson, G. (2006). Realistic monism: why physicalism entails panpsychism. *Journal of Consciousness Studies*, *13*(10-11), 3–31.

Tang, N. K. Y., & Harvey, A. G. (2005). Time Estimation Ability and Distorted Perception of Sleep in Insomnia. *Behavioral Sleep Medicine*, *3*(3), 134–150. doi:10.1207/s15402010bsm0303

Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J.-B., Lebihan, D., & Dehaene, S. (2006). Inverse retinotopy: inferring the visual content of images from brain activation patterns. *NeuroImage*, *33*(4), 1104–1116. doi:10.1016/j.neuroimage.2006.06.062

Tononi, G, & Edelman, G. M. (1998). Consciousness and complexity. *Science*, *282*(5395), 1846–1851. doi:10.1126/science.282.5395.1846

Tononi, G, & Edelman, G. M. (2000). Schizophrenia and the mechanisms of conscious integration. *Brain Research. Brain Research Reviews*, *31*(2-3), 391–400. doi:10.1016/S0165-0173(99)00056-9

Tononi, G, Sporns, O, & Edelman, G. M. (1992). Reentry and the problem of integrating multiple cortical areas: simulation of dynamic integration in the visual system. *Cerebral Cortex*, *2*(4), 310–35.

Tononi, G, McIntosh, a. R., Russell, D. P., & Edelman, G. M. (1998). Functional clustering: identifying strongly interactive brain regions in neuroimaging data. *NeuroImage*, *7*(2), 133–49. doi:10.1006/nimg.1997.0313

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, *5*, 42. doi:10.1186/1471-2202-5-42

Tononi, G. (2007). Consciousness and the Brain [Video file]. Retrieved from http://video.google.com/videoplay?docid=-7502852812875314243

Tononi, G. (2008, December). Consciousness as integrated information: a provisional manifesto. *The Biological Bulletin*, *215*(3), 216–42.

Tononi, G., & Cirelli, C. (2006). Sleep function and synaptic homeostasis. *Sleep Medicine Reviews*, *10*(1), 49–62. doi:10.1016/j.smrv.2005.05.002

Tononi, G., & Koch, C. (2008, March). The neural correlates of consciousness: an update. *Annals of the New York Academy of Sciences*, *1124*, 239–61. doi:10.1196/annals.1440.004

Tononi, G., & Sporns, O. (2003). Measuring information integration. *BMC Neuroscience*, *4*, 31. doi:10.1186/1471-2202-4-31

Uhlhaas, P. J., & Singer, W. (2010). Abnormal neural oscillations and synchrony in schizophrenia. *Nature Reviews Neuroscience*, *11*(2), 100–13. doi:10.1038/nrn2774

van Gerven, M. a. J., Cseke, B., de Lange, F. P., & Heskes, T. (2010). Efficient Bayesian multivariate fMRI analysis using a sparsifying spatio-temporal prior. *NeuroImage*, *50*(1), 150–61. doi:10.1016/j.neuroimage.2009.11.064

Vanable, P. A., Aikens, J. E., Tadimeti, L, Caruana-Montaldo, B, & Mendelson, W. B. (2000). Sleep latency and duration estimates among sleep disorder patients: variability as a function of sleep disorder diagnosis, sleep history, and psychological characteristics. *Sleep*, *23*(1), 71–79.

William, R. (2011). Eliminative materialsim. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Spring 2011 ed.). Retrieved from http://plato.stanford.edu/archives/spr2011/entries/materialism-eliminative/

Wittgenstein, L. (2005). *Tractatus Logico-Philosophicus*. New York: Routledge.

Wolfe, J. M. (1999). Inattentional Amnesia. In V. Coltheart (Ed.), *Fleeting memories* (Vol. 17, *5*, pp. 71–94). Cambridge, MA: MIT Press.

Woodward, J. (2008). Mental causation and neural mechanisms. In *Being reduced: new essays on reduction, explanation and causation* (pp. 1–57). Oxford University Press.