Information Closure Theory of Consciousness

Acer Y.C. Chang*, Martin Biehl†, Yen Yu‡ and Ryota Kanai§

ARAYA, Inc., Tokyo, Japan

Contents

Abstract

Information processing in neural systems can be described and analysed at multiple spatiotemporal scales. Generally, information at lower levels is more fine-grained but can be coarse-grained at higher levels. However, only information processed at specific scales of coarse-graining appears to be available for conscious awareness. We do not have direct experience of

*Corresponding author, acercyc@araya.org
†martin@araya.org
‡yen.yu@araya.org
§kanair@araya.org

information available at the scale of individual neurons, which is noisy and highly stochastic. Neither do we have experience of more macro-scale interactions, such as interpersonal communications. Neurophysiological evidence suggests that conscious experiences co-vary with information encoded in coarse-grained neural states such as the firing pattern of a population of neurons. In this article, we introduce a new informational theory of consciousness: Information Closure Theory of Consciousness (ICT). We hypothesise that conscious processes are processes which form non-trivial informational closure (NTIC) with respect to the environment at certain coarse-grained scales. This hypothesis implies that conscious experience is confined due to informational closure from conscious processing to other coarse-grained scales. ICT proposes new quantitative definitions of both conscious content and conscious level. With the parsimonious definitions and a hypothesise, ICT provides explanations and predictions of various phenomena associated with consciousness. The implications of ICT naturally reconcile issues in many existing theories of consciousness and provides explanations for many of our intuitions about consciousness. Most importantly, ICT demonstrates that information can be the common language between consciousness and physical reality.

Keywords:

Keywords: theory of consciousness, non-trivial informational closure, NTIC, coarse-graining, level of analysis

1 Introduction

Imagine you are a neuron in Alice's brain. Your daily work is to collect neurotransmitters through dendrites from other neurons, accumulate membrane potential, and finally send signals to other neurons through action potentials along axons. However, you have no idea that you are one of the neurons in Alice's supplementary motor area and are involved in many motor control processes for Alice's actions, such as grabbing a cup. You are ignorant of intentions, goals, and motor plans that Alice has at any moment, even though you are part of the physiological substrate responsible for all these actions. A similar story also happens in Alice's conscious mind. To grab a cup, for example, Alice is conscious of her intention and visuosensory experience of this action. However, her conscious experience does not reflect the dynamic of your membrane potential or the action potentials you send to other neurons every second. That is, not all the information you have is available to Alice's conscious mind.

It appears to be true that we do not consciously access information processed at every scale in the neural system. There are both more microscopic and more macroscopic scales than the scale corresponding to the conscious contents. On the one hand, the dynamics of individual neurons are stochastic (Goldwyn & Shea-Brown, 2011; White *et al.*, 2000). However, what we are aware of in our conscious mind shows astonishing stability and robustness against the ubiquitous noise in the neural system (Mathis & Mozer, 1995). In addition, some parts of the neural system contribute very little to conscious experience (the cerebellum for example (Lemon & Edgley, 2010)), also suggesting that conscious contents do not have one-to-one mapping to the entire state of the neural system. On the other hand, human conscious experience is more detailed than just a simple (e.g. binary) process can represent, suggesting that the state space of conscious experience is much larger than what a single overly coarse-grained binary variable can represent. These facts suggest that conscious processes occur at a particular scale. We currently have possess only a few theories (e.g., Integrated Information Theory (Hoel *et al.*, 2016) and Geometric Theory of Consciousness (Fekete & Edelman, 2011, 2012)) to identify the scale to which conscious processes correspond (also see discussion in Fekete *et al.* (2016)). We refer to this notion as **the scale problem of consciousness** (Fig. 1).

In this article, we propose a new information-based theory of consciousness, called the Information Closure Theory of Consciousness (ICT). We argue that every process with a positive non-trivial information closure (NTIC) has consciousness. This means that the state of such a process corresponds one-to-one to conscious content.[1]. We further postulate that the *level* of consciousness corresponds to the degree of NTIC. (For a discussion of the distinction between level versus content of consciousness see Laureys (2005); Overgaard & Overgaard (2010)).

In the following, we first introduce non-trivial informational closure and argue for its importance to information processing for human scale agents (Sec. 2). We next argue that through coarse-graining the neural system can form informational closure and a high degree of NTIC at a specific scale of coarse-graining (Sec. 3). In Sec. 4, we propose a new theory of consciousness (ICT). We also illustrate how ICT can parsimoniously explain empirical findings from previous consciousness studies (Sec. 5) and reconcile several current major theories of consciousness (Sec.6). Finally, we discuss the current theoretical and empirical limitations of ICT and propose the implications of ICT on the current consciousness science (Sec.7).

[1] In the following IC stands for "informational closure" or "informationally closed" and NTIC stands for "non-trivial informational closure" or "non-trivially informationally closed".

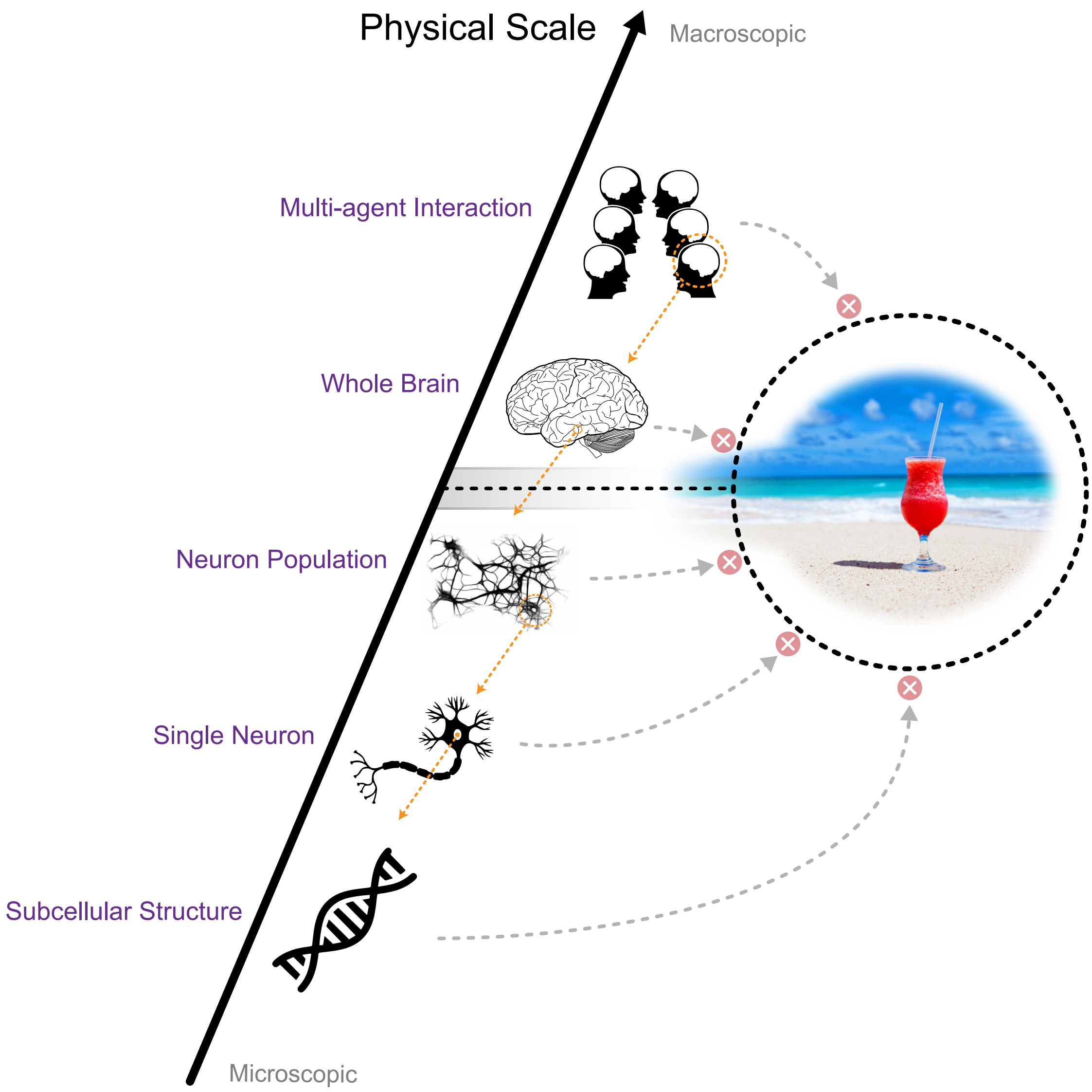

Figure 1: The scale problem of consciousness: Human conscious experience does not reflect information from every scale. Only information at a certain coarse-grained scale in the neural system is reflected in consciousness.

2 Non-trivial Informational Closure

The notion of non-trivial informational closure (NTIC) was introduced by Bertschinger *et al.* (2006). The concept of closure is closely related to system identification in systems theory. One can distinguish a system from its environment by computing the closedness of the system (Luhmann, 1995; Maturana & Varela, 1991; Pattee, 2012; Rosen, 1991). Closedness itself can be further quantified by information theory.

Consider two processes, the environment process $(E_t)_{t\in\mathbb{N}}$ and the system process $(Y_t)_{t\in\mathbb{N}}$ and let their interaction be described by the Bayesian network with the sensor channel $\hat{e}_t$ and the action $\hat{y}_t$ channel in Fig. 2. Information flow J_t from the environment E to a system S at time t can then be defined as the conditional mutual information I between the current environment state E_t and the future system state Y_{t+1} given the current system state Y_t

$$\begin{aligned} J_t(E \to Y) \quad &:= I(Y_{t+1}; E_t | Y_t) \\ &= I(Y_{t+1}; E_t) - (I(Y_{t+1}; Y_t) - I(Y_{t+1}; Y_t | E_t)) \end{aligned} \tag{1}$$

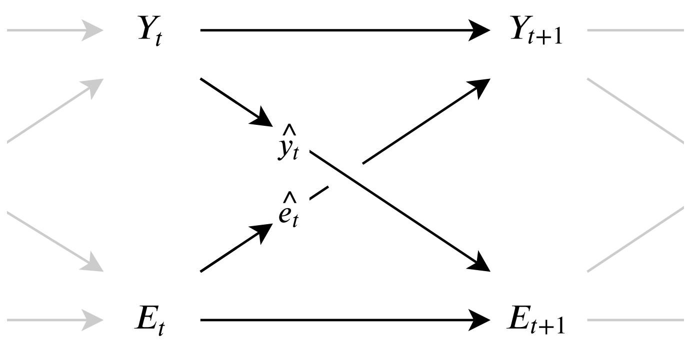

Figure 2: Dependencies between a system Y and its environment E through the channels $\hat{y}_t$ and $\hat{e}_t$.

Bertschinger *et al.* (2006) defines a system as informationally closed when information flow from the environment to the system is zero.

$$J_t(E \to Y) = 0 \tag{2}$$

Information closure (minimising J_t) is trivial if the environment and the system are entirely independent of each other.

$$I(Y_{t+1}; E_t) = 0 \quad \Rightarrow \quad J_t(E \to Y) = 0 \tag{3}$$

However, informational closure can be formed non-trivially. In the non-trivial case, even though a system contains (or encodes) information about the environmental dynamics, the system can still be informationally closed. In such cases, the mutual information between the current states of the environment and the future state of the system is larger than zero.

$$I(Y_{t+1}; E_t) > 0 \tag{4}$$

This also implies

$$I(Y_{t+1}; Y_t) - I(Y_{t+1}; Y_t|E_t) > 0 \tag{5}$$

And, non-trivial informational closure can be defined as

$$\begin{aligned} NTIC_t(E \to Y) :&= I(Y_{t+1}; Y_t) - I(Y_{t+1}; Y_t|E_t) \qquad (6)\\ &= I(Y_{t+1}; E_t) - I(Y_{t+1}; E_t|Y_t) \qquad (7) \end{aligned}$$

Hence, maximising $NTIC_t(E \to Y)$ amounts to

$$\begin{array}{lll} \text{maximising} & I(Y_{t+1}; Y_t) & \text{and} \\ \text{minimising} & I(Y_{t+1}; Y_t|E_t) & \end{array} \tag{8}$$

One can also maximise $NTIC_t(E \to Y)$ by

$$\begin{array}{lll} \text{maximising} & I(Y_{t+1}; E_t) & \text{and} \\ \text{minimising} & I(Y_{t+1}; E_t|Y_t) & \end{array} \tag{9}$$

This implies that the system contains within itself all the information about its own future and the self-predictive information contains the information about the environment.

2.1 Informational Closure Does not Imply Causality

A surprising result from the definition of information flow $J_t(E \to Y)$ (Eq. 1) is that information flow does not indicate causal dependency from E_t to Y_{t+1} or from Y_t to Y_{t+1}. Here we consider two scenarios, called *modelling* and *passive adaptation*, which previously noted by Bertschinger *et al.* (2006). In both scenarios, a process can form positive NTIC ($NTIC(E \to Y) > 0$) and informational closure ($J(E \to Y) = 0$), albeit via different causal dependencies.

In the scenario of *modelling*, to achieve positive NTIC and informational closure, a system can internalise and synchronise with the dynamics of the environment, e.g., model the environment. In this case, the future internal state Y_{t+1} of the system is driven by the current internal state Y_t and the system still retains mutual information with the environment. Having high degrees of NTIC then entails having high predictive power about the environment. This gives biological agents functional and evolutionary advantages.

In the *passive adaptation* scenario, counterintuitively, the future system states (Y_{t+1}) are entirely driven by the current environment states (E_t) but can also have positive NTIC and informational closure. This happens under the condition that the sensory process $\hat{e}_t$ is deterministic and the system simply copies the sensory values. The system is then a copy of another informationally closed process ($\hat{e}_t$) and is therefore closed itself. At the same time it has mutual information with the process that it is copying.

In most of the realistic cases, however, the environment is partially observable to the system, and thus the sensory process is usually not deterministic. Accordingly, it is difficult for the system to be informationally closed and have higher NTIC. More importantly, we argue in the Appendix that whenever the environment has itself more predictable dynamics than the observations, there is the potential exists for a process to achieve higher NTIC by modelling the environment rather than by copying the observations.

We will see that both scenarios are relevant to ICT in the following sections.

3 Coarse-graining in the Neural System

The formation of NTIC with a highly stochastic process is challenging. NTIC requires the predictability of the system state and is therefore impeded by noise in the system. Information processing at the microscopic scale (cellular scale) in neural systems suffers from multiple environmental noise sources such as sensor, cellular, electrical, and synaptic noises. For example, neurons exhibit large trial-to-trial variability at the cellular scale, and are subject to thermal fluctuations and other physical noises (Faisal *et al.*, 2008).

Nevertheless, it is possible that neural systems form NTIC at certain macroscopic scales through coarse-graining of microscopic neural states. Coarse-graining refers to many-to-one or one-to-one maps which aggregate microscopic states to a macroscopic state. In other words, a number of different micro-states correspond to the same value of the macro-variable (Price & Corry, 2007). Coarse-grainings can therefore form more stable and deterministic state transitions and more often form NTIC processes. For neural systems this means that a microscopically noisy neural system may still give rise to an NTIC process on a more macroscopic scale.

Indeed, empirical evidence suggests that coarse-graining is a common coding strategy of the neural system by which it establishes robustness against noise at microscopic scales. For instance, the inter-spike intervals of an individual neuron are stochastic. This implies that the state of an individual neuron does not represent stable information. However, the firing rate, i.e. the average spike counts over a given time interval, is more stable and robust against noise such as the variability in inter-spike intervals. Using this temporal coarse-graining strategy, known as rate coding (Adrian, 1926; Gerstner & Kistler, 2002; Maass & Bishop, 2001; Panzeri *et al.*, 2015; Stein *et al.*, 2005), neurons can encode stimulus intensity by increasing or decreasing their firing rate (Kandel *et al.*, 2000). (Stein *et al.*, 2005). The robustness of rate coding is a direct consequence of the many-to-one mapping (i.e., coarse-graining).

Population coding is another example of encoding information through coarse-graining in neural systems. In this coding scheme, information is encoded by the activation patterns of a set of neurons (a neuron population). In the population coding scheme, many states of a neuron population map to the same state of macroscopic variables which encode particular informational contents, thereby reducing the influence of noise in individual neurons. That is, stable representations can be formed through coarse-graining the high dimensional state space of a neuron population to a lower dimensional macroscopic state space (Binder *et al.*, 2009; Kristan Jr & Shaw, 1997; Pouget *et al.*, 2000; Quian Quiroga & Panzeri, 2009). Therefore, individual neuron states (microscopic scale) are not sufficiently informative about the complete encoded contents at the population scale (macroscopic scale). Instead, coarse-grained variables are better substrates for stably encoding information and allow the neural system to ignore noisy interactions at the fine-grained scale (Woodward, 2007).

These two examples show that the known coding schemes can be viewed as coarse-graining, and provide stochastic neural systems with the ability to form more stable and deterministic macroscopic processes for encoding and processing information reliably. We argue that coarse-graining allows neural systems to form NTIC processes at macroscopic scales. Based on the merit of coarse-graining in neural systems, we propose a new theory of consciousness in the next section.

4 Information Closure Theory of Consciousness

In this section, we propose a new theoretical framework of consciousness: the Information Closure Theory of Consciousness (ICT). The main hypothesis is that conscious processes are captured by what we call *C-processes*. We first define C-processes, then state our hypothesis and discuss its implications.

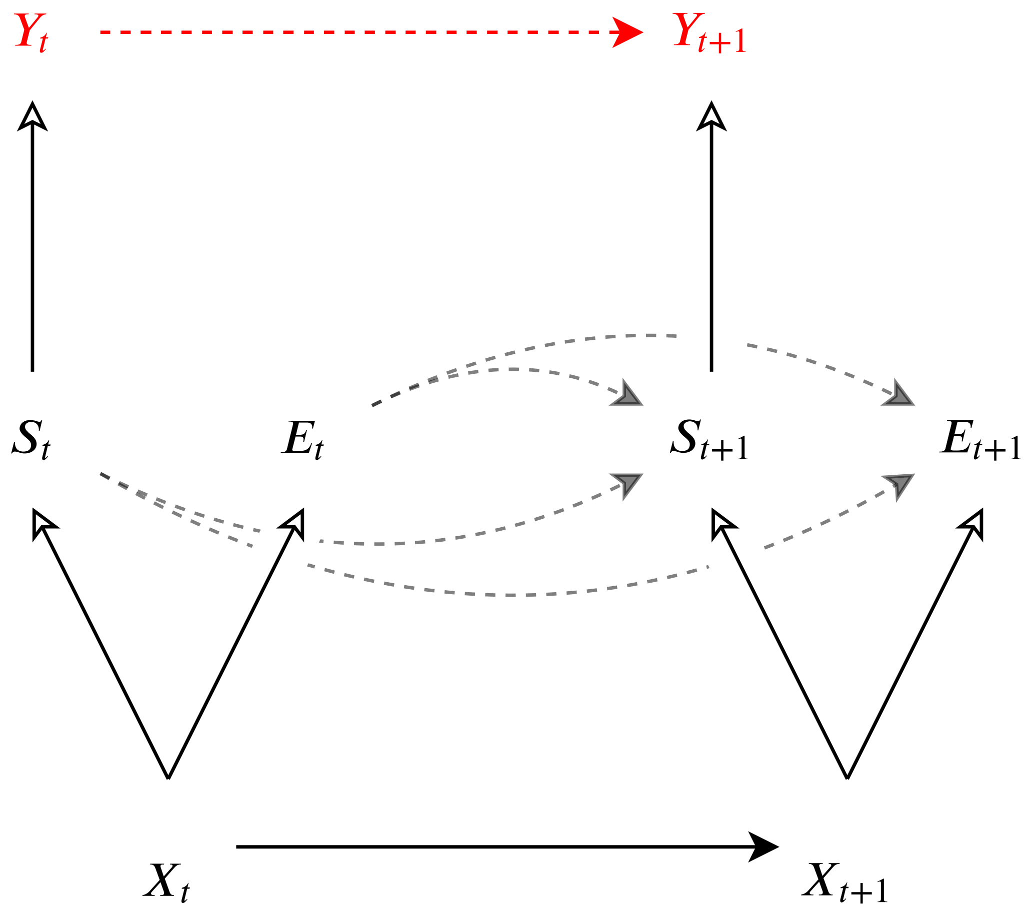

Figure 3: The information flow amounts the universe X, the system S, the environment of the system E, and the coarse-grained process Y of the system S. The solid line with a filled arrow from X_t to X_{t+1} represents the microscopic dynamic of the universe. The solid lines with a empty arrow represent directions of coarse-graining. The dashed lines represents virtual dependencies between two macroscopic variables. The red Y_t, Y_{t+1}, and the red dashed line in between represents a macroscopic process which forms informational closure at a certain coarse-grained scale.

To define C-processes we first need to define coarse-grainings. Every coarse-graining is characterised by a function that maps the microscopic process to the coarse-grained macroscopic process. More formally:

Definition 1. *Given a stochastic process X with state space $\mathcal{X}$, a* coarse-graining of X *is a stochastic process Y with state space $\mathcal{Y}$ such that there exists a function* [2] $f_Y : \mathcal{X} \to \mathcal{Y}$ *with* $Y_t = f_Y(X_t)$.

A more general definition of coarse-grainings that maps temporally extended sequences of the microscopic process to macroscopic states are possible, but for this first exposure of our theory the simpler definition above is sufficient.

Definition 2. *Given a stochastic process X called the universe process, a* C-process *is a coarse-graining Y of X such that the following two conditions are satisfied (see Fig. 3):*

1. *Y is informationally closed to X*
2. *there exists a pair (S, E) of coarse-grainings of X such that*
   - *Y is a coarse-graining of S,*
   - *the state space $\mathcal{X}$ of X is equal to the Cartesian product of the state spaces $\mathcal{S}$ and $\mathcal{E}$ of processes S and E respectively, formally $\mathcal{X} = \mathcal{S} \times \mathcal{E}$, and*
   - *Y is NTIC to E, formally:*

$$NTIC_t(E \to Y) > 0 \tag{10}$$

Note that, here we applied the same definitions of information flow (Eq. 1)

$$J_t(E \to Y) = I(Y_{t+1}; E_t | Y_t) \tag{11}$$

to the system-environment dependency and the micro-macro scale dependency

$$J_t(X \to Y) = I(Y_{t+1}; X_t | Y_t) \tag{12}$$

even though the Bayesian graphs differ in the two scenarios. Both these settings have been previously used in the literature (see Bertschinger *et al.*, 2006; Pfante *et al.*, 2014b).

With the two definitions we can state the main hypothesis of ICT

Hypothesis. *A process Y is conscious if and only if it is a C-process of some process X. Also the content of consciousness $C_t^{Content}$ at time t is the state y_t of the C-process at time t and the level of consciousness C_t^{Level} is the degree of NTIC of the process to the environment i.e. $NTIC_t(E \to Y)$:*

$$C_t^{Content} = y_t \tag{13}$$

$$C_t^{Level} = NTIC_t(E \to Y) \tag{14}$$

A concrete example in the context of neuroscience is that X represents the microscopic scale of the universe, S a cellular scale process in the neural system, Y a more macroscopic process of the neural system coarse-grained from the cellular scale process S, and E the environment which the cellular level process S interacts with. The environment E may include other processes in the neural system, the sensors for perception and interoception, and external physical worlds.

Based on the hypothesis, ICT leads to five core implications:

Implication 1. Consciousness is information. Here, "informative" refers to the resolution of uncertainty. Being in a certain conscious state rules out other possible conscious states. Therefore, every conscious percept resolves some amount of uncertainty and provides information. This implication is also in agreement with the "axiom" of *information* in Integrated Information Theory (IIT 3.0) which claims that "...an experience of pure darkness is what it is by differing, in its particular way, from an immense number of other possible experiences..." (Oizumi *et al.*, 2014, P. 2)

Implication 2. Consciousness is associated with physical substrates and the self-information of the conscious percept is equal to the self-information of the corresponding physical event. This is a direct implication from our hypothesis that every conscious percept $C_t^{Content}$ corresponds to a physical event y_t.

[2]Functions in the mathematical sense used here are always either one-to-one or many-to-one.

Implication 3. Conscious processes are self-determining. This is a direct implication of the requirement that Y is informationally closed with respect to X. To be informationally closed with respect to X, no coarse-graining knows anything about the conscious process' future that the conscious process does not know itself. This self-determining characteristics is also consistent with our daily life conscious experience which often shows stability and continuity and is ignorant of the stochasticity (e.g., noise) of the cellular scales.

Implication 4. Conscious processes encode the environmental influence on itself. This is due to the non-triviality of the informational closure of Y to E. At the same time all of this information is known to the conscious processes themselves since they are informationally closed with respect to their environments. This also suggests that conscious processes can model the environmental influence without knowing more information from the environment.

Implication 5. Conscious processes can model environmental information (by forming NTIC) but be ignorant to part of the information of more microscopic processes (from Implication 3 and 4). This is consistent with our conscious experience, namely that the information that every conscious percept provides represents rich and structured environmental states without involving all the information about microscopic activities.

4.1 Level of Consciousness is Equal to the Degree of NTIC of a Process

According to Eq. 8, ICT implies that conscious levels are determined by two quantities.

First, to form a high level of NTIC, one can increase the mutual information $I(Y_{t+1}; Y_t)$ between the current internal state Y_t and the future internal state Y_{t+1}. In other words, conscious levels are associated with the degree of self-predictive information (Bialek *et al.*, 2001). This mutual information term can be further decomposed to two information entropy quantities:

$$I(Y_{t+1}; Y_t) = H(Y_{t+1}) - H(Y_{t+1}|Y_t) \tag{15}$$

This implies that a highly NTIC process must have rich dynamics with self-predictability over time. Another implication is that complex systems can potentially attain higher levels of consciousness due to the greater information capacities needed to attain high mutual information. This outcome is consistent with the common intuition that conscious levels are often associated with the degree of complexity of a system.

Second, one can minimise the conditional mutual information $I(Y_{t+1}; Y_t|E_t)$ to increase the level of NTIC. If the mutual information term $I(Y_{t+1}; Y_t)$ is supposed to stay large, this quantity suggests that conscious level increases with the amount of information about the environment state E_t that the NTIC process encodes in its own state Y_t and Y_{t+1}. In other words, Y_t should not contain more information about Y_{t+1} than E_t. An important implication is that agents interacting with a complex environment have the chance to build a higher level of NTIC within their systems than those living in a simple environment. In other words, the level of consciousness is associated with environmental complexity.

It is important to note that NTIC can be a non-monotonic function of the scale of coarse-graining. At the finest scale we consider the whole universe X as the process Y. Then, since Y is a coarse-graining of S we have $Y = S = X$. In this case the environment E corresponding to the universe seen as a system is the constant coarse-graining[3] and therefore the mutual information $I(E_t; Y_{t+1})$ and the transfer entropy $I(Y_{t+1}; E_t|Y_t)$ are zero. The NTIC of the universe with respect to its environment is then zero, and X can never be a C-process.

If we now increase the scale of Y, this allows S to also reduce in scale and therefore E can become more and more fine-grained. This means that the mutual information $I(E_t; Y_{t+1})$ between E and Y can at least potentially become positive. Up to the point where E accounts for half of the bits of X and S for the other half the upper bound of the mutual information $I(E_t; Y_{t+1})$ achieved when $Y = S$ increases. Refining E even further again leads to a reduction of the upper bound of $I(E_t; Y_{t+1})$.

At the other extreme, when $E = X$ the system state space must be the singleton set and NTIC from E to Y must again be zero. Therefore, processes at intermediate scales of coarse-graining can form higher degrees of NTIC than those at the most microscopic or macroscopic scales (Fig. 4).

[3] Recall that, for a system with state space $\mathcal{S}$ the environment state space $\mathcal{E}$ must be such that $\mathcal{X} = \mathcal{S} \times \mathcal{E}$. If $\mathcal{S} = \mathcal{X}$ then we need $\mathcal{E}$ with $\mathcal{X} \times \mathcal{E} = \mathcal{X}$ such that $\mathcal{E}$ must be a singleton set. All coarse-grainings mapping $\mathcal{X}$ to a singleton set are constant over $\mathcal{X}$.

ICT suggests that human consciousness occurs at a scale of coarse-graining where high NTIC is formed within the neural system.

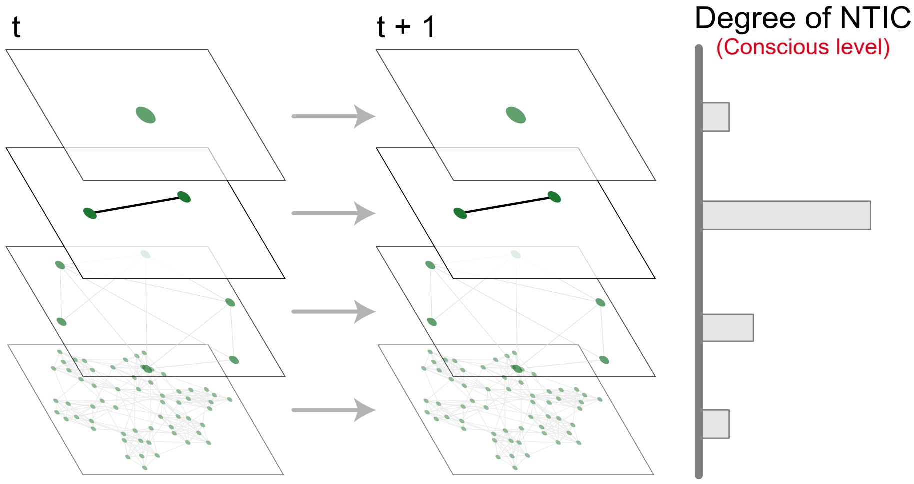

Figure 4: A non-monotonic relationship between the scale of coarse-graining and level of consciousness.

4.2 Conscious Contents Corresponding to States of an NTIC Process

ICT proposes that conscious contents correspond to the states of NTIC processes (Eq. 13). This implies that the size of the state space of an NTIC process is associated with the richness of the conscious contents that the process can potentially have. Accordingly, a complex NTIC process with a high dimensional state space can have richer conscious experience than a simple NTIC process. This outcome is consistent with the intuition that the richness of conscious contents is associated with the complexity of a system.

Informational closure can happen between scales of coarse-graining within a single system. Thus, a macroscopic NTIC process can be ignorant of its microscopic states. ICT argues that human conscious contents do not reflect cellular scale activity because the conscious process which corresponds to a macroscopic NTIC process is informationally closed to the cellular scale in the human neural system. Further more, since NTIC processes are informationally closed, each of them can be considered as a reality. When the information flow from its microscopic processes (and from the environment) to it is zero (Eq. 2), the future states of the process can be entirely self-determined by its past states.

Importantly, in most realistic cases, NTIC processes internalise the environmental dynamics in its states (see Sec. 2.1 and also Bertschinger *et al.* (2006)). This suggests that an NTIC process can be considered as a process that models the environmental dynamics. This implication fits well with several theories of consciousness (for example, world simulation metaphor (Revonsuo, 2006)). Note that ICT does not assume that generative models are necessary for consciousness. The implication is a natural result of processes with NTIC.

Finally, a coarse-graining can be a many-to-one map from microscopic to macroscopic states and ICT proposes that conscious contents $C^{Content}$ is the state of the NTIC process Y. ICT therefore implies the multiple realisation thesis of consciousness (Bechtel & Mundale, 1999; Putnam, 1967), which suggests that different physical implementations could map to the same conscious experience.

4.3 Reconciling the Levels and Contents of Consciousness

While it is useful to distinguish the levels and contents of consciousness at the notion level, whether they can be clearly dissociated has been a matter of debate (Bayne *et al.*, 2016; Fazekas & Overgaard, 2016). In ICT, conscious levels and conscious contents are simply two different properties of NTIC processes, and the two aspects of consciousness are therefore naturally reconciled. In an NTIC process with a large state space, conscious contents should also consist of rich and high dimensional information. This framework therefore integrates the levels and the contents of consciousness in a coherent fashion by providing explicit formal definitions of the two notions.

According to Sec. 4.1 and Sec. 4.2, an important implication from ICT is that both conscious levels and conscious contents are associated with the state space of an NTIC process Y. A larger state space of Y contributes conscious levels through the mutual information $I(Y_{t+1}; Y_t)$ and also contributes richer conscious contents by providing a greater number of possible states of conscious processes. ICT therefore explains why, in normal physiological states, conscious levels and conscious contents are often positively correlated (Laureys, 2005). This implication is also consistent with the intuition that consciousness is often associated with complex systems.

5 Conscious Versus Unconscious Processing

In this section, we show how ICT can explain and make predictions about which processes are more conscious than others. ICT is constructed using information theory and can provide predictions based on mathematical definitions.

5.1 Unconscious Processing

In this section we highlight two scenarios in which ICT predicts that processes remain unconscious.

Processes that are not Informationally Closed

The first scenario is built upon the assumption that sensor processes are non-deterministic [4] and that process dynamics are passively driven by environmental inputs. Such processes cannot be informationally closed and are, therefore, unconscious.

Reflexive behaviours (Casali *et al.*, 2013) can be considered an example of this scenario. In ICT, we can view reflexive behaviours as situations in which (Fig. 5) the internal state Y_t, which triggers reflexive action $\hat{y}_t$, is determined by the environment state E_{t-1}, overruling the influences from its own past Y_{t-1}. Such interpretation of reflexive behaviour from the viewpoint of ICT naturally explains why reflexes involve less or no conscious experience of external stimuli.

[4] Non-deterministic sensor processes here means $H(\hat{e}_{t+1}|\hat{e}_t) > 0$.

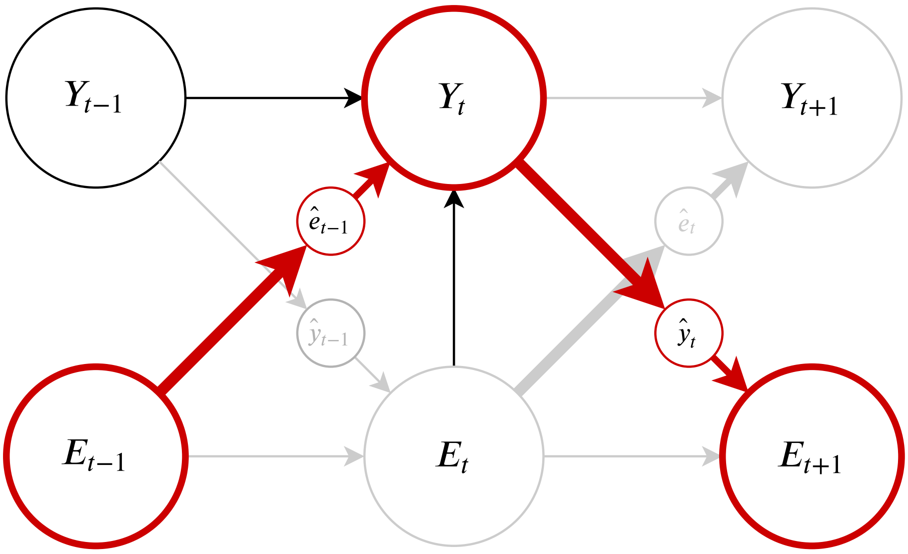

Figure 5: Schema depicting the information flow in reflexive behaviours (shown by the red nodes and arrows) happening through the interaction between a process Y and its environment E. When the sensor process $\hat{e}_t$ is non-deterministic and the internal state Y_t is mostly dependent on the sensor state $\hat{e}_t$ driven by the environment E_{t-1} but less on its past state Y_{t-1}, as a consequence, Y is unable to form informational closure and, therefore, remain unconscious.

The same principle can be applied to interpret blindsight (Humphrey, 1970, 1999, 1974) and procedural memory (Ashby *et al.*, 2010; Doyon *et al.*, 2009) which are often considered unconscious processes. Blindsight patients are able to track objects, avoid obstacles, and make above chance-level visual judgements with degraded or missing visual experience.(However, in some cases, they may still preserve some forms of conscious experience; See Mazzi *et al.* (2016); Overgaard (2011)). We argue that blindsight-guided actions are a result of stimulus-response mapping. The corresponding neural circuits are driven passively and therefore are not informationally closed. According to ICT we therefore have no conscious visual experience of visual stimuli.

Similarly, for procedural memory, the state transitions of corresponding neural circuits determining the action sequences largely depend on sensory inputs. This prevents the neural processes of procedural memory from informational closure and being conscious. ICT also offers an interpretation as to why patients with visual apperceptive agnosia (James *et al.*, 2003) can perform online motor controls without visual awareness of action targets (Whitwell *et al.*, 2014).

Note that, not all processes that are driven by the environment (passive adaptation) are unconscious. As mentioned in Sec. 2.1, when the sensor processes are deterministic, a system can still have positive NTIC and achieve informational closure via passive adaptation. Therefore, some passive system (for example pure feedforward networks) can potentially be conscious.[5]

For agents such as human beings, the environment is often informationally rich but only partially observable in such a way that the current sensory inputs are insufficient to predict the next inputs and to form deterministic sensor processes. In this situation, the system cannot become informationally closed by passive adaptation (e.g., simply copying the sensory values to the system). ICT predicts that, in most realistic cases, processes with passive adaptation are unconscious. On the other hand, networks with recurrent loops employing information stored in their own past states have the potential to achieve higher NTIC by modelling the environment. If it turns out to be true that for every pure feed-forward network there are non-feed-forward systems achieving higher NTIC, then ICT predicts that the latter systems achieve higher levels of consciousness. This implication coincides with theories of consciousness emphasising the importance of recurrent circuits to consciousness (Edelman, 1992; Lamme, 2006; Tononi & Koch, 2008).

[5]Since an n-layer feedforward network is a system with n-step memory it is technically appropriate to use the n-step memory definition of NTIC, i.e. $NTIC_t^m(E \to Y) := I(Y_{t+1} : E_t, \ldots, E_{t-n+1}) - I(Y_{t+1} : E_t, \ldots, E_{t-n+1} | Y_t)$ (Bertschinger *et al.*, 2006), for such systems. In this case the notion of non-deterministic input processes should be generalised to input processes with $H(\hat{e}_t | \hat{e}_{t-1}, \ldots, \hat{e}_{t-n}) > 0$.

Processes that are Trivially Closed

The second scenario is that when encoded information in a process is trivial, i.e. there is no mutual information between the process states and the environment states $I(Y_{t+1}; E_t)$ (Eq. 9), this leads to non-positive NTIC. In such cases, the process is considered to be unconscious. This implies that an isolated process which is informationally closed is insufficient to be conscious. This mathematical property of ICT is relevant for dealing with the boundary and individuality problems of consciousness[6] (Raymont & Brook, 2006). Consider an NTIC process Y and an isolated informationally closed process $\hat{Y}$ with only trivial information. Adding $\hat{Y}$ to Y can still maintain informational closure but does not increase non-trivial information, i.e., consciousness is unaffected.

$$\begin{aligned} I(Y, \hat{Y}; E) &= H(Y, \hat{Y}) - H(Y, \hat{Y}|E) \\ &= H(Y) + H(\hat{Y}|Y) - (H(Y|E) + H(\hat{Y}|Y, E)) \\ &= H(Y) + H(\hat{Y}) - (H(Y|E) + H(\hat{Y})) \\ &= H(Y) - H(Y|E) \\ &= I(Y; E) \end{aligned} \tag{16}$$

This implies that isolated processes with trivial information do not contribute consciousness and should be considered as being outside the informational boundary of the conscious processing. This property also implies that consciousnesses do not emerge from simple aggregation of informationally closed (isolated) processes which contain trivial information. In the future we hope to adapt the procedures for boundary detection proposed in Krakauer *et al.* (2014, 2020) to ICT.

5.2 Conscious Processing

In accordance with ICT, we claim that any process, system, or cognitive function which involves any NTIC process should be accompanied by conscious experience.

Previous consciousness research has identified a number of diverse cognitive processes which are often accompanied by conscious experience. ICT provides an integrated account of why these processes involve conscious experience. As mentioned above, an NTIC process can be seen as an internal modelling engine for agent-environmental interactions (Bertschinger *et al.*, 2006). Therefore, information encoded in NTIC processes is essential for several cognitive processes.

Among the most valuable types of information are predictions about environmental states. Cognitive functions requiring agent-scale environmental predictions are likely to recruit NTIC processes, and to therefore be accompanied by conscious experience; examples include planning and achieving long term goals.

Second, as a modelling engine, an NTIC process with a given initial state can self-evolve and simulate the environmental transitions. Cognitive functions involving simulations are expected to involve NTIC processes. Consequently, mental simulation, imagination, computing alternative realities, and generating counterfactuals often come with conscious experience.

Third, as an informationally closed system, an NTIC process can still provide environmental information without new sensory inputs. This is crucial for many types of off-line processing. Therefore, in contrast to reflexive-like behaviours, such as those mentioned above (Sec. 5.1), behaviours requiring off-line computations (Himmelbach & Karnath, 2005; Milner *et al.*, 1999; Revol *et al.*, 2003) often involve conscious experience.

Finally, for agents adapting to complex environments (e.g., human beings), any state of the NTIC process can be seen as an integration of high dimensional information. To accurately encode information about complex environmental states and transitions, the NTIC process requires knowledge about the complex causal dependencies involved in the environment. Cognitive functions requiring larger scale integration are therefore likely to involve NTIC processes and accompanied by conscious experience.

Note that many of the claims above are compatible with several theories of consciousness which highlight the connection between consciousness and internal simulation, predictive mechanism, or generative models inside a system (e.g. world simulation metaphor (Revonsuo, 2006), predictive

[6] The boundary problem of consciousness refers to identifying physical boundaries of conscious processes and the individuality problem of consciousness refers to identifying individual consciousnesses in the universe.

processing and Bayesian brain (Clark, 2013; Hohwy, 2013; Seth, 2014), generative model and information generation (Kanai *et al.*, 2019)). Instead of relating functional or mechanistic aspects of a system to consciousness, ICT captures common informational properties underlying those cognitive functions which are associated with consciousness. As such, ICT does not assume any functionalist perspectives of consciousness, which associate specific functions to consciousness. That is to say, since ICT associates information with consciousness, functional features accompanied by consciousness are collateral consequences of neural systems which utilise NTIC processes for adaptive functions.

In sum, we argue that cognitive functions involving the NTIC process are inevitably accompanied by consciousness. Having an NTIC process is potentially an effective approach to increasing fitness in the evolutionary process. It is likely that biological creatures evolve NTIC processes at some point during their evolution. Due to the fundamental relation between information and consciousness, biological creatures also evolve different degrees of consciousness depending on the physical scale and complexity of the environments they adapt to.

Although it starts with a non-functional hypothesis, ICT accounts for the association between function and consciousness. Further, ICT demonstrates remarkable explanatory power for various findings concerning conscious and unconscious processing.

6 Comparison with Other Relevant Theories of Consciousness

In this section, we compare ICT with other relevant theories of consciousness.

6.1 Multilevel Views on Consciousness and Cognition

ICT proposes that conscious processes can occur at any scale of coarse-graining which forms NTIC within a system. This suggests that the scale of coarse-graining is critical for in searching for and identifying the information corresponding to consciousness. A few number of versions of multilevel views on consciousness have previously been (explicitly or implicitly) proposed. To our knowledge, Pennartz's neurorepresentational theory (also called Neurorepresentationalism, (Pennartz, 2015, 2018)) is closest to the multilevel view of ICT. Similar to Neurorepresentationalism, the concept of levels in ICT is relevant to Marr's level of analysis (Marr, 1982; Pennartz, 2015, 2018). However, ICT suggests that coarse-graining is necessary only when the microscopic processes are not NTIC (e.g. the neural system). An NTIC process can be formed in a noise-free deterministic system without coarse-graining. According to ICT, this NTIC process is sufficient for consciousness. Another fundamental difference between ICT and Neurorepresentationalism is that Neurorepresentationalism takes a functionalist perspective and suggests that consciousness should serve high-level world-modelling and make a best guess about the interaction between the body and the environment. In contrast, however, ICT is grounded by a non-functional informational hypothesis. Therefore, ICT provides a non-functional and fundamental explanation for the scale problem of consciousness.

Another well-known proposal based on multilevel views is the Intermediate Level Theory of Consciousness (Jackendoff, 1987; Prinz, 2007, ILT). ILT proposes that conscious experience is only associated with neural representations at intermediate **levels of the sensory processing hierarchy** (e.g., the 2.5D representation of visual processing), and not with lower (e.g., pixel) or higher (e.g., abstract) levels of the sensory hierarchy.

Here, we want to make clear that "level" in ICT refers to the **scale of coarse-graining**, rather than "level" in cortical anatomy or sensory processing. It is important to note that the coarse-graining direction is an orthogonal dimension irrespective of the level of anatomy or of information processing hierarchy in the neural system (see Fig. 6). Because ILT focuses the levels of the sensory processing hierarchy and ICT focus on informational closure among the levels of coarse-graining, the two theories are fundamentally different.

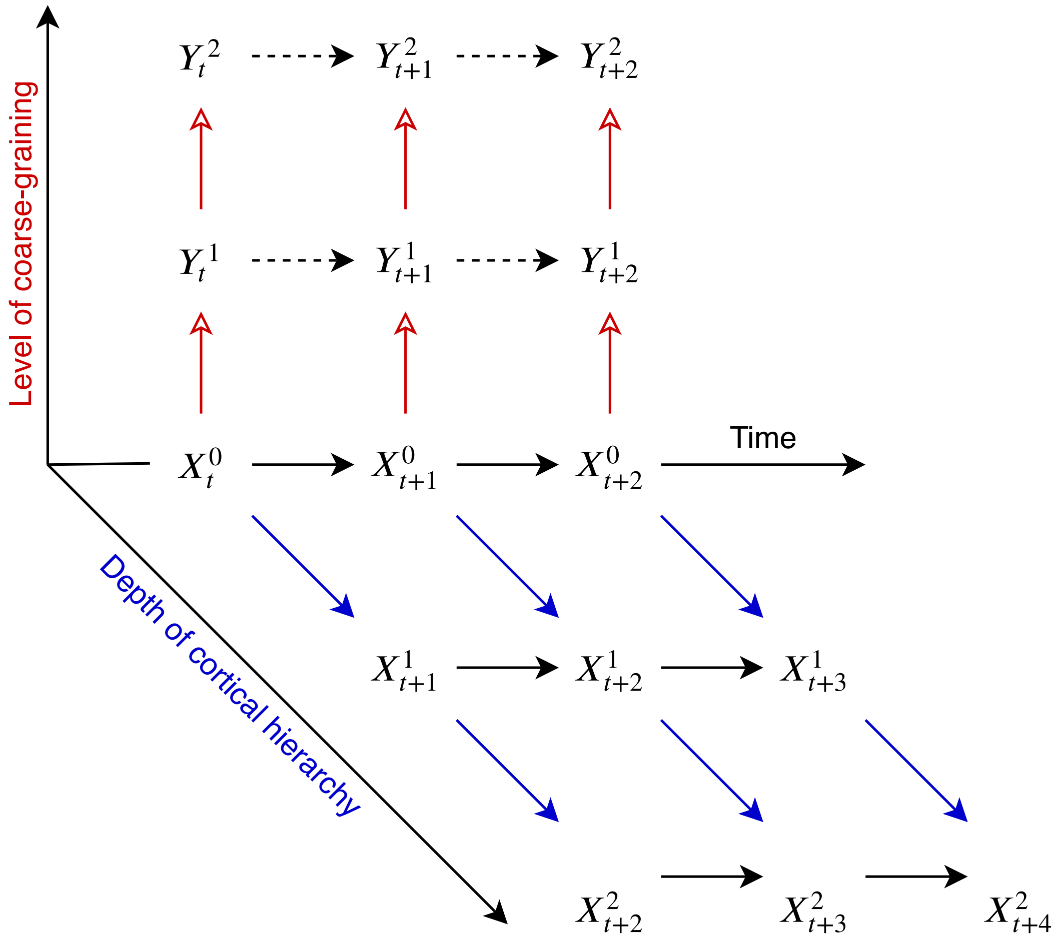

Figure 6: Distinction between the level of coarse-graining and the level of cortical hierarchy. X and Y represent the microscopic and macroscopic coarse-grained variables, respectively. X^0 represents microscopic states upstream of the cortical hierarchy. The red empty arrows represents the directions of coarse-graining and the blue arrows represent the directions of the physical dependencies in the cortical hierarchy from upstream to downstream. (Some variables and dependencies are omitted for clarity.)

6.2 Integrated Information Theory

Integrated information theory (IIT) states that consciousness is integrated information and that a system's consciousness is determined by its causal properties (Tononi *et al.*, 2016). ICT is consistent with IIT in that informational properties are thought to underlie consciousness. In this section, we will discuss ICT in the light of IIT.

The concept of "information": In IIT, information refers to "integrated information", namely "Information that is specified by a system that is irreducible to that specified by its parts." (Tononi *et al.* 2016) In ICT, information refers to "self-information", i.e. information about the states of conscious experience and the physical states of a process. Therefore, IIT focuses more on the relationships between consciousness and causal interactions among elements within a system, whereas ICT focuses more on the informational relationships between conscious experience and being in a certain state of a process.

The "Exclusion" axiom in IIT: In IIT, the Exclusion axiom claims that among all overlapping sets of elements, only one set, having maximal integrated information, can be conscious. The exclusion axiom should be applied over elements, space, time, and scales (Hoel *et al.*, 2016; Oizumi *et al.*, 2014). Differing from IIT, ICT allows multiple consciousnesses to coexist across different scales of coarse-graining within a system if they are informationally closed from to each other. The two distinctive predictions decisively pinpoint the core concepts of the two theories.

The concept of "integration": In IIT, integrated information is a core concept in defining conscious individuals. In the present paper, we do not include the notion of integrated information within ICT. However, this represents one of the current weaknesses of ICT, namely that it in some cases it lacks the ability to individuate NTIC processes (i.e., the problem of individuality). We discuss this weaknesses in Sec. 7.

Prediction after system damage: Prediction after system damage: ICT and IIT lead to different predictions when a system suffers from damage. Consider for example a densely connected network whose dynamics forms an NTIC process. If we cut the network in half, IIT predicts that this would result in two consciousnesses because elements in both networks still maintain high degrees of interaction. In contrast, ICT would predict that this operation might completely destroy informational closure of the network, and thereby render both parts unconscious. Nevertheless, this prediction is relatively premature. In the future, rigorous modelling studies will allow systematic comparisons between model predictions.

6.3 Predictive Processing

Predictive processing (PP) is a powerful framework which integrates several ideas from neuroscience. This emerging theoretical framework posits that neural systems constantly generate predictions about incoming sensory signals and update predictions based on prediction errors between predictions and sensory signals. According to PP, neural systems constantly perform unconscious statistical inference about hidden causes in the external environment. The perceptual contents are the "best guess" about those environment states which include these hidden causes (Clark, 2013; Hohwy, 2013). PP is well integrated with Bayesian brain hypothesis and has been used to interpret conscious perception in many domains (Hohwy, 2013; Seth, 2014).

PP is a powerful explanatory framework for diverse brain functions. However, to serve as a theory of consciousness, PP is still incomplete due to two explanatory gaps. First, the neural system is equipped with multiple predictive mechanisms, but it appears that not all of these predictive mechanisms are involved in conscious processes (e.g. mismatch negativity, Näätänen *et al.* (2007)). PP needs to explain the difference between conscious and unconscious predictive mechanisms.

Second, PP can be considered as a sophisticated computation for perceptual inference. It takes von Helmholtz's conception of perception as unconscious inference. Thus, only the most probable outcome computed by the inference processes can be conscious, while other details of the computation remain unconscious. PP also needs to explain how unconscious inferences are able to give rise to conscious results. In short, while PP is often discussed in the context of consciousness, these explanatory gaps prevent PP from being a theory of consciousness.

ICT is well compatible with PP. Crucially, ICT further provides natural and fundamental explanations to fill the two explanatory gaps which hamper PP. According to the definition of NTIC, a process with high NTIC can be regarded as a powerful predictive machine which has accurate self-predictive information ($I(Y_{t+1}; Y_t)$, E.q. 6) and concurrently incorporates environmental information into its dynamic ($I(Y_{t+1}; Y_t|E_t)$, E.q. 6). This predictive nature of NTIC processes is in agreement with the core notion of PP in which the conscious contents are always the predicted (inferred) outcome of our predictive mechanisms. Second, due to the informational closure to the environment, the encoded information about its environment in an NTIC process can appear to be as "the best guess" about the external environment in the context of Bayesian inference.

Finally, therefore, why is some predictive information conscious and some are not? ICT predicts that only the predictions generated from mechanisms involving the NTIC process are conscious. Note that it is not necessary for predictive processes to involve NTIC processes. A predictive process can make a prediction about the future state of its environment solely based on the current sensor states when the current sensor states and future sensor states have positive mutual information. However, this is not sufficient for a process to be informationally closed and, therefore, be conscious.

Also in accordance with ICT, we further propose that we can only be aware of the predictions of predictive processes due to informational closure to computational details of microscopic predictive processes. Acquisition by the macroscopic NTIC process is limited to the coarse-grained summary statistics of the microscopic processes. In other words, we predict that the computation of the statistical inferences of PP is implemented at microscopic (cellular) scales in the neural system.

Finally, we consider that PP is a potential empirical implementation of NTIC processes. To maintain accurate information about the environment encoded in an NTIC process, one can open an information channel between the process and the environment to allow the minimal flow of information required to correct the divergence between them. This proposal is compatible with PP, which suggests that PP systems update (correct) the current estimations by computing prediction errors between predicted and real sensory inputs.

6.4 Sensorimotor Contingency

The sensorimotor contingency (SMC) theory of consciousness proposes that different types of SMCs give rise to different characteristics of conscious experience (O'Regan & Noë, 2001). The theory radically rejects the view that conscious content is associated with the internal representations of a system. Rather, the quality of conscious experience depends on the agent's mastery of SMCs. SMC emphasizes that the interaction between a system and its environment determines conscious experience.

ICT is not compatible with SMC. As mentioned in Sec. 5, a process which directly maps the sensory states to the action states is insufficient to be NTIC. Therefore, learning contingencies between sensory inputs and action outputs do not imply NTIC. Hence, ICT predicts that having sensorimotor contingencies is neither a necessary nor a sufficient condition for consciousness. In fact, empirically, with extensive training on a sensorimotor task with a fixed contingency, the task can be gradually performed unconsciously. This indicates that strong SMCs do not contribute conscious contents. In contrast, ICT suggests that, with extensive training, the neural system establishes a neural mapping from sensory inputs to action outputs. This decreases the level of informational closure and, as a result, decrease the consciousness level of this process. This outcome better supports ICT than SMC.

Nevertheless, ICT does appreciate the notion that interactions between a process and its environment are crucial to shaping conscious experience. As mentioned above, to form NTIC, a process needs to encode environmental transitions into its own dynamic. Therefore, information of agent-environment interaction should also be encoded in the NTIC process, and thereby shape conscious contents in a specific way.

Different to classical SMC, a new version of SMC proposed by Seth (2014, 2015), namely Predictive Processing of SensoriMotor Contingencies (PPSMC), combines SMC and the predictive processing framework together. PPSMC emphasises the important role of generative models in computing counterfactuals, inferring hidden causes of sensory signals, and linking fictive sensory signals to possible actions. According to ICT, if the generative model involves the NTIC process in the computation of counterfactuals, PPSMC will be compatible with our theory and may have strong explanatory power for some specific conscious experience.

6.5 Global Workspace Theory

Global workspace theory (GWT; Baars (1988, 1997, 2002)) and Global Neuronal Workspace theory (GNWT; Dehaene & Changeux (2011); Dehaene & Naccache (2001); Dehaene *et al.* (1998)) state that the neural system consists of several specialised modules, and a central global workspace (GW) which integrates and broadcasts information gathered from these specialised modules. Only the information in the global workspace reaches conscious awareness, while information outside of it remains unconscious. These modules compete with each other to gain access to the GW, and the information from the winner triggers an all-or-none "ignition" in the GW. Information in the GW is broadcast to other modules. Conscious contents are then associates with the information that gains access to the internal global workspace (Dehaene *et al.*, 2017).

While GWT emphasises the importance of global information sharing as a basis of consciousness, the precise meaning of information broadcasting remains somewhat unclear if one tries to describe it more formally in the language of information theory. ICT offers one possible way to consider the meaning of broadcasting in GWT. Specifically, one could interpret the global workspace as the network of nodes wherein information is shared at the scale of NTIC and where communication is performed through macro-variables that are linked via mutual predictability. In other words, the global workspace should also be NTIC. While this link remains speculative, this interpretation encourages empirical studies into the relationship between the contents of consciousness and macrostate neural activities that are mutually predictive of each other.

7 Limitations and Future Work

As a completely new theory of consciousness, ICT is still far from completion. In the following, we discuss the current limitations and challenges of ICT and point out some potential future research directions.

It is important to clarify that ICT does not intend to solve the hard problems of consciousness (Chalmers, 1995). Knowing the state of a conscious process does not allow us to answer "What is it like to be in this state of this process" (Nagel, 1974). Instead, ICT focuses more on bridging consciousness and the physical world using information theory as a common language between them.

The current version of ICT cannot entirely solve the problem of individuality. The main issue with identifying individual consciousnesses using ICT is that at the moment the environment is not uniquely defined. Once we have identified processes that are informationally closed with respect to X we still have to find the environment process E with respect to which we compute NTIC. However, there are usually multiple system processes S of which a given Y is a coarse-graining in which case there are also multiple environment processes E with respect to which we could compute NTIC.

A more general problem of NTIC-based individuality is that we can define a new process Y and also its environment E by recruiting two independent NTIC processes Y^1 & Y^2 and their environments E^1 & E^2, respectively. Accordingly, $Y = (Y^1, Y^2)$ and $E = (E^1, E^2)$. In such a case, the new process Y will also be NTIC to E. The current version of ICT is therefore unable to determine whether there are two smaller consciousnesses or one bigger consciousness (or for that matter 3 coexisting consciousnesses). The problem of individuality is a significant theoretical weakness of the current version of ICT. The notion of integration is a possible remedy for this issue, and we will address it explicitly in our future work using the concept of synergy.

The current version of ICT assumes that consciousness receives contribution from only non-trivial information, rather than trivial information encoded in a process. In other words, the amount of information about environmental states and dynamics encoded in a process is a key quantity for consciousness. However, we do not exclude the possibility that environmental information may simply be a proxy for other informational quantities. More theoretical work is needed to elucidate the role of environments. This issue will also be discuss in our future theoretical paper.

In this article, we do not use a state-dependent formulation of NTIC. However, we believe that state-dependent NTIC is essential to describing the dynamics of conscious experience. The next version of ICT therefore requires further research using point-wise informational measures to construct state-dependent NTIC.

Explaining conscious experience during dreaming is always a challenge to theories of consciousness. ICT currently does not have a specific answer to dreaming. However, we wish to emphasize that not all processes in the neural system are NTIC since some processes are not informationally closed. They mainly passively react to sensory inputs or other processes in the neural system. To the conscious (NTIC) process, the rest of the neural system and the body should also be considered as part of the environment. They retain some degree of activity during sleep and dreaming. We speculate that, during dreaming, the neural system stably forms an NTIC process with respect to its environment, i.e. the other parts of the neural system. At present, however, this remains mere speculation. Identification of the NTIC process(es) during dreaming is an important milestone in extending the scope of ICT.

Empirically, a major challenge to ICT is to find appropriate coarse-graining functions which map microscopic processes to macroscopic NTIC processes. This issue will become imperative in the search for neurological evidence supporting ICT. Identifying such coarse-graining functions among infinite candidates (Price & Corry, 2007) appears to be very challenging. Nevertheless, recent theoretical and technical progress may contribute to solving this issue. For example, the concept of *causal emergence* proposed by Hoel (Hoel, 2018; Hoel *et al.*, 2013) has been further developed recently. Causal emergence is highly relevant to the relationship between informational closure and coarse-graining. In their new study, Klein & Hoel (2019), start to compare how different coarse-graining functions influence causal emergence at macroscopic scales. Pfante *et al.* (2014a,b) provide a thorough mathematical analysis of level identification, including informational closure. In neuroscience, an understanding of neural population codes has also made a tremendous progress due to advance in recording technique and data science (Kohn *et al.*, 2016; Panzeri *et al.*, 2015). Gamez (2016) has also systematically described relevant issues in finding data correlates of consciousness among different levels of abstraction. We believe that interdisciplinary research is required to narrow down the scope of search for coarse-graining functions and conscious processes at macro-scales in the neural system and beyond.

Finally, another empirical challenge to ICT is that of empirical supporting evidence. This is understandable because the concept of NTIC is relatively new in the history of information science, not to mention in neuroscience. Very few experiments and data collections examining NTIC properties in neural systems have yet appeared. To our knowledge, only two studies (Palmer *et al.*, 2015; Sederberg *et al.*, 2018) coincidentally examined relevant properties in salamander retina; these found that a large group of neural populations of retinal ganglion cells encoded predictive information about external stimuli and also had high self-predictive information about their own future states. This result is consistent with the characteristic of NTIC. We expect that there will be more empirical studies examining relevant neural properties of NTIC.

8 Conclusions

In this paper, we introduce the **Information Closure Theory of Consciousness (ICT)**, a new informational theory of consciousness. ICT proposes that a process which forms **non-trivial informational closure (NTIC)** is conscious and through coarse-graining the neural system can form NTIC processes, i.e., conscious processes, at certain macroscopic scales. ICT considers that information is a common language to bridge the gap between conscious experience and physical reality. Using information theory, ICT proposes computational definitions for both conscious level and conscious content. This allows ICT to be generalised to any system beyond the human brain.

ICT provides an explanation for various findings from research into conscious and unconscious processing. The implications of ICT indicate that the scales of coarse-graining play a critical role in the search for neural substrates of consciousness. Improper measurement of neurophysiological signals, such as those which are excessively fine or coarse in scale, may lead to misleading results and misinterpretations.

ICT reconciles several theories of consciousness. ICT indicates that they conditionally coincide with ICT's implications and predictions but, however, not the fundamental and sufficient conditions for consciousness. Example theories include those which emphasise recurrent circuits (Edelman, 1992; Lamme, 2006); highlight the internal simulation, predictive mechanisms, and generative models (Clark, 2013; Hohwy, 2013; Kanai *et al.*, 2019; Revonsuo, 2006; Seth, 2014, 2015); and relate to multilevel view of consciousness (Jackendoff, 1987; Pennartz, 2015, 2018; Prinz, 2007). Notably, while ICT is proposed based on the non-functional hypothesis, its implications for the functional aspects of a system fit several functionalist proposals well.

Regarding philosophy of mind, ICT connects several distinct arguments together. First, ICT can be seen as an identity theory because it assumes a fundamental relation between consciousness and information. Second, the implications of ICT tightly link consciousness to several cognitive functions in the context of evolution. This explains why people might intuitively have a functionalist point of view of consciousness. ICT emphasises that informational closure between scales of coarse-graining is critical to form NTIC processes in some stochastic systems. In this case, especially for the neural system, forming conscious processes at macroscopic scales coincides with the perspective of emergentism. Finally, forming NTIC (conscious) processes through many-to-one maps, i.e., coarse-graining, implies multiple realisability of consciousness. As a result, ICT provides an integrated view for these arguments and is further capable of indicating how and why they are conditionally true.

The current version of ICT is still far from completion, and several outstanding issues mandate further theoretical and empirical research. Nevertheless, ICT offers an explanation and a prediction for consciousness science. We hope that ICT will provide a new way of thinking about and understanding of neural substrates of consciousness.

Acknowledgements

A.C., Y.Y, and R.K. are funded by the Japan Science and Technology Agency (JST) CREST project. Work by M.B. and R.K. on this publication was made possible through the support of a grant from Templeton World Charity Foundation, Inc. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of Templeton World Charity Foundation, Inc. This manuscript has been released as a Pre-Print at arXiv (Chang *et al.*, 2019).

Author Contributions Statement

A.C. conceived and developed the theory. M.B. and A.C. contributed the mathematical formalisation of the theory. A.C., M.B, and R.K wrote the manuscript, based on a first draft by A.C. with extensive comments from Y.Y. All authors contributed to manuscript revision, read and approved the submitted version.

Conflict of Interest Statement

All authors were employed by Araya Inc. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix

Let us assume that the system only observes a part of the environment state.

We can represent the part of the environment that we observe by the value of a function f applied to the environment state. In this case we get for the transfer entropy

$$I(S_{t+1} : E_t|S_t) = I(S_{t+1} : f(E_t)|S_t). \tag{17}$$

If the system only copies the observation we then get for the transfer entropy

$$I(S_{t+1} : f(E_t)|S_t) = I(f(E_t) : f(E_t)|f(E_{t-1})) = H(f(E_t)|f(E_{t-1})) \tag{18}$$

and for the mutual information

$$I(S_{t+1}; E_t) = I(f(E_{t+1}); E_t) = H(f(E_t)) \tag{19}$$

such that

$$NTIC_t(E \to S) = I(f(E_t); f(E_{t-1})). \tag{20}$$

This shows that whenever there is mutual information between subsequent observations a process that only copies the observations has positive NTIC. Note that any additional (internal) processing of the observation without reference to an additional internal state using a function g can only reduce this mutual information:

$$I(f(E_t); g(f(E_{t-1}))) \leq I(f(E_t); f(E_{t-1})). \tag{21}$$

However, ignoring restrictions due to a possibly fixed choice of the universe process X we find that for each such system there are other systems that achieve higher NTIC. For example, if we define the system to be the "mirrored" and synchronized environment by setting $S_t := E_t$, then the transfer entropy vanishes

$$I(S_{t+1} : E_t|S_t) = I(E_{t+1} : E_t|E_t) = 0 \tag{22}$$

and the mutual information is equal to the mutual information between the current and next environment state:

$$I(S_{t+1}; E_t) = I(E_{t+1}; E_t). \tag{23}$$

In cases where the environment has itself higher predictive mutual information than the observations it produces - in other words, when

$$I(E_{t+1}; E_t) \geq I(f(E_{t+1}); f(E_t)) \tag{24}$$

there is then potential for a predictive process to achieve higher NTIC than a copying system or any system that only processes its last observations without taking account of other internal memory (i.e. those systems also applying g to their observations). Note that this also holds true in cases where the observations are themselves closed. If there is a more complex environment behind them, the mirrored and synchronised system has higher NTIC with respect to that environment than the system copying the observations.

References

Adrian, Edgar D. 1926. The impulses produced by sensory nerve endings. *The Journal of physiology*, **61**(1), 49–72.

Ashby, F Gregory, Turner, Benjamin O, & Horvitz, Jon C. 2010. Cortical and basal ganglia contributions to habit learning and automaticity. *Trends in cognitive sciences*, **14**(5), 208–215.

Baars, Bernard J. 1988. *A cognitive theory of consciousness.* Cambridge University Press.

Baars, Bernard J. 1997. In the theatre of consciousness. Global workspace theory, a rigorous scientific theory of consciousness. *Journal of Consciousness Studies*, **4**(4), 292–309.

Baars, Bernard J. 2002. The conscious access hypothesis: origins and recent evidence. *Trends in cognitive sciences*, **6**(1), 47–52.

Bayne, Tim, Hohwy, Jakob, & Owen, Adrian M. 2016. Are there levels of consciousness? *Trends in cognitive sciences*, **20**(6), 405–413.

Bechtel, William, & Mundale, Jennifer. 1999. Multiple realizability revisited: Linking cognitive and neural states. *Philosophy of science*, **66**(2), 175–207.

Bertschinger, Nils, Olbrich, Eckehard, Ay, Nihat, & Jost, Jürgen. 2006. Information and closure in systems theory. *Pages 9–21 of: Explorations in the Complexity of Possible Life. Proceedings of the 7th German Workshop of Artificial Life.*

Bialek, William, Nemenman, Ilya, & Tishby, Naftali. 2001. Predictability, complexity, and learning. *Neural computation*, **13**(11), 2409–2463.

Binder, Marc D, Hirokawa, Nobutaka, & Windhorst, Uwe. 2009. Encyclopedia of neuroscience.

Casali, Adenauer G, Gosseries, Olivia, Rosanova, Mario, Boly, Mélanie, Sarasso, Simone, Casali, Karina R, Casarotto, Silvia, Bruno, Marie-Aurélie, Laureys, Steven, Tononi, Giulio, *et al.* 2013. A theoretically based index of consciousness independent of sensory processing and behavior. *Science translational medicine*, **5**(198), 198ra105–198ra105.

Chalmers, David J. 1995. Facing up to the problem of consciousness. *Journal of consciousness studies*, **2**(3), 200–219.

Chang, Acer Y. C., Biehl, Martin, Yu, Yen, & Kanai, Ryota. 2019. *Information Closure Theory of Consciousness.*

Clark, Andy. 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, **36**(3), 181–204.

Dehaene, Stanislas, & Changeux, Jean-Pierre. 2011. Experimental and theoretical approaches to conscious processing. *Neuron*, **70**(2), 200–227.

Dehaene, Stanislas, & Naccache, Lionel. 2001. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, **79**(1-2), 1–37.

Dehaene, Stanislas, Kerszberg, Michel, & Changeux, Jean-Pierre. 1998. A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the national Academy of Sciences*, **95**(24), 14529–14534.

Dehaene, Stanislas, Lau, Hakwan, & Kouider, Sid. 2017. What is consciousness, and could machines have it?

Doyon, Julien, Bellec, Pierre, Amsel, Rhonda, Penhune, Virginia, Monchi, Oury, Carrier, Julie, Lehéricy, Stéphane, & Benali, Habib. 2009. Contributions of the basal ganglia and functionally related brain structures to motor learning. *Behavioural brain research*, **199**(1), 61–75.

Edelman, Gerald M. 1992. *Bright air, brilliant fire: On the matter of the mind.* Basic books.

Faisal, A Aldo, Selen, Luc PJ, & Wolpert, Daniel M. 2008. Noise in the nervous system. *Nature reviews neuroscience*, **9**(4), 292.

Fazekas, Peter, & Overgaard, Morten. 2016. Multidimensional Models of Degrees and Levels of Consciousness. *Trends in Cognitive Sciences*, **20**(10), 715–716.

Fekete, Tomer, & Edelman, Shimon. 2011. Towards a computational theory of experience. *Consciousness and cognition*, **20**(3), 807–827.

Fekete, Tomer, & Edelman, Shimon. 2012. The (lack of) mental life of some machines. *Being in Time: Dynamical Models of Phenomenal Experience*, **88**, 95–120.

Fekete, Tomer, van Leeuwen, Cees, & Edelman, Shimon. 2016. System, subsystem, hive: boundary problems in computational theories of consciousness. *Frontiers in psychology*, **7**, 1041.

Gamez, David. 2016. Are Information or Data Patterns Correlated with Consciousness? *Topoi*, **35**(1), 225–239.

Gerstner, Wulfram, & Kistler, Werner M. 2002. *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge university press.

Goldwyn, Joshua H., & Shea-Brown, Eric. 2011. The what and where of adding channel noise to the Hodgkin-Huxley equations. *PLoS Computational Biology*, **7**(11).

Himmelbach, Marc, & Karnath, Hans-Otto. 2005. Dorsal and ventral stream interaction: contributions from optic ataxia. *Journal of Cognitive Neuroscience*, **17**(4), 632–640.

Hoel, Erik P. 2018. *Agent Above, Atom Below: How Agents Causally Emerge from Their Underlying Microphysics*. Cham: Springer International Publishing. Pages 63–76.

Hoel, Erik P., Albantakis, Larissa, & Tononi, Giulio. 2013. Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, **110**(49), 19790–19795.

Hoel, Erik P, Albantakis, Larissa, Marshall, William, & Tononi, Giulio. 2016. Can the macro beat the micro? Integrated information across spatiotemporal scales. *Neuroscience of Consciousness*, **2016**(1), niw012.

Hohwy, Jakob. 2013. *The Predictive Mind*. Oxford University Press. 187 cites:.

Humphrey, N. K. 1970. What the Frog's Eye Tells the Monkey's Brain. *Brain, Behavior and Evolution*, **3**(1-4), 324–337.

Humphrey, Nicholas. 1999. *A History of the Mind: Evolution and the Birth of Consciousness*. Springer Science & Business Media.

Humphrey, Nicholas K. 1974. Vision in a monkey without striate cortex: a case study. *Perception*, **3**(3), 241–255.

Jackendoff, Ray. 1987. *Consciousness and the computational mind*. The MIT Press.

James, Thomas W, Culham, Jody, Humphrey, G Keith, Milner, A David, & Goodale, Melvyn A. 2003. Ventral occipital lesions impair object recognition but not object-directed grasping: an fMRI study. *Brain*, **126**(11), 2463–2475.

Kanai, Ryota, Chang, Acer, Yu, Yen, de Abril, Ildefons M, Biehl, Martin, & Guttenberg, Nicholas. 2019 (Jul). *Information Generation as a Functional Basis of Consciousness*.

Kandel, Eric R, Schwartz, James H, Jessell, Thomas M, of Biochemistry, Department, Jessell, Molecular Biophysics Thomas, Siegelbaum, Steven, & Hudspeth, AJ. 2000. *Principles of neural science*. Vol. 4. McGraw-hill New York.

Klein, Brennan, & Hoel, Erik. 2019. Uncertainty and causal emergence in complex networks. *arXiv preprint arXiv:1907.03902*.

Kohn, Adam, Coen-Cagli, Ruben, Kanitscheider, Ingmar, & Pouget, Alexandre. 2016. Correlations and Neuronal Population Information. *Annual Review of Neuroscience*, **39**(1), 237–256.

Krakauer, David, Bertschinger, Nils, Olbrich, Eckehard, Ay, Nihat, & Flack, Jessica C. 2014. The information theory of individuality. *arXiv preprint arXiv:1412.2447*.

Krakauer, David, Bertschinger, Nils, Olbrich, Eckehard, Flack, Jessica C, & Ay, Nihat. 2020. The information theory of individuality. *Theory in Biosciences*, 1–15.

Kristan Jr, William B, & Shaw, Brian K. 1997. Population coding and behavioral choice. *Current opinion in neurobiology*, **7**(6), 826–831.

Lamme, Victor AF. 2006. Towards a true neural stance on consciousness. *Trends in cognitive sciences*, **10**(11), 494–501.

Laureys, Steven. 2005. The neural correlate of (un) awareness: lessons from the vegetative state. *Trends in cognitive sciences*, **9**(12), 556–559.

Lemon, RN, & Edgley, SA. 2010. Life without a cerebellum. *Brain*, **133**(3), 652–654.

Luhmann, Niklas. 1995. Probleme mit operativer Schließung. *Soziologische Aufklärung*, **6**, 12–24.

Maass, Wolfgang, & Bishop, Christopher M. 2001. *Pulsed neural networks*. MIT press.

Marr, David. 1982. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. *New York, NY: W.H. Freeman and Company*, **8**(11).

Mathis, Donald W, & Mozer, Michael C. 1995. On the computational utility of consciousness. *Pages 11–18 of: Advances in neural information processing systems*.

Maturana, Humberto R, & Varela, Francisco J. 1991. *Autopoiesis and cognition: The realization of the living*. Vol. 42. Springer Science & Business Media.

Mazzi, Chiara, Bagattini, Chiara, & Savazzi, Silvia. 2016. Blind-sight vs. degraded-sight: different measures tell a different story. *Frontiers in psychology*, **7**, 901.

Milner, AD, Paulignan, Y, Dijkerman, HC, Michel, F, & Jeannerod, M. 1999. A paradoxical improvement of misreaching in optic ataxia: new evidence for two separate neural systems for visual localization. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **266**(1434), 2225–2229.

Näätänen, Risto, Paavilainen, Petri, Rinne, Teemu, & Alho, Kimmo. 2007. The mismatch negativity (MMN) in basic research of central auditory processing: a review. *Clinical neurophysiology*, **118**(12), 2544–2590.

Nagel, Thomas. 1974. What is it like to be a bat? *The philosophical review*, **83**(4), 435–450.

Oizumi, Masafumi, Albantakis, Larissa, & Tononi, Giulio. 2014. From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. *PLoS computational biology*, **10**(5), e1003588.

O'Regan, J Kevin, & Noë, Alva. 2001. A sensorimotor account of vision and visual consciousness. *Behavioral and brain sciences*, **24**(5), 939–973.

Overgaard, Morten. 2011. Visual experience and blindsight: a methodological review. *Experimental Brain Research*, **209**(4), 473–479.

Overgaard, Morten, & Overgaard, Rikke. 2010. Neural correlates of contents and levels of consciousness. *Frontiers in psychology*, **1**, 164.

Palmer, Stephanie E., Marre, Olivier, Berry, Michael J., & Bialek, William. 2015. Predictive information in a sensory population. *Proceedings of the National Academy of Sciences*, **112**(22), 6908–6913.

Panzeri, Stefano, Macke, Jakob H, Gross, Joachim, & Kayser, Christoph. 2015. Neural population coding: combining insights from microscopic and mass signals. *Trends in cognitive sciences*, **19**(3), 162–172.

Pattee, Howard Hunt. 2012. Evolving self-reference: matter, symbols, and semantic closure. *Pages 211–226 of: Laws, language and life.* Springer.

Pennartz, Cyriel MA. 2015. *The brain's representational power: on consciousness and the integration of modalities.* MIT Press.

Pennartz, Cyriel MA. 2018. Consciousness, representation, action: the importance of being goal-directed. *Trends in cognitive sciences*, **22**(2), 137–153.

Pfante, Oliver, Olbrich, Eckehard, Bertschinger, Nils, Ay, Nihat, & Jost, Jürgen. 2014a. Closure measures for coarse-graining of the tent map. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, **24**(1), 013136.

Pfante, Oliver, Bertschinger, Nils, Olbrich, Eckehard, Ay, Nihat, & Jost, Jürgen. 2014b. Comparison between different methods of level identification. *Advances in Complex Systems*, **17**(02), 1450007.

Pouget, Alexandre, Dayan, Peter, & Zemel, Richard. 2000. Information processing with population codes. *Nature Reviews Neuroscience*, **1**(2), 125.

Price, Huw, & Corry, Richard. 2007. *Causation, physics, and the constitution of reality: Russell's republic revisited.* Oxford University Press.

Prinz, Jesse. 2007. The intermediate level theory of consciousness. *The Blackwell companion to consciousness*, 257–271.

Putnam, Hilary. 1967. Psychological predicates. *Art, mind, and religion*, **1**, 37–48.

Quian Quiroga, Rodrigo, & Panzeri, Stefano. 2009. Extracting information from neuronal populations: Information theory and decoding approaches. *Nature Reviews Neuroscience*, **10**(3), 173–185.

Raymont, Paul, & Brook, Andy. 2006. Unity of Consciousness. *Pages 565–577 of:* Beckermann, Ansgar, & McLaughlin, Brian P. (eds), *The Oxford Handbook of Philosophy of Mind.* Oxford University Press.

Revol, P, Rossetti, Y, Vighetto, A, Rode, G, Boisson, D, & Pisella, L. 2003. Pointing errors in immediate and delayed conditions in unilateral optic ataxia. *Spatial vision*, **16**(3-4), 347–364.

Revonsuo, Antti. 2006. *Inner presence: Consciousness as a biological phenomenon.* Mit Press.

Rosen, Robert. 1991. *Life itself: a comprehensive inquiry into the nature, origin, and fabrication of life.* Columbia University Press.

Sederberg, Audrey J, MacLean, Jason N, & Palmer, Stephanie E. 2018. Learning to make external sensory stimulus predictions using internal correlations in populations of neurons. *Proceedings of the National Academy of Sciences*, 201710779.

Seth, Anil K. 2014. A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive neuroscience*, **5**(2), 97–118.

Seth, Anil K. 2015. Presence, objecthood, and the phenomenology of predictive perception. *Cognitive neuroscience*, **6**(2-3), 111–117.

Stein, Richard B, Gossen, E Roderich, & Jones, Kelvin E. 2005. Neuronal variability: noise or part of the signal? *Nature Reviews Neuroscience*, **6**(5), 389.

Tononi, Giulio, & Koch, Christof. 2008. The neural correlates of consciousness: an update. *Annals of the New York Academy of Sciences*, **1124**(1), 239–261.

Tononi, Giulio, Boly, Melanie, Massimini, Marcello, & Koch, Christof. 2016. Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience*, **17**(7), 450.

White, John A., Rubinstein, Jay T., & Kay, Alan R. 2000. Channel noise in neurons. *Trends in Neurosciences*, **23**(3), 131–137.

Whitwell, Robert L., Milner, A. David, & Goodale, Melvyn A. 2014. The Two Visual Systems Hypothesis: New Challenges and Insights from Visual form Agnosic Patient DF. *Frontiers in Neurology*, **5**, 255.

Woodward, James. 2007. Causation with a Human Face. *In:* Price, Huw, & Corry, Richard (eds), *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Oxford University Press.

Figure Legends