Integrated Information in Process Theories: Towards Categorical IIT

Sean Tull[1] and Johannes Kleiner[2]

[1]*Cambridge Quantum Computing*
[2]*Munich Center for Mathematical Philosophy*
sean.tull@cambridgequantum.com, johannes.kleiner@lmu.de

Abstract

We demonstrate how integrated information and other key notions from Tononi et al.'s Integrated Information Theory (IIT) can be studied within the simple graphical language of process theories (symmetric monoidal categories). This allows IIT to be generalised to a broad range of physical theories, including as a special case the Quantum IIT of Zanardi, Tomka and Venuti, and sets the foundation for a categorical definition of IIT.

Keywords: *Integrated Information Theory, Process theory, Monoidal Category, Consciousness, Quantum Integrated Information Theory*

1. Introduction

Integrated Information Theory (IIT) is a theory of consciousness proposed and developed by Giulio Tononi and collaborators (Tononi, 2008; Oizumi et al., 2014). Originally defined in terms of a numerical measure Φ representing the level of phenomenal consciousness of a system (Tononi, 2004; Mediano et al., 2019), the most recent version of the theory, IIT 3.0, now employs an algorithm which claims to determine in addition which part of a system is conscious, and what it is conscious of.

Received 28th February 2021; Revised 25th May 2021; Accepted 30th May 2021
Journal of Cognitive Science 22(2): 92-123 June 2021

In this article we show how the key concepts of IIT, including systems, integration and causation, can be studied naturally in the language of physical *process theories*, which are mathematically described as *symmetric monoidal categories*. Process theories come with an intuitive but rigorous graphical calculus (Selinger, 2011) which allows us to present many aspects of IIT in a simple pictorial fashion.

The constructions we provide in this article can be applied to any suitable process theory to yield a notion of *generalised IIT* as defined by the authors in a companion article (Kleiner and Tull, 2021). This allows us to extend IIT to new physical settings. As special cases, choosing the process theory of classical probabilistic processes essentially yields the usual IIT 3.0 in the sense of (Oizumi et al., 2014). Starting instead from the theory of quantum processes gives the *Quantum Integrated Information Theory* defined by Zanardi, Tomka and Venuti (Zanardi et al., 2018), which was another motivation for this work.

Independently of consciousness itself, our constructions provide a possible foundation for a general theory of integrated or 'holistic' behaviour within process theories, i.e. monoidal categories, which may be of interest to a broad range of fields. For example, neural net-like systems that achieve a task using a high degree of integration should be more efficient than fully modular ones, in that they require fewer neurons for the same task, and indeed integrated behaviour has been shown to evolve in simple models of biological organisms (Albantakis et al., 2014). The methods of IIT have been applied generally in the study of integration in information processing systems, including treatments of autonomy (Marshall et al., 2017), causation (Albantakis et al., 2017), and state differentiation (Marshall et al., 2016).

1.1 Background: Mathematical Consciousness Science

The background for our work is in the growing field of Mathematical Consciousness Science (MCS), which aims to apply formal and mathematical tools in order to resolve open problems in the scientific study of consciousness. One major goal thereby is to expose and improve the mathematical structure of neuroscientific theories of consciousness so as to allow quantifiable comparison between competing models, generate novel experimental predictions, and to provide a thorough foundation for further development

and combination of theories. More foundationally, it aims to uncover how consciousness relates to the physical world in terms of empirically grounded and philosophically motivated scientific theories. Progress in this direction is essential for resolving medical challenges (most notably, improving the understanding of neurological, psychiatric and psychological disorders (Michel et al., 2019)) and ethical reasons (for example the detection of consciousness in anesthetized or non-communicating patients (Alkire et al., 2008; Fink et al., 2018)), and could generate new advances in AI (artificial implementation of consciousness-related functions, for example (McDermott, 2007)).

A crucial cornerstone in this program is the representation of conscious experience in terms of a mathematical spaces, and to expound theories of consciousness as mappings from a mathematical description of physical systems to these spaces. Early precursors of the former are quality spaces (Beals et al., 1968; Clark, 1996, 2000) which make use of just noticeable difference between stimuli to construct a representation of mental qualities and similarities between them. In the companion article (Kleiner and Tull, 2021), we provide a definition of an *experience space* that builds upon quality spaces while being geared at precisely what is required to flash out IIT as a mathematical mapping of the just-mentioned kind.

This contributes to the exploration and application of category theory as a framework for theories of consciousness (Tsuchiya et al., 2016; Northoff et al., 2019; Ehresmann, 2012). Category theory itself provides a natural language for describing mappings between scientific domains, such as domains of physical systems and those modelling phenomenal experiences. Its emphasis on processes between systems in particular makes it ideal for describing theories and experimental findings which relate consciousness to dynamical processes, as discussed for example in (Fekete and Edelman, 2011; Wiese and Friston, 2020; Grindrod, 2018). The use of monoidal categories in this article additionally allows us to treat compositional aspects of systems and processes, which are central to theories such as IIT.

1.2 A Primer on Integrated Information Theory

Though the majority of the article is self-contained and requires no prior knowledge of the theory, for context we include here a short introduction

to IIT 3.0 (Oizumi et al., 2014), as formulated in its general form in our companion article (Kleiner and Tull, 2021) to which we refer for a more detailed presentation of the theory.

Any generalised IIT, including IIT 3.0, takes as input a given class of physical systems S, each with a given state space $\mathrm{St}(S)$, and specifies a map $\mathbb{E}$ which provides each system with a space describing its possible conscious experiences. Additionally, for each state $s \in \mathrm{St}(S)$ the theory specifies a particular experience $\mathbb{E}(s) \in \mathbb{E}(S)$ which the system will have in that state:

$$\boxed{\text{Physical systems and states}} \xrightarrow{\mathbb{E}} \boxed{\text{Spaces and states of conscious experience}}$$

In IIT 3.0 the nature of this mapping derives from a number of essential properties–so called 'axioms'–which are postulated to characterize every conscious experience. Next to integration and information, these axioms include intrinsic existence, composition and exclusion (Tononi, 2015). These axioms are being translated into formal requirements. To this end, comparably simple physical systems are considered. These consist of a set of elements (or 'nodes'), each usually with only two states (on or off), and come with a discrete Markovian time evolution which is often described via a given causal graph. The prototypical example would be a human brain, in which the nodes represent neurons and their firing. The result of the translation process is the algorithm of IIT 3.0, i.e. the map $\mathbb{E}$ when applied to classical physical systems.

Starting from such a system S along with its current state s, the theory then specifies a set of probability distributions known as the *cause-effect repertoire*. For each pair of subsystems M, P ('mechanism' and 'purview') of S, the cause repertoire $\mathsf{caus}(M, P)$ is a distribution specifying how the current state of M constrains the state of P in the previous time-step, and similarly the effect repertoire $\mathsf{eff}(M, P)$ addresses the next time-step instead.

In the IIT algorithm one goes on how to calculate how 'integrated' each of these repertoires are by comparing them against repertoires obtained instead by 'cutting' the (evolution of the) system into various parts, by removing causal connections between them. For each mechanism M one determines which purviews give the most integrated values of $\mathsf{caus}(M, P)$ and $\mathsf{eff}(M, P)$, and these repertoire values (along with their level of integration) determine a *concept* for that mechanism. The weighted collection of these concepts

determines the entity $\mathbb{E}(s)$, also known as the *Q-shape* of the system, which is claimed to specify its total conscious experience. In particular this Q-Shape comes with its own level of integration, denoted $\Phi(s)$, which describes 'how conscious' the system is as a whole. A final 'exclusion' step enforces that only the subsystem of S with the highest Φ value will in fact be conscious.

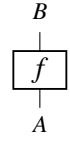
In the article (Kleiner and Tull, 2021) we show how to define a broad class of generalisations of IIT, in which for example the repertoires need no longer be described by probability distributions, but the states of a general physical theory. In the present article we describe how such IITs may be defined starting from any physical process theory. To do so we define the key notions of any IIT within such a setting, namely causal relations and their integration.

1.3 Structure of Article

The article is structured as follows. We introduce process theories in Section 2 and then use them to describe the key notions from IIT – decompositions of objects (Section 3), systems (Section 4) and cause and effect repertoires (Section 5). We summarise how to define a generalised IIT from a process theory in Section 6 before giving examples in Section 7 and discussing future work in Section 8. The appendix contains some initial steps in developing a general study of integration in monoidal categories.

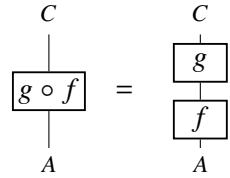
2. Process Theories

We begin by introducing the framework of *process theories* used throughout this work; for more detailed introductions we refer to (Coecke and Paquette, 2010; Coecke and Kissinger, 2017). The basic ingredients of such a theory are *objects* and *processes* between them. We depict a process from the object A to the object B as a box:

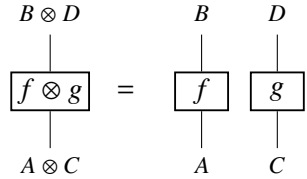


These processes may be *composed* together to form new ones in several ways. Firstly, given a process such as f above, and any other process g from

B to C, we may compose them 'in sequence' to form a new one from A to C, denoted:

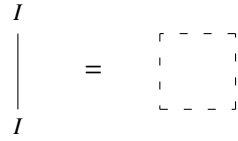


Secondly, we may compose processes in parallel. Any two objects A, B may be combined into a single object $A \otimes B$. Moreover any processes f from A to B, and g from C to D may be placed 'side-by-side' to form a new process:



from $A \otimes C$ to $B \otimes D$. More generally, by combining these operations, many processes may all be plugged together to form more complex diagrams describing a single composite process.

As a convenience, any process theory is taken to come with the following. Firstly, any object A come with an *identity process*, depicted as a blank wire on A, which 'does nothing' in that composing with it via $\circ$ leaves any process as it is. Secondly, it has a *trivial object*, denoted I, which leaves objects alone when combining under $\otimes$. We depict I as empty space:



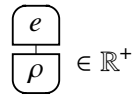
Finally, we formally assume the presence of a special process $\times$ which allows us to 'swap' any pair of wires over each other, along with a set of rules saying roughly that diagrams in the above sense are well-defined.

Mathematically, all of this is summarised by saying that a process theory is precisely a *symmetric monoidal category* $(\mathbf{C}, \otimes, I)$ with the processes as

its *morphisms*. Our diagrammatic rules correspond to the precise *graphical calculus* for reasoning in such categories (Selinger, 2011).

We will often wish to refer to some special kinds of processes. Processes with 'no input' in diagrams (and so formally with input object I) are called *states*, and can be thought of as 'preparations' of the physical system given by their output object:

Processes with no output, called *effects*, may be thought of as 'observations' we may record on our system. Finally, processes with neither input nor output are called *scalars*. It is common for theories to come with a *probabilistic* interpretation meaning that each of their scalars p correspond to a probability, or more generally an 'unnormalised probability' $p \in \mathbb{R}^+$, with $r \otimes s = r \cdot s$ for scalars and the empty diagram given by 1. In particular, the composition of a state with an effect

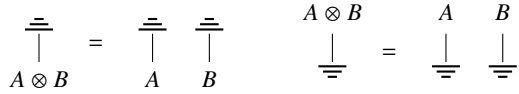


corresponds to the 'probability' of observing the effect e in the state ρ. Such 'generalised probabilistic theories' are a major focus of study in the foundations of physics (Barrett, 2007).

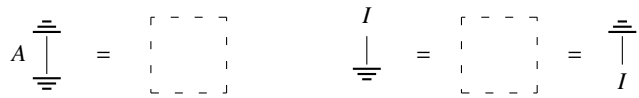
The theories we consider here will often come with further structure giving them a physical interpretation. Firstly, every object will come with a distinguished *discarding* effect depicted

which we think of as the process of simply 'throwing away' or 'ignoring' a physical system. Similarly, every object should come with a distinguished *completely mixed state* depicted as

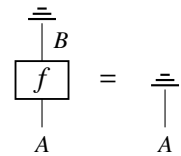
which corresponds to preparing the object in a maximally 'noisy' or 'random' state. These processes should satisfy



as well as



for all objects A, B. We then define a process f to be *causal* when it satisfies

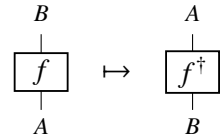


or similarly as *co-causal* if it preserves $\frac{\perp}{\top}$. Discarding processes are in fact closely related to physical notions of causality; see for example (Coecke, 2014; Chiribella et al., 2010).

In such a probabilistic theory there is a unique process between any two objects, the *zero process* 0, such that composing any process via $\circ, \otimes$ with 0 always yields 0.

At times we will assume our process theory also comes with a way of describing how similar any two causal states are. This amounts to a choice of *distance function* on the set $\mathrm{St}_{\mathsf{c}}(A)$ of causal states of each object A, providing a value $d(a, b) \in \mathbb{R}^+$ for each $a, b \in \mathrm{St}_{\mathsf{c}}(A)$. Often this map d will satisfy the axioms of a metric, but this is not required.

Our main examples of process theories will come with a notable extra feature, though this will not be necessary for our approach. In many theories it is possible to 'reverse' any process, in that for any process f there is another $f^\dagger$ in the opposite direction. We say a process theory has a *dagger* when it comes with such a mapping



which preserves composition and identity maps in an appropriate sense, and satisfies $f^{\dagger\dagger} = f$ for all f. The presence of a dagger is a common starting point in categorical approaches to quantum theory; see e.g. (Abramsky and Coecke, 2004; Selinger, 2007).

Let us now meet our main examples of process theories with the above features.

Example 1 (**Classical Probabilistic Processes**) In the process theory **Class** of finite-dimensional probabilistic classical physics, the objects are finite sets $A, B, C, \ldots$ and the processes f from A to B are functions sending each element $a \in A$ to a 'unnormalised probability distribution' over the elements of B, i.e functions $f\colon A \times B \to \mathbb{R}^+$. Composition of f from A to B and g from B to C is defined by

$$(g \circ f)(a, c) = \sum_{b \in B} f(a, b) \cdot g(b, c)$$

In this theory the trivial object is the singleton set $I = \{\star\}$, with $\otimes$ given by the Cartesian product $A \times B$ and $(f \times g)(a, c)(b, d) = f(a, b) \cdot g(c, d)$. This theory is probabilistic, with scalars $r \in \mathbb{R}^+$.

Here $\bar{\top}_A$ is the unique effect with $\bar{\top}_A(a) = 1$ for all $a \in A$. A process f is causal whenever it is stochastic, i.e. sends each element $a \in A$ to a (normalised) probability distribution over the elements of B. Applying the process $\bar{\top}$ to some output wire of a process corresponds to *marginalising* over the set which is discarded.

States of an object are '$\mathbb{R}^+$-distributions' over their elements, while causal states are normalised ones, i.e. probability distributions. The completely mixed state $\bar{\bot}_A$ is the uniform probability distribution, with $\bar{\bot}_A(a) = \frac{1}{|A|}$ for all $a \in A$. This theory also has a dagger by $f^\dagger(b, a) = f(a, b)$.

Similarly we define another process theory $\mathbf{Class}_\mathsf{m}$, in the same way, but with objects now being finite *metric spaces* (A, d). Each object A now comes with a metric d on its underlying set, with $A \otimes B = A \times B$ having the product metric. For each object A we extend d to a metric d_W on probability distributions over A, i.e. causal states of A, called the *Wasserstein metric* or *Earth Mover's Distance* (EMD), definable e.g. by

$$d_W(s, t) := \sup_f \{\sum_{a \in A} f(a) \cdot s(a) - \sum_{a \in A} f(a) \cdot t(a)\}$$

where the suprema is taken over all functions f satisfying $|f(a) - f(b)| \leq d(a, b)$ for all a, b. **Class** itself may be given a metric on causal states in the same way by taking each object A to have metric $d(a, b) = 1 - \delta_{a,b}$.

Example 2 (**Quantum Processes**) In the process theory **Quant** the objects are finite-dimensional complex Hilbert spaces $\mathcal{H}, \mathcal{K}, \ldots$ and the processes from $\mathcal{H}$ to $\mathcal{K}$ are *completely positive maps* $f\colon B(\mathcal{H}) \to B(\mathcal{K})$ between their spaces of operators. Here $I = \mathbb{C}$ and $\otimes$ is the usual tensor product of Hilbert spaces and maps. States ρ of an object $\mathcal{H}$ may be identified with (unnormalised) *density matrices*, i.e. quantum states in the usual sense, as may effects. The effect $\bar{\top}$ sends each operator $a \in B(\mathcal{H})$ to its *trace* $\mathrm{Tr}(a)$, and $\bar{\bot}$ is the maximally mixed state on $\mathcal{H}$, with density matrix $\frac{1}{\dim(\mathcal{H})} 1_{\mathcal{H}}$. Here a process is causal precisely when it is trace-preserving, and the dagger is given by the Hermitian adjoint.

Example 3 (**Quantum-Classical Processes**) To combine **Class** and **Quant** we may use the theory **CStar** whose objects are finite-dimensional C^*-*algebras* $A, B, \ldots$ and processes are completely positive maps $f\colon A \to B$, with $\otimes$ given by the standard tensor product, $I = \mathbb{C}$ and the dagger again by the Hermitian adjoint. Here $\bar{\top}$ sends each element $a \in A$ to its trace $\mathrm{Tr}(a) \in \mathbb{C}$, while $\bar{\bot}$ corresponds to the rescaling $\frac{1}{d}1$ of the element $1 \in A$, where $\mathrm{Tr}(1) = d$. Each C*-algebra comes with a metric induced by its norm, providing a metric on states in the theory.

Class may be identified with the sub-theory of **CStar** containing the commutative algebras, and **Quant** with those of the form $B(\mathcal{H})$ for some Hilbert space $\mathcal{H}$. More general algebras are 'quantum-classical', being given by direct sums of quantum algebras.

3. Decompositions

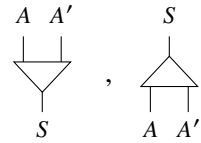
A central aspect of IIT is evaluating the level of integration of a process, and particularly of a state of some object. To do so we must compare the object in question against ways it may be *decomposed*, as follows.

Firstly, recall that a process f from A to B is an *isomorphism* when there is some (unique) f^{-1} from B to A for which $f^{-1} \circ f$ and $f \circ f^{-1}$ are both identities. We write $A \simeq B$ when such an isomorphism exists.

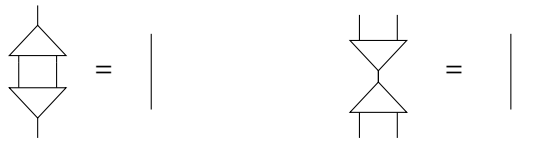
Definition 4 *In any process theory, a* decomposition *of an object S is a pair of objects A, A' along with an isomorphism $S \simeq A \otimes A'$.*

In a process theory with $\overset{\bullet}{\top}, \underset{\bullet}{\bot}$ *we will always consider decompositions whose isomorphisms are causal and co-causal. We also assume that decomposition isomorphisms preserve any distances between causal states.*

For short we often denote such a decomposition simply by (A, A') and depict its isomorphism and inverse by

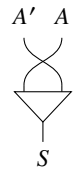


respectively. The fact that they form an isomorphism means that



One can go on to develop a general study of decompositions in process theories. Here we just note some of the basics, for more see Appendix A.

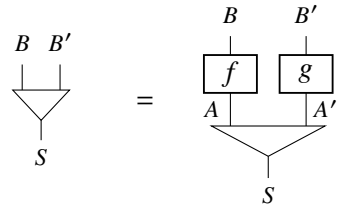
Firstly, any decomposition has an induced *complement* decomposition $(A, A')^{\perp} := (A', A)$, with isomorphism given by swapping its components:



All decompositions then satisfy $(A, A')^{\perp\perp} = (A, A')$. Moreover, any object always S always comes with *trivial decompositions* denoted $1 := (S, I)$ and $0 := (I, S)$ with $0 = 1^{\perp}$. Drawing either of their isomorphisms would just mean drawing a blank wire labelled by S.

It is also useful to note when two decompositions of an object are 'essentially the same'. We write $(A, A') \sim (B, B')$ and call both decompositions

equivalent when there exists isomorphisms f, g with



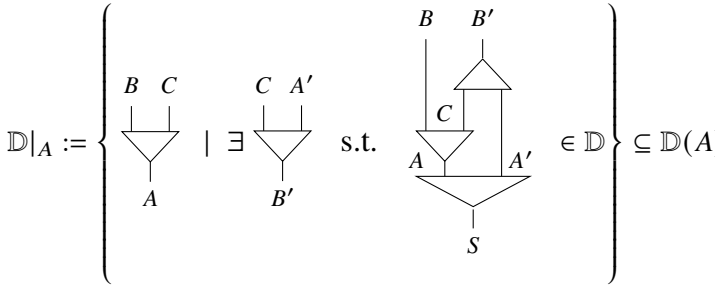
(1)

In a theory with $\overset{=}{\top}, \overset{=}{\bot}$ we require moreover that f, g are causal and co-causal.

We write $\mathbb{D}(S)$ for the set of all equivalence classes of decompositions of S under $\sim$ (we will ignore the fact that in full generality each equivalence class may be a proper class rather than a set). Often we abuse notation and denote the members of simply by (A, A') instead of as equivalence classes $[(A, A')]_\sim$. It is easy to see that if two decompositions are equivalent then so are their complements, so that $(-)^\perp$ is well-defined on $\mathbb{D}(S)$.

Definition 5 *By a* decomposition set *of an object S in a process theory we mean a subset $\mathbb{D}$ of $\mathbb{D}(S)$ containing* 1 *and closed under* $(-)^\perp$.

Given any decomposition set $\mathbb{D}$ of S and any $(A, A') \in \mathbb{D}$, we define the *restriction* of $\mathbb{D}$ to A via this decomposition to be the decomposition set



Intuitively $\mathbb{D}|_A$ consists of all decompositions of A which themselves can be extended to give a decomposition of S belonging to $\mathbb{D}$, via (A, A').

The most important examples of decomposition sets are the following.

Example 6 Let S be an object with a given isomorphism

$$S \simeq S_1 \otimes \cdots \otimes S_n$$

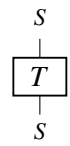
representing S as finite tensor of objects S_i which we may call *elements*. This induces a decomposition set $\mathbb{D}$ of S whose elements correspond to subsets J of the elements. For any such subset, defining $S_J := \bigotimes_J S_j$ we have a decomposition $S \simeq S_J \otimes S_{J'}$ where J' is the set of remaining elements. Then $\mathbb{D}|_{S_J}$ contains a decomposition for each $K \subseteq J$ in the same way.

Decompositions via elements as above are the only kinds appearing in classical or quantum IIT. However, more general ones allow us to treat systems which are not decomposable into any finite set of 'elementary' subsystems.

4. Systems

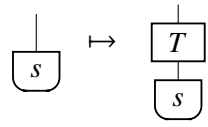
We now begin by seeing how each of the main components of IIT, or any 'generalised IIT' in the sense of (Kleiner and Tull, 2021), may be treated starting from any given process theory **C**. The focus will be on a class of *systems*, as follows.

Definition 7 *By a* system type *we mean a triple* $\underline{S} = (S, \mathbb{D}, T)$ *consisting of an object* S *with a decomposition set* $\mathbb{D}$ *and a causal process*

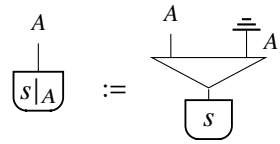


which we call its time evolution. *A* state *of* $\underline{S}$ *is simply a state of* S *in* **C**. *We typically refer to a system type simply as a* system.

The set $\mathbb{D}$ specifies the ways in which we will decompose our underlying system when assessing integration. The process T is intended to describe the way in which states of the system evolve over each single 'time-step', via



In what follows it will be useful to be able to restrict any state s of our system to the components of any decomposition $(A, A') \in \mathbb{D}$ by setting

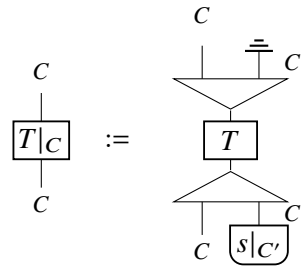


and defining $s|_{A'}$ similarly. We define the *trivial system* $\underline{I}$ to have object I, a single decomposition $1 = (I, I) = 0$, and time evolution being the identity.

4.1 Subsystems

There are several operations on systems one carries out in the context of IITs. The first is the taking of *subsystems*.

Definition 8 *For each object C belonging to some decomposition $(C, C') \in \mathbb{D}$, and each state s of $\underline{S}$, the corresponding* subsystem *of $\underline{S}$ is defined to be the system type $\underline{C}^s := (C, \mathbb{D}|_C, T|_C)$ with time evolution*



The above definition of $T|_C$ is from (Oizumi et al., 2014) and aims to capture the evolution of a state of C conditioned on the state of C' being $s|_{C'}$.

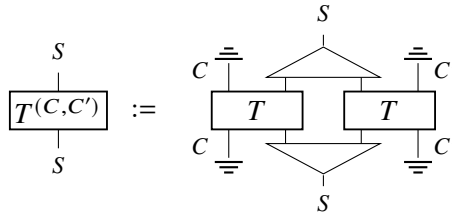
4.2 Cutting

A second important operation involves removing (some or all) causal connections between the two different components of a decomposition of a

system. For any system $\underline{S} = (S, \mathbb{D}, T)$ and decomposition $(C, C') \in \mathbb{D}$, we should be able to form a new such *cut* system of the form

$$\underline{S}^{(C,C')} = (S, \mathbb{D}, T^{(C,C')})$$

with the new evolution $T^{(C,C')}$ removing some influence between these regions. The most straightforward form of cutting is a *symmetric cut*, in which both components are fully disconnected from each other, with evolution

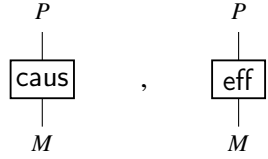


(2)

(where the triangle denotes $(C, C')^{\perp}$). However, later we will see that some IITs use additional structure to carry out alternative notions of system cut.

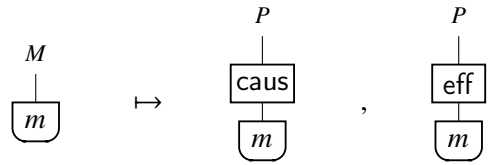
5. Cause and Effect

Central to any IIT is a notion of causal influence between any two possible subsystems of a system. These influences are captured in a pair of assignments called the *cause repertoire* and *effect repertoire* of the system. In IIT 3.0 these contain probability distributions describing how the present state of each subsystem constrains the past and future states of each other subsystem (Oizumi et al., 2014). For our purposes it suffices to note that such cause and effect repertoires amount to specifying a pair of processes



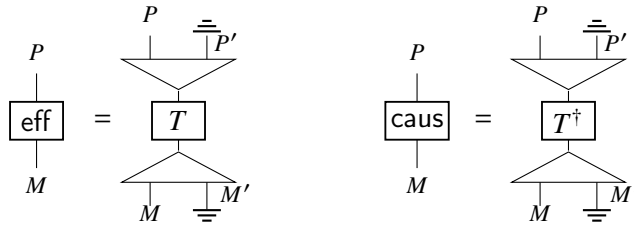
for each pair of underlying objects M, P of subsystems $\underline{M}, \underline{P}$ of $\underline{S}$ via some state s. In this setting M is typically called the 'mechanism' and P the 'purview', and the above processes should capture the way in which the

current state m of M constrains the previous or next state of P, respectively. These constraints are captured by the pair of states of P given by plugging in the 'current' state m of M:



We will additionally require the processes caus, eff to be *weakly causal* in the sense that whenever the state m is causal then each of the above states must either be causal or 0.

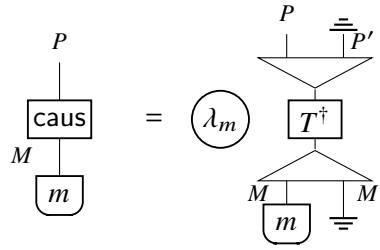
Example 9 For any process theory (resp. with a dagger) there is a simple choice of effect (resp. cause) repertoire given by



(3)

Note however that this definition of caus may not be weakly causal in our above sense if $T^\dagger$ is not causal.

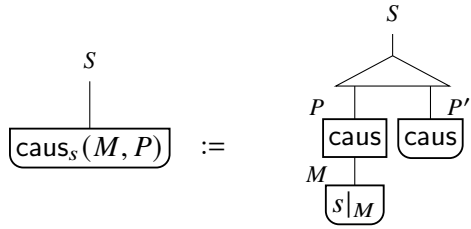
In a probabilistic process theory we should instead have that



(4)

where λ_m is the unique *normalisation* scalar for the right-hand state, making it causal if it is non-zero (and being zero otherwise). It is not in general possible to define a process caus in terms of its action on states m in this way, but this is possible for example in **Class**, **Quant** or **CStar**.

However the repertoires are specified, we will need to compare their values in a fixed state while varying P. To do so, for each state s of $\underline{S}$ and each such M, P we define the *cause repertoire at s* to be the state of S given by



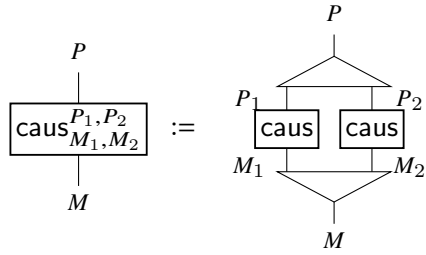
(5)

The features of this diagram have special names in (Oizumi et al., 2014); the right-hand caus state above, given by taking mechanism $M = I$, is called the *unconstrained* cause repertoire, and the whole process above $s|_M$ in the diagram is called the *extended cause repertoire* at M, P. Defining them in this way allows us to compare the repertoire values for varying M, P.

Similarly, $\mathsf{eff}_s(M, P)$, the *effect repertoire at s*, and the *unconstrained* and *extended effect repertoire* are all defined in terms of eff in the same way.

5.1 Decomposing Repertoires

In an IIT we must assess how integrated each of these repertoire values are at a given state . This involves comparing the repertoires with how they behave under decomposing each of M and P. For any decompositions $(M_1, M_2) \in \mathbb{D}|_M$ of M and $(P_1, P_2) \in \mathbb{D}|_P$ of P, the *decomposed* cause repertoire process is defined by



(6)

We then define the state $\mathsf{caus}^{P_1,P_2}_{s,M_1,M_2}(M, P)$ just like (5) but replacing caus with the process (6). We decompose the effect repertoire in just the same way in terms of eff.

6. Generalised IITs

In summary, let **C** be a process theory coming with the features $\overline{\top}, \underline{\top}, d$ of Section 2. To define an integrated information theory we must specify:

1. a class **Sys** of system types, closed under subsystems;
2. a definition of system cuts, under which **Sys** is closed;
3. a choice of weakly causal processes $\mathsf{caus}, \mathsf{eff}$ between the underlying objects M, P of each pair of subsystems $\underline{M}, \underline{P}$ via some state s, of any system $\underline{S}$.

More precisely, this provides the *data* of a generalised integrated information theory in the sense of (Kleiner and Tull, 2021). From this data we may now use the *IIT algorithm* from (Oizumi et al., 2014) to calculate the usual objects of interest in IIT.

6.1 The IIT Algorithm

We now briefly summarise this algorithm as treated in the general setting in (Kleiner and Tull, 2021), to which we refer for more details. Let us fix a 'current' state s of a system $\underline{S}$. Firstly, the level of *integration* of each value of the cause repertoire is defined by

$$\phi(\mathsf{caus}_s(M,P)) \;:=\; \min d(\mathsf{caus}_s(M,P)\;,\;\mathsf{caus}^{P_1,P_2}_{s,M_1,M_2}(M,P)) \qquad (7)$$

where the minima is taken over all pairs of decompositions of M, P which are not both trivial, i.e. equal to 1. [1] The integration level $\phi(\mathsf{eff}_s(M,P))$ is defined similarly in terms of eff.

For each choice of mechanism M, its *core cause* P^c and *core effect* P^e are the purviews P with maximal ϕ values for $\mathsf{caus}, \mathsf{eff}$ respectively. The minima of their corresponding ϕ values is then denoted by $\phi(M)$. We then associate to M and object called its *concept* $\mathbb{C}(M)$, essentially defined as the triple

$$(\mathsf{caus}_s(M,P^c), \mathsf{eff}_s(M,P^e), \phi(M))$$

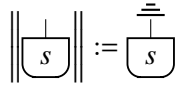
[1] When $\mathsf{caus}_s(M,P) = 0$ we alternatively set $\phi = 0$.

More precisely, in (Kleiner and Tull, 2021), $\mathbb{C}(M)$ is given by the pair of above repertoire values with each 'rescaled' by $\phi(M)$.

The tuple $\mathbb{Q}(s)$ of all these concepts, for varying M, is called the *Q-shape* $\mathbb{Q}(s)$ of the state s. The collection of all possible such tuples is denoted $\mathbb{E}(\underline{S})$. The level of integration of $\mathbb{Q}(s)$ is calculated similarly to (7) by considering all possible cuts of the system. The subsystem $\underline{M}$ of $\underline{S}$ whose Q-shape is itself found to be most integrated is called the *major complex*. Rescaling this Q-shape $\mathbb{Q}(\underline{M}, s|_M)$ according to its level of integration, and using an embedding $\mathbb{E}(\underline{M}) \hookrightarrow \mathbb{E}(\underline{S})$ we finally obtain a new element $\mathbb{E}(s) \in \mathbb{E}(\underline{S})$.

The claim of an IIT with regards to consciousness is that $\mathbb{E}(\underline{S})$ is the space of all possible conscious experiences of the system $\underline{S}$, and that $\mathbb{E}(s)$ is the particular experience attained when it is in the state s, with intensity $\Phi(s) := \| \mathbb{E}(s) \|$.

Remark 10 Let us make explicit how the specification of 1, 2, 3 above provides the data of an IIT in the sense of (Kleiner and Tull, 2021). The system class of the theory is **Sys**, and $\mathsf{caus}_s(M, P), \mathsf{eff}_s(M, P)$ and their decompositions are as outlined in Section 5.1. When **C** is probabilistic and has distances $d(a, b)$ defined for *arbitrary* states a, b of an object A, we may define the space of *proto-experiences* $\mathbb{PE}(\underline{S})$ of a system $\underline{S}$ to be simply its set of states, with



However, if d is only defined on causal states, as in classical IIT, to follow the algorithm from (Kleiner and Tull, 2021) one must instead set $\mathbb{PE}(\underline{S}) := \mathrm{St}_{\mathsf{c}}(S) \times \mathbb{R}^+$ as in Example 3 of (Kleiner and Tull, 2021). For either choice, for any subsystem $\underline{M}$ of $\underline{S}$ we obtain an embedding $\mathbb{PE}(\underline{M}) \hookrightarrow \mathbb{PE}(\underline{S})$ by composing alongside $\frac{\perp}{=}_{M^\perp}$, and this can be seen to provide a further embedding $\mathbb{E}(\underline{M}) \hookrightarrow \mathbb{E}(\underline{S})$.

7. Examples

Let us now meet several examples of IITs defined from process theories.

7.1 Generic IITs

Let $\mathbf{C}$ be any process theory coming with the features outlined in Section 2, including a dagger on processes. We define a generalised IIT denoted $\mathsf{IIT}(\mathbf{C})$ by taking as systems all tuples $\underline{S} = (S, \mathbb{D}, T)$ of an object S in $\mathbf{C}$ along with a causal process T and a decomposition set $\mathbb{D}$ induced by a single isomorphism $S \simeq \bigotimes_{i=1}^{n} S_i$ in terms of elements S_i, as in Example 6. As before each partition of these elements gives a decomposition of S. We define system cuts to be symmetric as in (2) and the repertoires in the straightforward sense of (3).

Remark 11 We can extend this example in to ways. Firstly we may allow systems $\underline{S}$ to come with arbitrary finite decomposition sets $\mathbb{D}$ of S. Secondly, we may extend the definition to theories without daggers by instead simply requiring each system $\underline{S}$ to come with a process T^{-} describing 'reversed time evolution', and then define the cause repertoire by replacing $T^{\dagger}$ with T^{-}.

7.2 Classical IIT

The 'classical' IIT version 3.0 of Tononi and collaborators (Oizumi et al., 2014) is built on the process theory $\mathbf{Class}_{\mathsf{m}}$. As such a toy model of the theory is provided by $\mathsf{IIT}(\mathbf{Class}_{\mathsf{m}})$. However IIT 3.0 itself differs from this theory, using some more specific features of the process theories $\mathbf{Class}$ and $\mathbf{Class}_{\mathsf{m}}$ which we now describe.

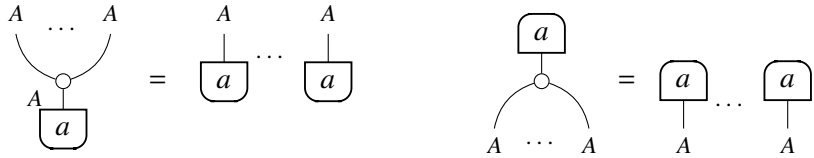
Firstly, note that in these classical process theories, for each object A, each element $a \in A$ corresponds to a unique state given by the point distribution at a, as well as a unique effect, namely the map sending a to 1 and all other elements of A to 0. We denote this state and effect both simply by a.[2]

Any process f from A to B is determined entirely by its compositions with these special states and effects since plugging in such a state a and effect b yields its value $f(a, b)$.

Another special feature of these classical process theories is that each object A comes with a distinguished *copying* process from A to $A \otimes \cdots \otimes A$,

[2]Typically these are the only kinds of 'state' considered, e.g. in (Oizumi et al., 2014) and even in our related article (Kleiner and Tull, 2021). In contrast here the term 'state' would include all distributions over A, i.e. all states of the process theory $\mathbf{Class}_{\mathsf{m}}$.

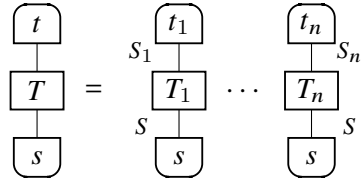
for any number of copies of A, as well as a *comparison* process in the opposite direction. We denote and define these respectively by the rules



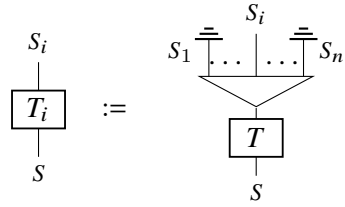
for all $a \in A$. Abstractly, these operations form a canonical commutative *Frobenius algebra* on each object, and there is no such canonical algebra on each object in **Quant** due to the *no-cloning* theorem (Coecke et al., 2013). We may now describe IIT 3.0 itself as follows.

7.2.1 Systems

In this theory systems are defined similarly to $\mathsf{IIT}(\mathbf{Class}_{\mathsf{m}})$, being given by a finite metric space S given as a product of elements $S \simeq \bigotimes_{i=1}^{n} S_i$, along with a causal (i.e. stochastic) evolution T on S. Additionally in (Oizumi et al., 2014) each evolution T is required to satisfy *conditional independence*, which states that for all $s, t \in S$, with $t = (t_1, \ldots, t_n)$ for some $t_i \in S_i$ we have

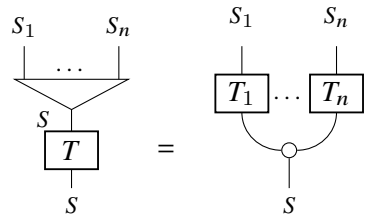


where for each element S_i we define the process T_i by



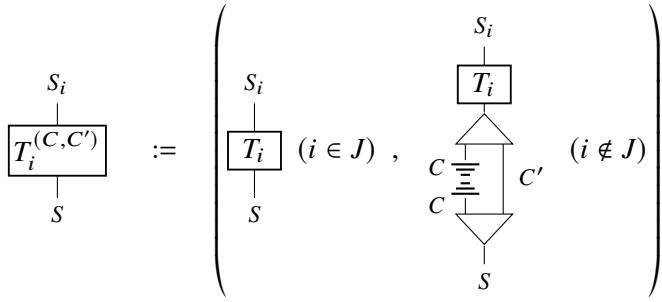
having depicted the isomorphism $S \simeq \bigotimes_{i=1}^{n} S_i$ by the triangle above. In other words, conditional independence states that the probabilities for the

next state of each element S_i are independent. Equivalently, T must satisfy



7.2.2 Cuts

Rather than our earlier symmetric cuts, the system cuts used in IIT 3.0 are *directional*. For any decomposition (C, C') of S with $C = \bigotimes_{j \in J} S_j$ for some subset of notes indexed by $J \subseteq \{1, \ldots, n\}$, we define the cut evolution $T^{(C,C')}$ using conditional independence by setting



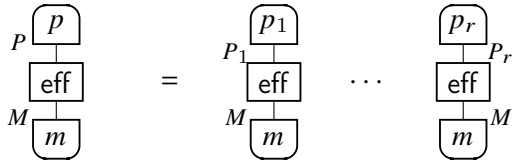
In other words, in the cut system all causal connections $C \to C'$ are replaced by noise, while all those into C remain intact.

7.2.3 Repertoires

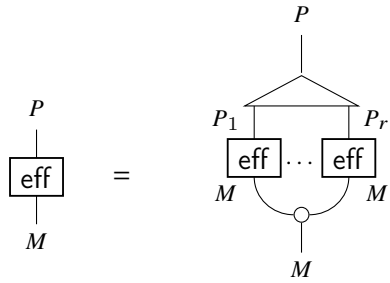
Let us now define the processes caus, eff between a pair of objects M and P, with $M = \bigotimes_{i=1}^{k} M_i$ and $P = \bigotimes_{j=1}^{r} P_j$ for some subsets $\{M_1, \ldots, M_k\}$ and $\{P_1, \ldots, P_r\}$ of elements of the system.

We begin with eff. When P is simply a single element P_j, eff is defined exactly as in (3). For more general P we define eff to again satisfy a form of

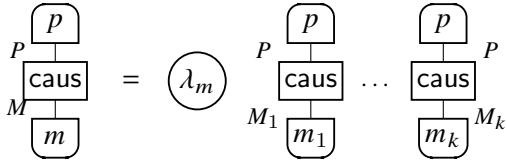
conditional independence, so that



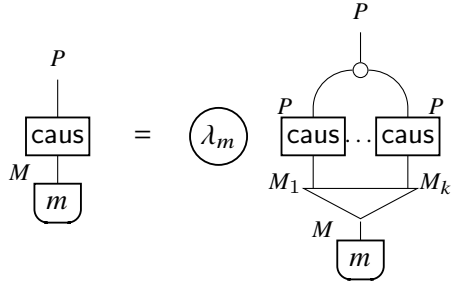
for all $m \in M$, $p = (p_1, \ldots, p_r) \in P$. Equivalently, we have that



In a similar fashion, whenever M is a single element M_i we define caus from M to P as in (4), while for more general M we require that



for all $m = (m_1, \ldots, m_k) \in M$ and $p \in P$, where λ_m is the normalisation scalar making $\mathsf{caus} \circ m$ a causal state (probability distribution) if it is non-zero, or $\lambda_m = 0$ otherwise. Equivalently, this means that



for each $m \in M$. This concludes the data of classical IIT.

7.3 Quantum IIT

Zanardi, Tomka and Venuti have proposed a quantum extension of classical IIT (Zanardi et al., 2018). In fact it is comparatively much simpler to describe in our approach, being precisely the theory $\mathsf{IIT}(\mathbf{Quant})$.

Explicitly, systems in this theory are given by finite-dimensional complex Hilbert spaces $\mathcal{H}$ along with a given decomposition into elements $\mathcal{H} \simeq \bigotimes_{i=1}^{n} \mathcal{H}_i$ and a completely positive trace-preserving map T on $B(\mathcal{H})$. States and repertoire values are given by density matrices ρ. In this theory each Q-shape $\mathbb{Q}(\rho)$ may be encoded as a single positive semi-definite operator on the space $(\mathbb{C}^2)^{\otimes n} \otimes \mathbb{C}^2 \otimes \mathcal{H}$, as discussed in (Zanardi et al., 2018).

7.4 Quantum-Classical IIT

We may now define a version of *quantum-classical IIT* as $\mathsf{IIT}(\mathbf{CStar})$. This synthesizes quantum IIT with the toy version $\mathsf{IIT}(\mathbf{Class}_{\mathsf{m}})$ of classical IIT, containing both kinds of systems. In future it would be desirable to synthesise quantum IIT with IIT 3.0 proper. Since the latter relies on the presence of copying maps, this may be achievable using the more general notion of a *leak* on a C*-algebra (Selby and Coecke, 2017).

8. Outlook

In this article we have taken first steps to show how Integrated Information Theory, and its generalisations to other domains of physics, may be studied categorically. There are many avenues for future work.

Firstly, we have so far made no requirements on the cause and effect repertoire processes caus, eff. To be fit for their name these processes should be required to satisfy axioms which ensure they have a causal interpretation, ideally determining them uniquely within any given process theory. Monoidal categories provide a natural setting for the study of causality, a major contemporary topic in the foundations of physics (Kissinger and Uijlen, 2017).

At a higher level, it seems natural for the class of systems **Sys** of a generalised IIT to itself form a category. The theory itself should then give a functor into another category **Exp** of (spaces of) phenomenal experiences; a formalization of the latter is for example given in (Kleiner and Tull, 2021).

Making IIT functorial in this way will likely involve modifying it to be more natural from a categorical perspective. Developing a useful notion of integration applicable to any monoidal category may also help to resolve mathematical problems of the IIT algorithm, for example its relying on the unique existence of core purviews which are not guaranteed*.

References

Abramsky, Samson and Bob Coecke. 2004. A categorical semantics of quantum protocols. In *Proceedings of the 19th Annual IEEE Symposium on Logic in Computer Science, 2004.*, pages 415–425. IEEE.

Albantakis, Larissa, Arend Hintze, Christof Koch, Christoph Adami, and Giulio Tononi. 2014. Evolution of integrated causal structures in animats exposed to environments of increasing complexity. *PLoS Comput Biol* 10 (12): e1003966.

Albantakis, Larissa, William Marshall, Erik Hoel, and Giulio Tononi. 2017. What caused what? An irreducible account of actual causation. *arXiv:1708.06716* .

Alkire, Michael T, Anthony G Hudetz, and Giulio Tononi. 2008. Consciousness and anesthesia. *Science* 322 (5903): 876–880.

Barrett, Jonathan. 2007. Information processing in generalized probabilistic theories. *Physical Review A* 75 (3): 032304.

Beals, Richard, David H Krantz, and Amos Tversky. 1968. Foundations of multidimensional scaling. *Psychological review* 75 (2): 127.

Chiribella, Giulio, Giacomo Mauro D'Ariano, and Paolo Perinotti. 2010. Probabilistic theories with purification. *Physical Review A* 81 (6): 062348.

Clark, Austen. 1996. *Sensory Qualities*. Oxford Scholarship Online.

———. 2000. *A theory of sentience*. Clarendon press.

Coecke, Bob. 2014. Terminality implies non-signalling. *arXiv preprint arXiv:1405.3681* .

Coecke, Bob and Aleks Kissinger. 2017. *Picturing quantum processes*. Cambridge University Press.

*ACKNOWLEDGEMENTS: We would like to thank the organizers and participants of the *Workshop on Information Theory and Consciousness* at the Centre for Mathematical Sciences of the University of Cambridge, of the *Modelling Consciousness Workshop* in Dorfgastein and of the *Models of Consciousness Conference* at the Mathematical Institute of the University of Oxford for discussions on this topic. Much of this work was carried out while Sean Tull was under the support of an EPSRC Doctoral Prize at the University of Oxford, from November 2018 to July 2019, and while Johannes Kleiner was under the support of postdoctoral funding at the Institute for Theoretical Physics of the Leibniz University of Hanover. We would like to thank both institutions.

Coecke, Bob and Eric Oliver Paquette. 2010. Categories for the practising physicist. In *New structures for physics*, pages 173–286. Springer.

Coecke, Bob, Dusko Pavlovic, and Jamie Vicary. 2013. A new description of orthogonal bases. *Mathematical Structures in Computer Science* 23 (3): 555–567.

Ehresmann, Andrée C. 2012. Mens: from neurons to higher mental processes up to consciousness. In *Integral Biomathics*, pages 29–30. Springer.

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

Fink, Sascha Benjamin, Wanja Wiese, and Jennifer Michelle Windt. 2018. *Philosophical and Ethical Aspects of a Science of Consciousness and the Self*. Frontiers in Psychology.

Grindrod, Peter. 2018. On human consciousness: A mathematical perspective. *Network neuroscience* 2 (1): 23–40.

Kissinger, Aleks and Sander Uijlen. 2017. A categorical semantics for causal structure. In *2017 32nd Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*, pages 1–12. IEEE.

Kleiner, Johannes and Sean Tull. 2021. The Mathematical Structure of Integrated Information Theory. *Frontiers in Applied Mathematics and Statistics* .

Marshall, William, Jaime Gomez-Ramirez, and Giulio Tononi. 2016. Integrated information and state differentiation. *Frontiers in psychology* 7: 926.

Marshall, William, Hyunju Kim, Sara I Walker, Giulio Tononi, and Larissa Albantakis. 2017. How causal analysis can reveal autonomy in models of biological systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 375 (2109): 20160358.

McDermott, Drew. 2007. Artificial intelligence and consciousness. *The Cambridge handbook of consciousness* pages 117–150.

Mediano, Pedro AM, Anil K Seth, and Adam B Barrett. 2019. Measuring integrated information: Comparison of candidate measures in theory and simulation. *Entropy* 21 (1): 17.

Michel, Matthias, Diane Beck, Ned Block, Hal Blumenfeld, Richard Brown, David Carmel, Marisa Carrasco, Mazviita Chirimuuta, Marvin Chun, Axel Cleeremans, et al. 2019. Opportunities and challenges for a maturing science of consciousness. *Nature human behaviour* 3 (2): 104–107.

Northoff, Georg, Naotsugu Tsuchiya, and Hayato Saigo. 2019. Mathematics and the brain: A category theoretical approach to go beyond the neural correlates of consciousness. *Entropy* 21 (12): 1234.

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

Selby, John and Bob Coecke. 2017. Leaks: quantum, classical, intermediate and

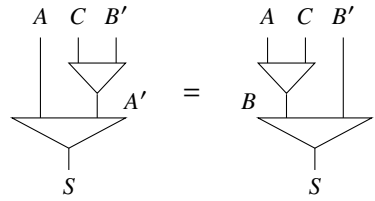
more. *Entropy* 19 (4): 174.
Selinger, Peter. 2007. Dagger compact closed categories and completely positive maps. *Electronic Notes in Theoretical computer science* 170: 139–163.
———. 2011. A survey of graphical languages for monoidal categories. In *New structures for physics*, pages 289–355. Springer.
Tononi, Giulio. 2004. An information integration theory of consciousness. *BMC neuroscience* 5 (1): 42.
———. 2008. Consciousness as integrated information: a provisional manifesto. *The Biological Bulletin* 215 (3): 216–242.
———. 2015. Integrated information theory. *Scholarpedia* 10 (1): 4164.
Tsuchiya, Naotsugu, Shigeru Taguchi, and Hayato Saigo. 2016. Using category theory to assess the relationship between consciousness and integrated information theory. *Neuroscience research* 107: 1–7.
Wiese, Wanja and Karl Friston. 2020. The neural correlates of consciousness under the free energy principle: From computational correlates to computational explanation .
Zanardi, Paolo, Michael Tomka, and Lorenzo Campos Venuti. 2018. Quantum integrated information theory. *arXiv preprint arXiv:1806.01421* .

A. Decompositions and Integration

Here we briefly mention a few further results about decompositions of objects in process theories; we leave a detailed study of their properties to future work.

Our earlier definition of $\mathbb{D}|_A$ was based on an idea of one decomposition as being 'contained in' another. Let us make this precise.

Definition 12 *Let S be an object in a process theory and (A, A'), (B, B') two decompositions. We write that $(A, A') \preceq (B, B')$ whenever there exists an object C and decompositions (A, C) of B and (B', C) of A' such that*

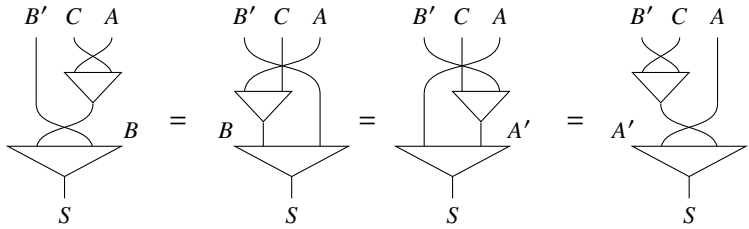


(8)

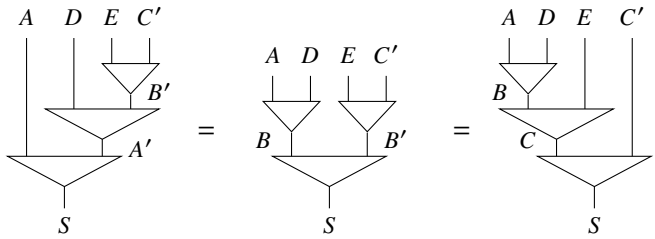
Intuitively, this states that A is contained in B (as is B' within A') in a way compatible with these decompositions.

Lemma 13 *Let S be an object in a process theory. Then $\preceq$ forms a pre-order on the set of decompositions of S, with top element 1 and bottom element 0, and $(-)^{\perp}$ as an involution.*

Proof. We always have $(A, A') \preceq (A, A')$ by taking $C = I$ and using the decompositions 1 and 0 on A in (8). Similarly $(A, A') \preceq 1$ by taking $C = A'$. To see that $(-)^{\perp}$ is an involution, suppose that $(A, A') \preceq (B, B')$ as above. Then we have $(B, B')^{\perp} \preceq (A, A')^{\perp}$ since



Hence we always have $0 = 1^{\perp} \preceq (A, A')$ for all (A, A'). For transitivity, note that whenever $(A, A') \preceq (B, B') \preceq (C, C')$ via some respective objects D, E then we have

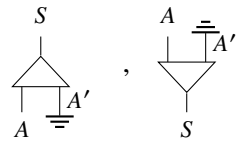


so that $(A, A') \preceq (C, C')$ via the above decompositions $(D \otimes E, C')$ of A' and $(A, D \otimes E)$ of C. ■

Recall that in any category, a *sub-object* of an object A is an (isomorphism class of a) monomorphism $m\colon M \to A$. It is *split* when $e \circ m = \mathrm{id}_M$ for some e. The sub-objects of A form a partial order $\mathrm{Sub}(A)$.

Lemma 14 *In any process theory with* $\overline{\top}, \overline{\perp}$*, for any object* S*:*

1. *Any decomposition* (A, A') *of* S *makes* A *a split sub-object of* S *via*

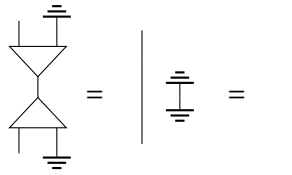


(9)

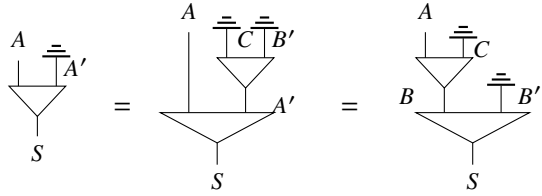
Moreover if $(A, A') \preceq (B, B')$ *then* $A \leq B$ *in* $\mathrm{Sub}(S)$.

2. $\preceq$ *restricts to a partial order* $\leq$ *on* $\mathbb{D}(S)$*, again with top element* 1*, bottom* 0 *and involution* $(-)^{\perp}$.

Proof. 1: We have

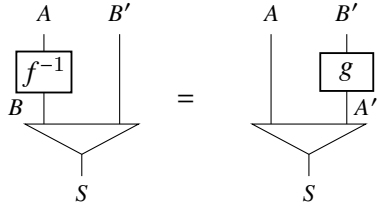


If $(A, A') \preceq (B, B')$ then the splitting for A factors over that for B since:



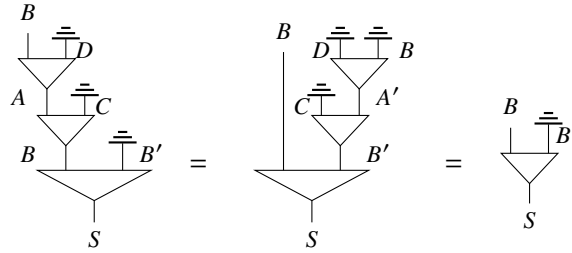
It follows that $A \leq B$ in $\mathrm{Sub}(S)$.

2: We need to show that any two decompositions (A, A') and (B, B') are equivalent under $\preceq$ precisely when they are equivalent in the sense of (1). Firstly, if there exists causal and co-causal isomorphisms f, g making (1) hold, then we have

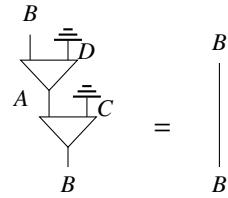


Viewing f^{-1} and g as decompositions (A, I) of B and (I, B') of A', respectively, this gives that $(B, B') \preceq (A, A')$. Then $(A, A') \preceq (B, B')$ holds similarly.

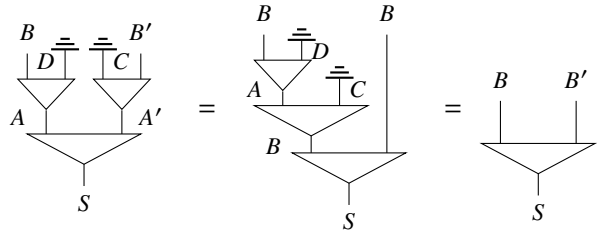
Conversely, if $(A, A') \preceq (B, B') \preceq (A, A')$, via respective objects C, D then



Since the right-hand map is an epimorphism by the first part, this gives that



Dually, composing in the other order gives the identity on A, making these causal and co-causal isomorphisms $A \simeq B$. Similarly we obtain such isomorphisms $A' \simeq B$'. Then we have

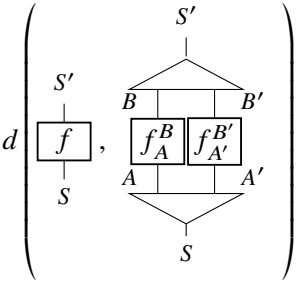


as required. Now 2 follows since any pre-order restricts to a partial order on its set of equivalence classes, and so $\preceq$ becomes a partial order $\leq$ on $\mathbb{D}(S)$. It is easy to see that the earlier properties of $1, 0, (-)^{\perp}$ carry over to $\leq$. ■

A.1 Integration

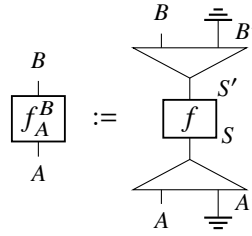
Let us briefly allude to how integration may generally be studied and quantified using decomposition sets.

Suppose we have objects S, S' with given decomposition sets $\mathbb{D}, \mathbb{D}'$ and for each $(A, A') \in \mathbb{D}$ and $(B, B') \in \mathbb{D}'$ a process f_A^B from A to B. We denote $f_S^{S'}$ simply by f. Whenever we have a given distance function d on the set of processes from S to S', we may define the level of *integration* of the family $(f_A^B)_{A,B}$ as

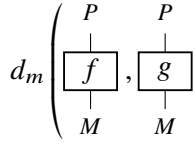
$\phi(f) := \min_{\mathbb{D}\times\mathbb{D}'} d$ 

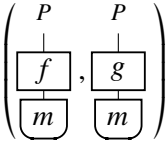
where we exclude the top element $(1, 1)$ of $\mathbb{D} \times \mathbb{D}'$ in the minimisation.

Example 15 Given any process f from S to S' we may define such a family $(f_A^B)_{A,B}$ with $f_S^{S'} = f$ by setting



Example 16 Our earlier description of the IIT algorithm precisely includes evaluating the integration level of each of the families of processes $(\mathsf{caus})_{M,P}$ and $(\mathsf{eff})_{M,P}$ using the state-dependent distance

d_m 

$:= d$ 

where $m = s|_M$ and d is the distance on $\mathrm{St}(S)$.