Integrated Information Theory:
Theoretical Developments & Empirical Applications

By

William G. P. Mayner

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Neuroscience)

at the

UNIVERSITY OF WISCONSIN–MADISON

2023

Date of final oral examination: 8/28/2023

The dissertation is approved by the following members of the Final Oral Committee:

Matthew Banks, Professor, Anesthesiology
Chiara Cirelli, Professor, Psychiatry
Bradley Postle, Professor, Psychology
Giulio Tononi, Professor, Pyschiatry
Justin Williams, Professor, Biomedical Engineering

For Brittany

ACKNOWLEDGMENTS

I am deeply indebted to the many people who have made this work possible.

First, I must give my thanks to Giulio Tononi for taking in an idealistic, passionate, and rather naïve young student, and showing him how real science is done. I also thank the members of my thesis committee for their guidance. To Matthew Banks, I cannot thank you enough for your mentorship and support. To my fellow students in the NTP, thank you for providing a sense of community and camaraderie.

I thank all my collaborators at the Allen Institute for Brain Science, notably Saurabh Gandhi, Yazan Billeh, Jerome Lecoq, Christof Koch, and Anton Arkhipov. To Anton, thank you for your optimism, patience, and graciousness as a leader.

Science is a social endeavor, and I have been fortunate in finding an intellectually vibrant community of brilliant and kind-hearted people in this lab. I am particularly indebted to Larissa Albantakis and William Marshall, who have been invaluable mentors, collaborators, and friends. I could not have asked for better scientific role models. To Bjørn Juel, I am so grateful for your enthusiasm, kindness, and friendship, without which I could not have completed this dissertation. I am also especially grateful to Benjamin Baird, Leonardo Barbosa, Melanie Boly, Tom Bugnon, Benedetta Cecconi, Mattia Chini, Renzo Comolatti, Graham Findlay, Keiko Fujii, Olivia Gosseries, Matteo Grasso, Andrew Haun, Elsa Juan, Jonathan Lang, Sophia Loschky, Robert Marchman, Armand Mensen, Garrett Mindt, Shuntaro Sasai, Francesca Siclari, Joanna Szczotka, Noémi De Wispelaere, and Alireza Zaeemzadeh. To all my colleagues, thank you for your friendship and support throughout my time here.

I am profoundly grateful for my friends, both near and far, and to my community in Madison, for making it such a cherished place to call home. Matt, I am lucky to have found such a kindred spirit on day one. Kacy, thank you for your companionship and inimitable wit. Tom, thank you for *la fraternité* and our adventures in the early years. Fitz, thank you for the fellowship and conversations we shared on our many strolls. Beny, Robby, Joe, Sam, Matt, and the rest of the jam crew, thank you for all the good vibes and funky riffs. Lea, thank you for the levity and good sense you bring to everything. Margaux, thank you for your encouragement and advice. Matteo, thank you for your friendship and for always lending a kind ear. Graham, thank you for inspiring me to take up cooking and cycling. To everyone in the Pod, thank you for being an anchor of happiness in a dark and strange time. Kala, thank you for all your positivity and support. Jonathan, thank you for your comradeship and humor; you are the best storyteller I know. Mattia & Noémi, *grazie mille* for the good times in Madison and in the mountains. Caitlin, Zarmeen, and Ziyad, thank you for being there during the hard times. Liz, thank you for all your support as I started this journey. Kevin, Kaizo, Jonny, Varun, Jaime, and Jesse, thank you for being a source of joy every time we get together again. And to the many others who have touched my life during these years—thank you.

Mom and Dad, I am truly lucky to be your son. I could not have followed this path without you, and I will always be grateful for your love and support during my Ph.D. and throughout my life—and, of course, for the greatest gift of all: consciousness. David, thank you for all your support and for encouraging me to take the first step, all those years ago. Stacey and Gabrielle, thank you for making our home such a happy place. To all my family, thank you for being there for me.

I also wish to remember my grandfather, Gerald Mayner, who had a deep love of science and learning. I know he would be proud.

Finally, I can scarcely find the words to express my love and gratitude for my partner, Brittany. Your wisdom, strength, and humor have sustained me during these last few years. No scientific achievement could ever match the joy and fulfillment I have found in you.

CONTENTS

CHAPTER 1

Introduction

Consciousness is an enduring mystery. Why do some physical systems, such as the brain, support subjective experience, while others do not? What accounts for the particular qualities of an experience? How do these subjective qualities relate to objective phenomena? In recent decades, these fundamental questions have been taken up as subjects of scientific inquiry. A scientific explanation for consciousness would have profound implications in several areas, including the treatment of disorders of consciousness; the use of animals in research and in society at large; the use of brain organoids in research (Lavazza & Massimini, 2018); the ethics of neurotechnology (Goering & Yuste, 2016; Yuste et al., 2017) the ethics of artificial intelligence (Findlay et al., 2019; Tononi & Koch, 2015); the question of free will (Tononi, 2013; Tononi, Albantakis, et al., 2022); and, ultimately, our understanding of our place in the cosmos (Koch, 2019). Yet the emerging science of consciousness still lacks a principled, parsimonious, and comprehensive theory that can fully account for the presence and quality of experience (Cogitate Consortium et al., 2023; Koch et al., 2016).

Integrated information theory (IIT) is a promising candidate for such a theory. It identifies the essential properties of experience (*axioms*), infers the necessary and sufficient properties that its substrate must satisfy (*postulates*), and expresses them mathematically. It aims to thereby account for the properties of experience in physical terms. In principle, the postulates can be applied to any system of units in a state to determine whether it is conscious, to what degree, and in what way. IIT offers a parsimonious explanation of empirical evidence, makes testable predictions concerning both the presence and the quality of experience, and permits inferences and extrapolations (Ellia et al., 2021; Grasso et al., 2021; Haun & Tononi, 2019; Oizumi et al., 2014; Tononi, 2008; Tononi, Albantakis, et al., 2022; Tononi et al., 2016).

Specific aims

Aim 1: Improve and refine IIT's mathematical formalism and develop an open-source software package that implements it.

In Chapter 2, I present *PyPhi*, the open-source Python software package that provides a reference implementation of IIT's mathematical formalism. PyPhi has become the main tool for researchers undertaking theoretical work involving IIT both within our laboratory and in the wider community, and has provided a crucial means of developing and testing the ideas that lead to the latest version of the theory, IIT 4.0.

IIT 4.0 is presented in Chapter 3. It is the result of a concerted collaborative effort and incorporates several developments of the past ten years. These include a more accurate translation of axioms into postulates and mathematical expressions, the introduction of a unique measure of intrinsic information that is consistent with the postulates, an updated treatment of background conditions, and an explicit assessment of causal relations.

Aim 2: Measure area- and layer-specific neurophysiological differentiation with cellular-level resolution in mouse visual cortex.

Despite significant progress in understanding neural coding (Hubel & Wiesel, 1959; Quiroga & Panzeri, 2009), it remains unclear how the coordinated activity of large populations of neurons relates to what an observer actually perceives. Since neurophysiological differences must underlie differences among percepts, *differentiation analysis*—quantifying distinct patterns of neurophysiological activity—has been proposed as an "inside-out" approach that addresses this question. This methodology contrasts with "outside-in" approaches such as feature tuning and decoding analyses, which are defined in terms of extrinsic experimental variables (Brette, 2019; Buzsáki, 2019).

Theoretical work in our laboratory has shown a formal relationship between neurophysiological differentiation and a system's integrated information (Marshall et al., 2016). Empirical studies using functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) in humans have demonstrated that neurophysiological differentiation is greater when subjects view naturalistic stimuli that were subjectively rated as "meaningful" compared to artificial, "meaningless" stimuli (Boly et al., 2015; Mensen et al., 2017, 2018). However, the low spatial resolution of fMRI and EEG precluded identifying the cell populations underlying this correspondence. Also, these studies did not contrast differentiation analysis with stimulus-decoding approaches.

To address these gaps, we used two-photon calcium imaging in mice to systematically survey stimulus-evoked neurophysiological differentiation in excitatory neuronal populations in layers 2/3, 4, and 5 across five visual cortical areas (primary, lateromedial, anterolateral, posteromedial, and anteromedial) in response to naturalistic and phase-scrambled movie stimuli (Chapter 4). Consistent with the previous findings in humans, we found that unscrambled stimuli evoke greater neurophysiological differentiation than scrambled stimuli specifically in layers 2/3 of the anterolateral and anteromedial areas, and that this effect is modulated by arousal state and locomotion. By contrast, decoding performance was far above chance and did not vary substantially across areas and layers. Differentiation also differed within the unscrambled stimulus set, suggesting that differentiation analysis may be used to probe the subjective meaningfulness of individual stimuli at the scale of local neuronal populations.

Aim 3: Extend IIT's mathematical framework to characterize the relationship between intrinsic meaning and environmental stimuli.

IIT argues that the substrate of consciousness is a complex of units that is maximally irreducible. The complex's subsets specify a cause-effect structure, composed of distinctions and their relations, that accounts in full for the quality of experience. The meaning of a specific experience is thus defined intrinsically, regardless of whether the experience occurs in a dream or is triggered by processes in the environment.

But what does it mean for a stimulus to be perceived? To what extent does a given experience refer to something in the environment? In Chapter 5, we address these questions by extending IIT's framework to characterize the relationship between intrinsic meaning and environmental stimuli. First, we define *connectedness* as the extent to which the current state of the system's subsets is triggered by a stimulus. Next, we quantify *perception* as the extent to which the cause-effect structure is triggered by the stimulus. We then quantify the difference between the amount of intrinsic meaning triggered by a stimulus sampled from an environment E and a random stimulus sampled from independent noise sources N. The difference, called *stimulus-specific matching*, measures how much intrinsic meaning refers to regularities in E. Finally, we quantify the differentiation of unique distinctions and relations triggered by a set of stimuli from E and N. The difference, *stimulus sequence matching*, reflects how much the connectivity of a complex has internalized different aspects of causal processes in its environment. Matching thus provides a link between intrinsic meaning and extrinsic reference.

Integrated information theory

IIT was proposed by Tononi in 2004 and has been extended and refined in several iterations since, though the core ideas have remained the same (Balduzzi & Tononi, 2008, 2009; Oizumi et al., 2014; Tononi, 2008). It takes a novel approach to the "hard problem" (Chalmers, 1995) of providing a scientific explanation for consciousness: it starts from experience itself and identifies phenomenal properties that are essential to it (*axioms*), and then infers a corresponding set of physical properties that its physical substrate must possess (*postulates*). These physical properties are then expressed mathematically. The result is an algorithm that takes as input a discrete dynamical system and a complete description of the interactions of the system's elements, and yields the system's cause-effect structure (the set of irreducible causal distinctions and relations specified by subsets of the system) and a measure of this structure's irreducibility, a non-negative scalar quantity referred to as integrated information. IIT thus enables precise quantitative predictions about which physical systems are substrates of consciousness and what they are conscious of.

IIT proposes an fundamental explanatory identity between experience and the cause-effect structure specified by a maximally-irreducible substrate. According to IIT, if a substrate is maximally irreducible, then it has subjective experience, and the qualities of the experience—'what it is like' to be that system (Nagel, 1974)—are in one-to-one correspondence with the properties of the cause-effect structure it specifies. If a substrate is not maximally irreducible, then there is nothing it is like to be that substrate. This proposed identity and the phenomenological axioms undergirding it are what provide the "materials for the bridge" (Chalmers, 1995) that can span the "explanatory gap" (Levine, 1983) between experience and its physical substrate.

IIT offers principled and parsimonious explanations for various observations about consciousness. It can account for why the cerebellum appears to be unnecessary for the brain to support conscious experience (Boyd, 2010; Yu et al., 2015) despite having more neurons than cerebral cortex: the cerebellum has a modular, parallel, and largely feed-forward architecture that likely lacks the functional integration required for a high Φ value, in contrast to the functional specialization and recurrent connectivity of the cerebral cortex. IIT also suggests an explanation for the fading of consciousness in dreamless sleep, anesthesia, epilepsy, and coma. In wakefulness, intracortical electrical stimulation induces a chain of deterministic phase-locked activity, while in non-rapid eye movement (NREM) sleep, the response to the stimulation lacks a strong phase-locked component (Pigorini et al., 2015). This suggests that the bistable fluctuations between neuronal silence and wake-like activity that are characteristic of NREM sleep prevent sustained causal interactions among cortical areas, and that this correlates with the lack of subjective experience. Indeed, a study (Nieminen et al., 2016) that focused exclusively on dream reports upon awakening from NREM sleep (which are not uncommon; Nielsen, 1999, p. 107; Siclari et al., 2017) found that transcranial magnetic stimulation (TMS) pulses evoked a larger negative deflection (an indication of neuronal bistability) and shorter phase-locked response when subjects did not report a dream. The loss of sustained deterministic responses across cortex is likely associated with a low value of integrated information (Massimini et al., 2012).

Integrated information can be thought of as a measure of complexity; it is high for systems that are both integrated (each part influences many other parts) and differentiated (the system has a large repertoire of states), and low for systems that are either reducible to independent parts or that have few states. These considerations were applied in the development of the perturbational complexity index (PCI), which measures the Lempel-Ziv complexity (an approximation of lossless compressibility) of the EEG response to a TMS pulse (Casali et al., 2013). By design, PCI is sensitive to the joint presence of integration and differentiation and thus can reasonably be interpreted as a proxy for integrated information. PCI has consistently shown a remarkable sensitivity and specificity for detecting the presence of consciousness in a graded manner across diverse conditions: in healthy individuals during wakefulness, dreaming, dreamless sleep, and anesthesia (ketamine, midazolam, xenon, and propofol), and in brain-injured patients with clinical diagnoses ranging from unresponsive wakefulness syndrome to locked-in syndrome (Casali et al., 2013; Casarotto et al., 2016; Comolatti et al., 2019).

Differentiation analysis

Differentiation—the degree to which a system expresses a rich and varied repertoire of states—and integration—a system's irreducibility to its components—are common threads that run through the theoretical developments that have culminated in IIT's current form. An early proposal (Tononi et al., 1994) put forward a measure called neural complexity, defined as the average mutual information between bipartitions of a system, that attempted to capture the balance between functional segregation (the division of the brain into functionally specialized structures) and integration and suggested a fundamental role for these two principles of brain organization in supporting complex dynamics. The dynamic core hypothesis, proposed in subsequent work by Tononi and Edelman (1998), emphasized that differentiation and integration are properties of conscious experience itself, and suggested that the substrate of consciousness might therefore be identified with the functional cluster of neural ensembles that simultaneously maximizes differentiation and integration: the "dynamic core." In this proposal, differentiation was measured with neural complexity, while integration was measured with a related measure termed functional clustering (Tononi et al., 1998). These ideas led to the first iteration of IIT, initially described by Tononi and Sporns (2003) and then fully expounded as a theory of consciousness by Tononi (2004).

At this point, a key feature of IIT was introduced: the causal aspect of the analysis. In the dynamic core hypothesis, integration and differentiation are defined in terms of statistical dependencies in the observed dynamics of a system; in IIT, by contrast, it is causal dependencies that are important. This shift in emphasis represents a significant development as it requires a perturbational approach that is defined in terms of causal interventions (Pearl, 2000; Pearl, 2009) rather than statistical correlations. Furthermore, while the earlier formalisms involve averages across many states of the system, integrated information is defined for a particular state. Consequently, differentiation is not directly invoked in IIT's formalism (by contrast, neural complexity is defined directly in terms of the system's entropy and is trivially bounded by it).

However, differentiation is nonetheless deeply related to integrated information. Intuitively, low-entropy systems without rich dynamics will not exhibit the complex causal constraints captured by IIT's measure. This intuition was formalized and proved by Marshall et al. (2016). The authors define two measures of differentiation: $\mathcal{D}_1$, the number of potential states that the system can transition into following every possible perturbation, and $\mathcal{D}_2$, the cumulative variance of the system's elements. They establish an upper bound for the average integrated information in terms of $\mathcal{D}_1$ and $\mathcal{D}_2$. Moreover, they prove that the probability of observing a state in which integrated information differs from the average decreases as the size of the system increases.

These results provide a theoretical link between integrated information and the differentiation of observed states: a system with high integrated information necessarily exhibits high differentiation (the converse is not true, however; a non-integrated system can be highly differentiated). Crucially, this implies that differentiation can be used empirically as a proxy for integrated information when integration can be assumed. In the intact cortex, this assumption is reasonable, given its remarkable functional and anatomical connectivity (Sporns et al., 2005). Indeed, several studies have revealed a positive correlation between various measures of differentiation and the presence of consciousness (Barttfeld et al., 2015; Gosseries et al., 2011; Hudetz et al., 2014, 2016; Sarà et al., 2011). For example, a study using cellular-resolution techniques found that in both humans and mice, loss of consciousness during anesthesia was associated with a reduction in several measures of differentiation at the level of local neural populations (Wenzel et al., 2019).

Neurophysiological differentiation can also be measured in relation to specific stimuli. Several studies from our laboratory have investigated neurophysiological differentiation in humans using fMRI and EEG (Boly et al., 2015; Mensen et al., 2017, 2018). Boly et al. (2015) used fMRI to analyze the differentiation of the blood-oxygen-level-dependent (BOLD) signal in response to three movie stimuli of varying levels of subjective meaningfulness: a silent film, a spatially shuffled version of the film, and white noise. The authors computed the Lempel-Ziv complexity of the responses and estimated both neural complexity (Tononi et al., 1994) and Φ^* (the difference between the mutual information between the previous and current state of the system when it is considered as a whole vs. as independent elements; see Oizumi, Amari, et al., 2016). All measures were all highest for the unscrambled film, intermediate for the shuffled film, and lowest for white noise. Importantly, these results could not be explained by overall activity levels, which were comparable across conditions, nor by the differentiation of the stimuli themselves, which was ordered oppositely to that of the BOLD activity patterns.

Similar results were obtained in later studies using EEG. Mensen et al. (2017) recorded event-related potentials of presentations of images of natural scenes containing recognizable objects (‘meaningful’ condition) and artificially generated images, including phase-scrambled versions of the ‘meaningful’ stimuli and white noise (‘meaningless’ condition). Differentiation was computed as the mean pairwise within-category Euclidean distance and was found to be higher for the responses to meaningful stimuli. Moreover, differentiation correlated with subjects’ ratings of the meaningful differences within each category. A follow-up study (Mensen et al., 2018) used continuous audiovisual movie stimuli and analyzed differentiation at the single-trial level. Differentiation was highest for movies containing naturalistic structure (‘meaningful’ condition) compared to phase-scrambled and white noise movies, and furthermore was correlated with subjective ratings of the ‘meaningfulness’ of the stimuli and subjective estimates of how many distinct experiences they elicited.

Taken together, this body of work suggests that in addition to serving as an empirical proxy for integrated information, analysis of neurophysiological differentiation can be used to track the subjective meaningfulness of a stimulus sequence. This was borne out in our calcium imaging experiments in mice (Chapter 4), in which neurophysiological differentiation tracked putative stimulus meaningfulness within specific cortical layers and areas. Chapter 5 presents a novel formalism that provides a theoretical foundation for this correspondence. Importantly, differentiation analysis does not rely on behavioral report, and may thus support the inference of consciousness in cases where no report is available, such as in patients with disorders of consciousness, or in animals.

CHAPTER 2

PyPhi

A toolbox for integrated information theory

William G. P. Mayner[1,2], William Marshall[2], Larissa Albantakis[2], Graham Findlay[1,2], Robert Marchman[2], Giulio Tononi[2]

1 Neuroscience Training Program, University of Wisconsin–Madison, Madison, WI, USA

2 Psychiatry Department, University of Wisconsin–Madison, Madison, WI, USA

This chapter has been previously published as:

> Mayner, W. G. P., Marshall, W., Albantakis, L., Findlay, G., Marchman, R., & Tononi, G. (2018). PyPhi: A toolbox for integrated information theory. *PLoS computational biology*, *14*(7), e1006343

Acknowledgments

We thank Andrew Haun, Leonardo Barbosa, Sabrina Streipert, and Erik Hoel for their valuable feedback as early users of the software.

WGPM, WM, LA, GF, RM and GT received funding from the Templeton World Charity Foundation (Grant #TWCF 0067/AB41). WGPM and GF were also supported by National Research Service Award (NRSA) T32 GM007507. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

2.1 Introduction

Integrated information theory (IIT) has been proposed as a theory of consciousness. The central hypothesis is that a physical system has to meet five requirements (‘postulates’) in order to be a physical substrate of subjective experience: (1) *intrinsic existence* (the system must be able to make a difference to itself);, (2) *composition* (it must be composed of parts that have causal power within the whole);, (3) *information* (its causal power must be specific);, (4) *integration* (its causal power must not be reducible to that of its parts); and, and (5) *exclusion* (it must be maximally irreducible) (Balduzzi & Tononi, 2008; Oizumi et al., 2014; Tononi, 2004, 2015; Tononi et al., 2016).

From these postulates, IIT develops a mathematical framework to assess the cause-effect structure (CES) of a physical system that is applicable to discrete dynamical systems. This framework has proven useful not only for the study of consciousness but has also been applied in research on complexity (Albantakis & Tononi, 2015, 2017; Albantakis et al., 2014; Oizumi, Tsuchiya, & Amari, 2016), emergence (Hoel et al., 2013, 2016; Marshall et al., 2018), and certain biological questions (Marshall et al., 2017).

The main measure of cause-effect power, *integrated information* (denoted Φ), quantifies how irreducible a system’s CES is to those of its parts. Φ also serves as a general measure of complexity that captures to what extent a system is both integrated (Albantakis & Tononi, 2015) and differentiated (informative) (Marshall et al., 2016).

Here we describe *PyPhi*, a Python software package that implements IIT’s framework for causal analysis and unfolds the full CES of discrete Markovian dynamical systems of binary elements. The software allows users to easily study these CESs and serves as an up-to-date reference implementation of the formalisms of IIT.

Details of the mathematical framework are published elsewhere (Oizumi et al., 2014; Tononi et al., 2016); in § 2.2 Results, we describe the output and input of the software and give an overview of the main algorithm in the course of reproducing results obtained from an example system studied in (Oizumi et al., 2014). In § 2.3 Design and implementation, we discuss specific issues concerning the algorithm’s implementation. Finally in § 2.4 Availability and future directions, we describe how the software can be obtained and discuss new functionality planned for future versions.

2.2 Results

Output

The software has two primary functions: (1) to unfold the full CES of a discrete dynamical system of interacting elements and compute its Φ value, and, and (2) to compute the maximally-irreducible cause-effect repertoires of a particular set of elements within the system. The first is function is implemented by `pyphi.compute.major_complex()`, which returns a `SystemIrreducibilityAnalysis` object (Figure 2.1A). The system's CES is contained in the '`ces`' attribute and its Φ value is contained in '`phi`'. Other attributes are detailed in the online documentation.

The CES is composed of `Concept` objects, which are the output of the second main function: `Subsystem.concept()` (Figure 2.1B). Each `Concept` is specified by a set of elements within the system (contained in its '`mechanism`' attribute). A `Concept` contains a maximally-irreducible cause and effect repertoire ('`cause_repertoire`' and '`effect_repertoire`'), which are probability distributions that capture how the mechanism elements in their current state constrain the previous and next state of the system, respectively; a φ value ('`phi`'), which measures the irreducibility of the repertoires; and several other attributes discussed below and detailed in the online documentation.

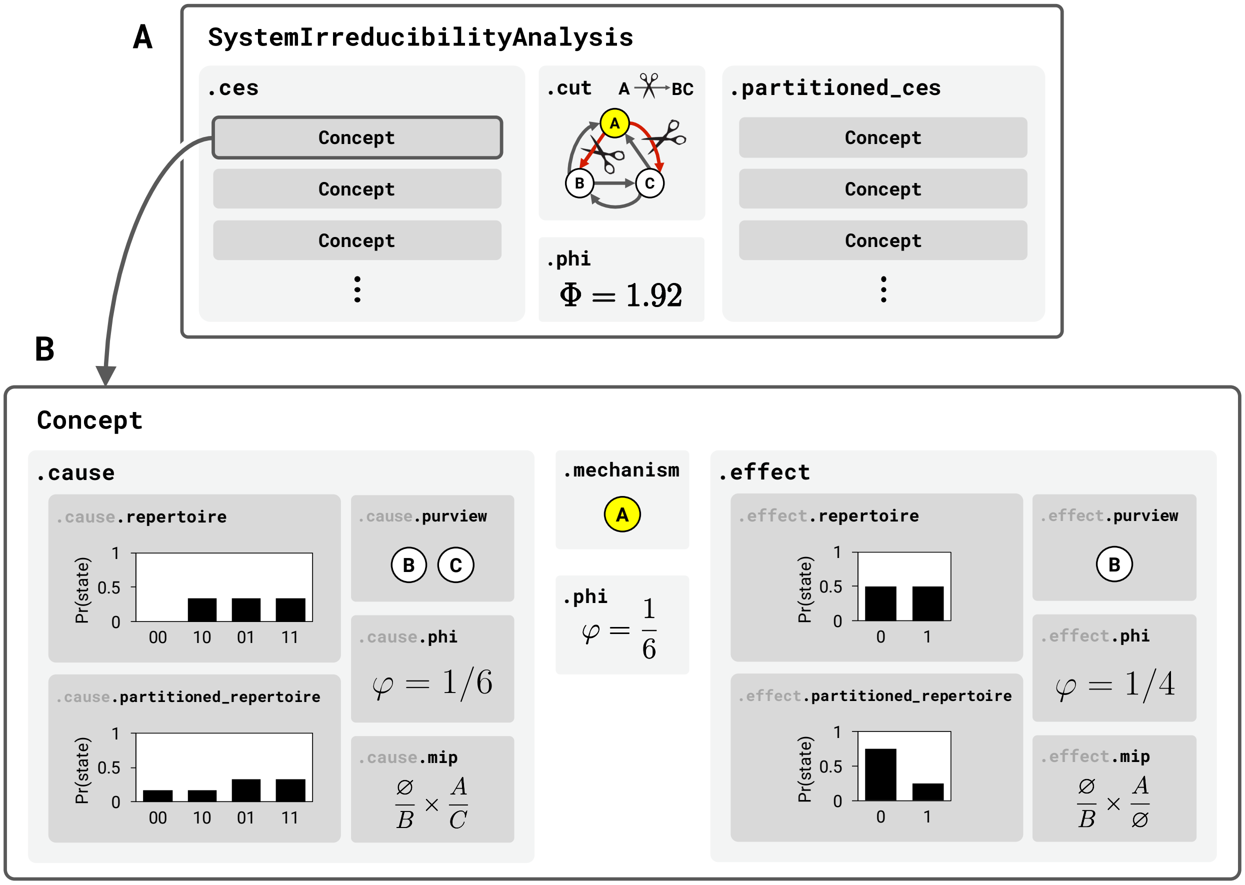
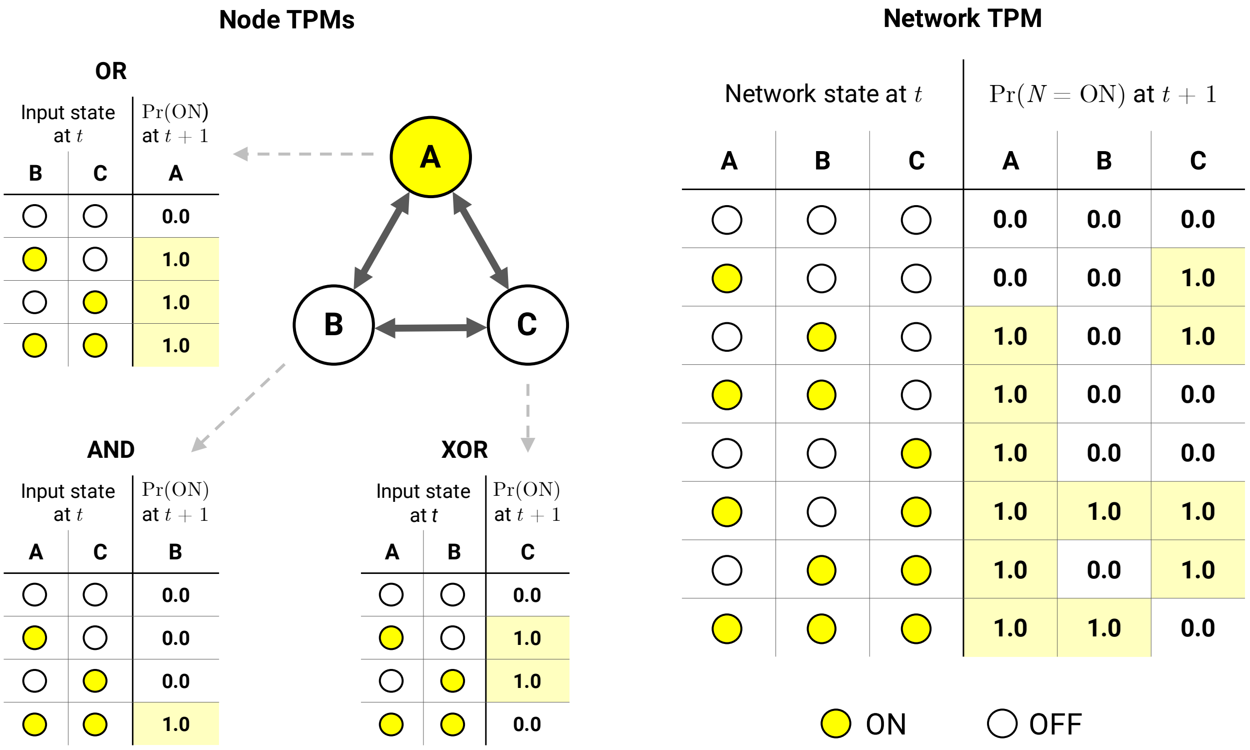


Figure 2.1. Output. (A) The `SystemIrreducibilityAnalysis` object is the main output of the software. It represents the results of the analysis of the system in question. It has several attributes (grey boxes): '`ces`' is a `CauseEffectStructure` object containing all of the system's `Concepts`; '`cut`' is a `Cut` object that represents the minimum-information partition (MIP) of the system (the partition of the system that makes the least difference to its CES); '`partitioned_ces`' is the `CauseEffectStructure` of `Concepts` specified by the system after applying the MIP; and '`phi`' is the Φ value, which measures the difference between the unpartitioned and partitioned CES. **(B)** A `Concept` represents the maximally-irreducible cause (MIC) and maximally-irreducible effect (MIE) of a mechanism in a state. The '`mechanism`' attribute contains the indices of the mechanism elements. The '`cause`' and '`effect`' attributes contain `MaximallyIrreducibleCause` and `MaximallyIrreducibleEffect` objects that describe the mechanism's MIC and MIE, respectively; each of these contains a purview, repertoire, MIP, partitioned repertoire, and φ value. The '`phi`' attribute contains the concept's φ value, which is the minimum of the φ values of the MIC and MIE.

Input

The starting point for the IIT analysis is a discrete Markovian dynamical system S composed of n interacting elements. Such a system can be represented by a directed graph of interconnected nodes, each equipped with a function that outputs the node's state at the next timestep $t + 1$ given the state of its parents at the previous timestep t (Figure 2.2). At present, PyPhi can analyze both deterministic and stochastic discrete Markovian dynamical systems consisting of elements with two states.



OR

Input state at t		$\Pr(\text{ON})$ at $t+1$
B	C	A
○	○	0.0
●	○	1.0
○	●	1.0
●	●	1.0

AND

Input state at t		$\Pr(\text{ON})$ at $t+1$
A	C	B
○	○	0.0
●	○	0.0
○	●	0.0
●	●	1.0

XOR

Input state at t		$\Pr(\text{ON})$ at $t+1$
A	B	C
○	○	0.0
●	○	1.0
○	●	1.0
●	●	0.0

Network state at t			$\Pr(N = \text{ON})$ at $t+1$		
A	B	C	A	B	C
○	○	○	0.0	0.0	0.0
●	○	○	0.0	0.0	1.0
○	●	○	1.0	0.0	1.0
●	●	○	1.0	0.0	0.0
○	○	●	1.0	0.0	0.0
●	○	●	1.0	1.0	1.0
○	●	●	1.0	0.0	1.0
●	●	●	1.0	1.0	0.0

● ON ○ OFF

Figure 2.2. A network of nodes and its TPM. Each node has its own TPM—in this case, the truth-table of a deterministic logic gate. Yellow signifies the "ON" state; white signifies "OFF". The system's TPM (right) is composed of the TPMs of its nodes (left), here shown in state-by-node form (see § 2.3, Representation of the TPM and probability distributions). Note that in PyPhi's TPM representation, the first node's state varies the fastest, according to the little-endian convention (see § 2.3, 2-dimensional state-by-node form).

Such a discrete dynamical system is completely specified by its transition probability matrix (TPM), which contains the probabilities of all state transitions from t to $t+1$. It can be obtained from the graphical representation of the system by perturbing the system into each of its possible states and observing the following state at the next timestep (for stochastic systems, repeated trials of perturbation/observation will yield the probabilities of each state transition). In PyPhi, the TPM is the fundamental representation of the system.

Formally, if we let S_t be the random variable of the system state at t, the TPM specifies the conditional probability distribution over the next state S_{t+1} given each current state s_t:

$$\Pr(S_{t+1} \mid S_t = s_t), \quad \forall\, s_t \in \Omega_S,$$

where Ω_S denotes the set of possible states. Furthermore, given a marginal distribution over the previous states of the system, the TPM fully specifies the joint distribution over state transitions. Here

IIT imposes uniformity on the marginal distribution of the previous states because the aim of the analysis is to capture direct causal relationships across a single timestep without confounding factors, such as influences from system states before $t - 1$ (Albantakis et al., 2019; Ay & Polani, 2008; Balduzzi & Tononi, 2008; Hoel et al., 2013; Oizumi et al., 2014). The marginal distribution thus corresponds to an interventional (causal), not observed, state distribution.

Moreover, IIT assumes that there is no instantaneous causation; that is, it is assumed that the elements of the dynamical system influence one another only from one timestep to the next. Therefore we require that the system satisfies the following Markov condition, called the *conditional independence property*: each element's state at $t + 1$ must be independent of the state of the others, given the state of the system at t (Pearl, 2009),

$$\Pr(S_{t+1} \mid S_t = s_t) \;=\; \prod_{N \in S} \Pr(N_{t+1} \mid S_t = s_t)\;, \quad \forall\, s_t \in S. \tag{2.1}$$

For systems of binary elements, a TPM that satisfies (2.1) can be represented in state-by-node form (Figure 2.2, right), since we need only store each element's marginal distribution rather than the full joint distribution.

In PyPhi, the system under analysis is represented by a `Network` object. A `Network` is created by passing its TPM as the first argument: `network = pyphi.Network(tpm)` (see § 2.2, Setup). Optionally, a connectivity matrix (CM) can also be provided, where

$$[\mathrm{CM}]_{i,j} = \begin{cases} 1 & \text{if there is an edge from element } i \text{ to element } j \\ 0 & \text{otherwise,} \end{cases}$$

via the `cm` keyword argument: `network = pyphi.Network(tpm, cm=cm)`. Because the TPM completely specifies the system, providing a CM is not necessary; however, explicit connectivity information can be used to make computations more efficient, especially for sparse networks, because PyPhi can rule out certain causal influences *a priori* if there are missing connections (see § 2.3, Connectivity optimizations). Note that this means providing an incorrect CM can result in inaccurate output. If no CM is given, PyPhi assumes full connectivity; *i.e.*, it assumes each element may have an effect on any other, which guarantees correct results.

Once the `Network` is created, a subset of elements within the system (called a *candidate system*), together with a particular system state, can be selected for analysis by creating a `Subsystem` object. Hereafter we refer to a candidate system as a *subsystem*.

Demonstration

The mathematical framework of IIT is typically formulated using graphical causal models as representations of physical systems of elements. The framework builds on the causal calculus of the $do(\cdot)$ operator introduced by Pearl (Pearl, 2009). In order to assess causal relationships among the elements, interventions (manipulations, perturbations) are used to actively set elements into a specific state, after which the resulting state transition is observed.

For reference, we define a set of graphical operations that are used during the IIT analysis. To *fix* an element is to use system interventions to keep it in the same state for every observation. To *noise* an element is to use system interventions to set it into a state chosen uniformly at random. Finally, to *cut* a connection from a source element to a target element is to make the source appear noised to the target, while the remaining, uncut connections from the source still correctly transmit its state.

In this section we demonstrate some of the capabilities of the software by unfolding the CES of a small deterministic system of logic gates as described in (Oizumi et al., 2014) while describing how the algorithm is implemented in terms of TPM manipulations, which we link to the graphical operations defined above. A schematic of the algorithm is shown in Figure 2.3 and Figure 2.4, and a more detailed illustration is given in Calculating Φ.

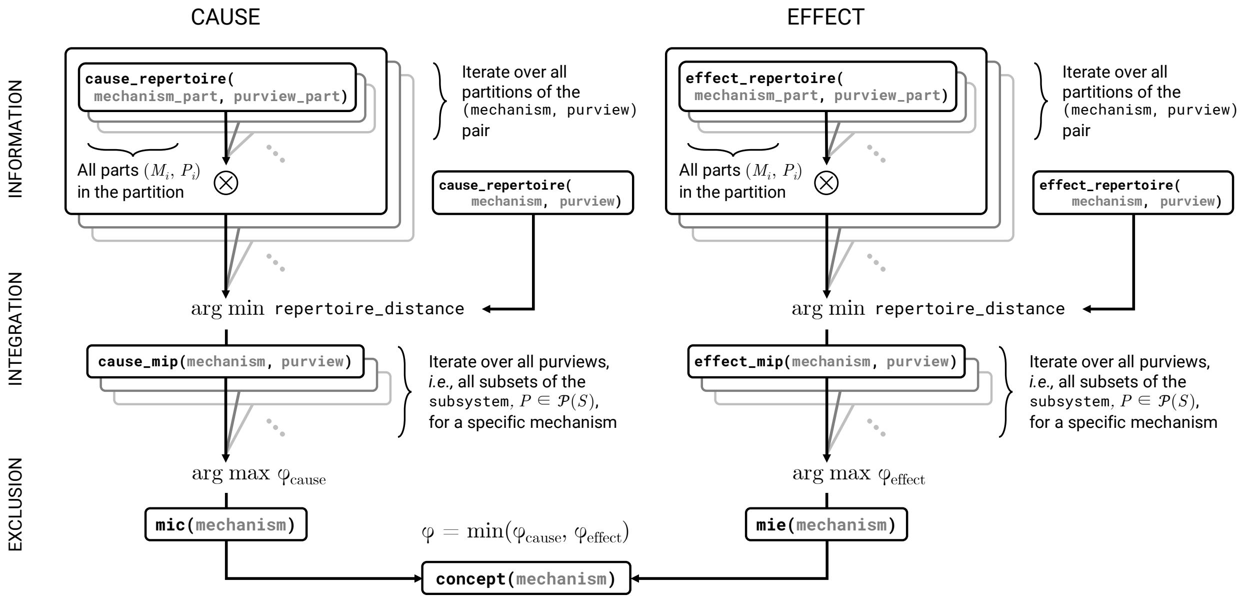

CAUSE
EFFECT
INFORMATION
INTEGRATION
EXCLUSION
cause_repertoire(mechanism_part, purview_part)
All parts (M_i, P_i) in the partition
Iterate over all partitions of the (mechanism, purview) pair
cause_repertoire(mechanism, purview)
arg min repertoire_distance
cause_mip(mechanism, purview)
Iterate over all purviews, i.e., all subsets of the subsystem, $P \in \mathcal{P}(S)$, for a specific mechanism
arg max φ_{cause}
mic(mechanism)
effect_repertoire(mechanism_part, purview_part)
All parts (M_i, P_i) in the partition
Iterate over all partitions of the (mechanism, purview) pair
effect_repertoire(mechanism, purview)
arg min repertoire_distance
effect_mip(mechanism, purview)
Iterate over all purviews, i.e., all subsets of the subsystem, $P \in \mathcal{P}(S)$, for a specific mechanism
arg max φ_{effect}
mie(mechanism)
$\varphi = \min(\varphi_{\text{cause}}, \varphi_{\text{effect}})$
concept(mechanism)

Figure 2.3. Algorithm schematic at the mechanism level. PyPhi functions are named in boxes, with arguments in grey. Arrows point from callee to caller. Functions are organized by the postulate they correspond to (left). $\otimes$ denotes the tensor product; $\mathcal{P}$ denotes the power set.

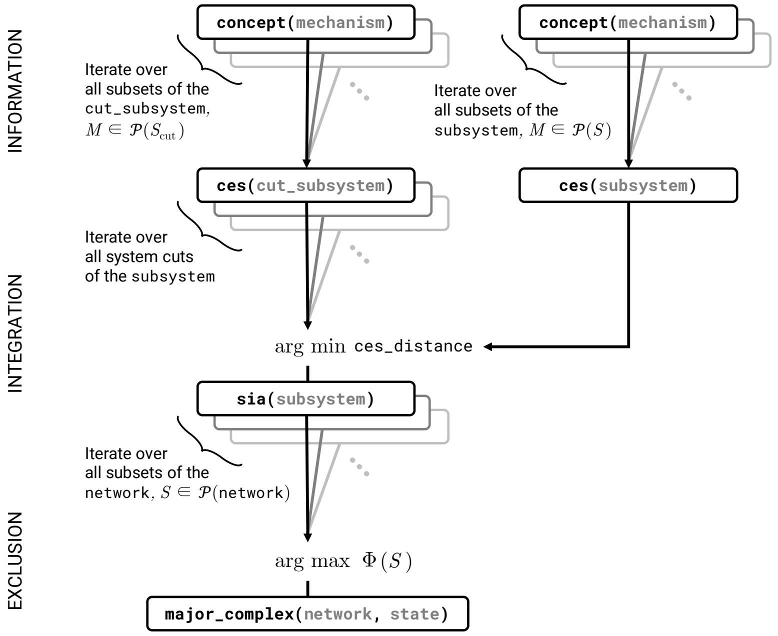

INFORMATION
INTEGRATION
EXCLUSION
concept(mechanism)
Iterate over all subsets of the cut_subsystem, $M \in \mathcal{P}(S_{\text{cut}})$
concept(mechanism)
Iterate over all subsets of the subsystem, $M \in \mathcal{P}(S)$
ces(cut_subsystem)
ces(subsystem)
Iterate over all system cuts of the subsystem
arg min ces_distance
sia(subsystem)
Iterate over all subsets of the network, $S \in \mathcal{P}$(network)
arg max $\Phi(S)$
major_complex(network, state)

Figure 2.4. Algorithm schematic at the system level. PyPhi functions are named in boxes, with arguments in grey. Arrows point from callee to caller. Functions are organized by the postulate they correspond to (left). $\mathcal{P}$ denotes the power set.

Setup

The first step is to create the `Network` object. Here we choose to provide the TPM in 2-dimensional state-by-node form (see § 2.3, 2-dimensional state-by-node form). The TPM is the only required argument, but we provide the CM as well, since we know that there are no self-loops in the system and PyPhi will use this information to speed up the computation. We also label the nodes A, B, and C to make the output easier to read.

```
>>> import pyphi
>>> import numpy as np
>>> tpm = np.array([
...     [0.0, 0.0, 0.0],
...     [0.0, 0.0, 1.0],
...     [1.0, 0.0, 1.0],
...     [1.0, 0.0, 0.0],
...     [1.0, 0.0, 0.0],
...     [1.0, 1.0, 1.0],
...     [1.0, 0.0, 1.0],
...     [1.0, 1.0, 0.0]
... ])
>>> cm = np.array([
...     [0, 1, 1],
...     [1, 0, 1],
...     [1, 1, 0],
... ])
>>> network = pyphi.Network(tpm, cm=cm, node_labels=['A', 'B', 'C'])
```

We select a subsystem and a system state for analysis by creating a `Subsystem` object. System states are represented by tuples of `1`s and `0`s, with `1` meaning "ON" and `0` meaning "OFF." In this case we will analyze the entire system, so the subsystem will contain all three nodes. The nodes to include can be specified with either their labels or their indices (note that in other PyPhi functions, nodes must be specified with their indices).

```
>>> state = (1, 0, 0)
>>> nodes = ('A', 'B', 'C')
>>> subsystem = pyphi.Subsystem(network, state, nodes)
```

If there are nodes outside the subsystem, they are considered as *background conditions* for the causal analysis (Oizumi et al., 2014). In the graphical representation of the system, the background conditions are *fixed* in their current state while the subsystem is perturbed and observed in order to derive its TPM. In the TPM representation, the equivalent operation is performed by *conditioning* the system

TPM on the state at t of the nodes outside the subsystem and then *marginalizing out* those nodes at $t+1$ (illustrated in Calculating Φ). In PyPhi, this is done when the subsystem is created; the subsystem TPM can be accessed with the `tpm` attribute, *e.g.* `subsystem.tpm`.

Cause/effect repertoires (mechanism-level information)

The lowest-level objects in the CES of a system are the *cause repertoire* and *effect repertoire* of a set of nodes within the subsystem, called a *mechanism*, over another set of nodes within the subsystem, called a *purview* of the mechanism. The cause (effect) repertoire is a probability distribution that captures the information specified by the mechanism about the purview by describing how the previous (next) state of the purview is constrained by the current state of the mechanism.

In terms of graphical operations, the effect repertoire is obtained by (1) *fixing* the mechanism nodes in their state at t;, (2) *noising* the non-mechanism nodes at time t, so as to remove their causal influence on the purview; and, and (3) observing the resulting state transition from t to $t+1$ while ignoring the state at $t+1$ of non-purview nodes, in order to derive a distribution over purview states at $t+1$.

The cause repertoire is obtained similarly, but in that case, the purview nodes at time $t-1$ are noised, and the resulting state transition from $t-1$ to t is observed while ignoring the state of non-mechanism nodes. Bayes' rule is then applied, resulting in a distribution over purview states at $t-1$. The corresponding operations on the TPM are detailed in § 2.3, Calculation of cause/effect repertoires from the TPM , and illustrated in Calculating Φ.

Note that, operationally, we enforce that each input from a noised node conveys *independent* noise during the perturbation/observation step. In this way, we avoid counting correlations from outside the mechanism-purview pair as constraints due to the current state of the mechanism. Graphically, this process would correspond to replacing each noised node that is a parent of multiple purview nodes (for the effect repertoire) or mechanism nodes (for the cause repertoire) with multiple, independent "virtual nodes" (as in Oizumi et al., 2014, Supplementary Methods). However, the equivalent definition of repertoires in (2.3) and (2.5) obviates the need to actually implement virtual nodes in PyPhi.

With the `cause_repertoire()` method of the `Subsystem`, we can obtain the cause repertoire of, for example, mechanism A over the purview ABC depicted in Figure 4 of (Oizumi et al., 2014):

```
>>> A, B, C = subsystem.node_indices
>>> print(subsystem.state)
(1, 0, 0)
>>> mechanism = (A,)
>>> purview = (A, B, C)
>>> cr = subsystem.cause_repertoire(mechanism, purview)
>>> print(cr)
[[[ 0.          0.16666667]
  [ 0.16666667  0.16666667]]

 [[ 0.          0.16666667]
  [ 0.16666667  0.16666667]]]
```

We see that mechanism A in its current state, ON (`1`), specifies information by ruling out the previous states in which B and C are OFF (`0`). That is, the probability that either `(0, 0, 0)` or `(1, 0, 0)` was the previous state, given that A is currently ON, is zero:

```
>>> print(cr[(0, 0, 0)])
0.0
>>> print(cr[(1, 0, 0)])
0.0
```

Note that repertoires are returned in multidimensional form, so they can be indexed with state tuples as above. Repertoires can be reshaped to be 1-dimensional if needed, *e.g.* for plotting, but care must be taken that NumPy's FORTRAN (column-major) ordering is used so that PyPhi's little-endian convention for indexing states is respected (see § 2.3, 2-dimensional state-by-node form). PyPhi provides the `pyphi.distribution.flatten()` function for this:

```
>>> flat_cr = pyphi.distribution.flatten(cr)
>>> print(flat_cr)
[ 0.          0.          0.16666667  0.16666667  0.16666667  0.16666667
  0.16666667  0.16666667]
```

Minimum-information partitions (mechanism-level integration)

Having assessed the information of a mechanism over a purview, the next step is to assess its *integrated information* (denoted φ) by quantifying the extent to which the cause and effect repertoires of the mechanism-purview pair can be reduced to the repertoires of its parts.

In terms of graphical operations, the irreducibility of a mechanism-purview pair is tested by partitioning it into parts and *cutting* the connections between them. By applying the perturbation/observation procedure after cutting the connections we obtain a *partitioned repertoire*. Since the partition renders the parts independent, in terms of TPM manipulations, the partitioned repertoire can be calculated as the product of the individual repertoires for each of the parts. If the partitioned repertoire is no different than the original unpartitioned repertoire, then the mechanism as a whole did not specify integrated information about the purview. By contrast, if a repertoire cannot be factored in this way, then some of its selectivity is due to the causal influence of the mechanism *as an integrated whole* on the purview, and the repertoire is said to be *irreducible.*

The amount of irreducibility of a mechanism over a purview with respect to a partition is quantified as the distance between the unpartitioned repertoire and the partitioned repertoire (calculated with `pyphi.distance.repertoire_distance()`). There are many ways to divide the mechanism and purview into parts, so the irreducibility is measured for every partition and the partition that yields the minimum irreducibility is called the *minimum-information partition* (MIP). The integrated information (φ) of a mechanism-purview pair is the distance between the unpartitioned repertoire and the partitioned repertoire associated with the MIP. PyPhi supports several distance measures and partitioning schemes (see § 2.3, Configuration).

The MIP search procedure is implemented by the `Subsystem.cause_mip()` and `Subsystem.effect_mip()` methods. Each returns a `RepertoireIrreducibilityAnalysis` object that contains the MIP, as well as the φ value, mechanism, purview, temporal direction (cause or effect), unpartitioned repertoire, and partitioned repertoire. For example, we compute the effect MIP of mechanism ABC over purview ABC from Figure 6 of (Oizumi et al., 2014) as follows:

```
>>> mechanism = (A, B, C)
>>> purview = (A, B, C)
>>> mip = subsystem.effect_mip(mechanism, purview)
>>> print(mip)
Repertoire irreducibility analysis
  φ = 1/4
  Mechanism: [A, B, C]
  Purview = [A, B, C]
  Direction: EFFECT
  Partition:
     Ø    A,B,C
    ─── × ─────
     B     A,C
  Repertoire:
    ┌───────────────┐
    │  S     Pr(S)  │
    │ ------------- │
    │ 000    0      │
    │ 100    0      │
    │ 010    0      │
    │ 110    0      │
    │ 001    1      │
    │ 101    0      │
    │ 011    0      │
    │ 111    0      │
    └───────────────┘
  Partitioned repertoire:
    ┌───────────────┐
    │  S     Pr(S)  │
    │ ------------- │
    │ 000    0      │
    │ 100    0      │
    │ 010    0      │
    │ 110    0      │
    │ 001    3/4    │
    │ 101    0      │
    │ 011    1/4    │
    │ 111    0      │
    └───────────────┘
```

Here we can see that the MIP attempts to factor the repertoire of ABC over ABC into the product of the repertoire of ABC over AC and the repertoire of the empty mechanism $\varnothing$ over B. However, the repertoire cannot be factored in this way without information loss; the distance between the unpartitioned and partitioned repertoire is nonzero ($\varphi = 1/4$). Thus mechanism ABC over the purview ABC is irreducible.

Maximally-irreducible cause-effect repertoires (mechanism-level exclusion)

Next, we apply IIT’s postulate of exclusion at the mechanism level by finding the *maximally-irreducible cause* (MIC) and *maximally irreducible effect* (MIE) specified by a mechanism. This is done by searching over all possible purviews for the `RepertoireIrreducibilityAnalysis` object with the maximal φ value. The `Subsystem.mic()` and `Subsystem.mie()` methods implement this search procedure; they return a `MaximallyIrreducibleCause` and a `MaximallyIrreducibleEffect` object, respectively. The MIC of mechanism BC, for example, is the purview AB (Figure 8 of Oizumi et al., 2014). This is computed like so:

```
>>> mechanism = (B, C)
>>> mic = subsystem.mic(mechanism)
>>> print(mic)
Maximally-irreducible cause
  φ = 1/3
  Mechanism: [B, C]
  Purview = [A, B]
  Direction: CAUSE
  MIP:
     B     C
    ─── × ───
     Ø    A,B
  Repertoire:

     S     Pr(S)
     ------------
     00    2/3
     10    0
     01    0
     11    1/3

  Partitioned repertoire:

     S     Pr(S)
     ------------
     00    1/2
     10    0
     01    0
     11    1/2
```

Concepts

If the mechanism's MIC has $\varphi_{\text{cause}} > 0$ and its MIE has $\varphi_{\text{effect}} > 0$, then the mechanism is said to specify a *concept*. The φ value of the concept as a whole is the minimum of φ_{cause} and φ_{effect}.

We can compute the concept depicted in Figure 9 of (Oizumi et al., 2014) using the `Subsystem.concept()` method, which takes a `mechanism` and returns a `Concept` object containing the φ value, the MIC (in the '`cause`' attribute), and the MIE (in the '`effect`' attribute):

```
>>> mechanism = (A,)
>>> concept = subsystem.concept(mechanism)
>>> print(concept)
═════════════════════════════════════════════════════════
          Concept: Mechanism = [A], φ = 1/6
═════════════════════════════════════════════════════════
            MIC                            MIE
┌───────────────────────────┐┌───────────────────────────┐
│  φ = 1/6                  ││  φ = 1/4                  │
│  Purview = [B, C]         ││  Purview = [B]            │
│  MIP:                     ││  MIP:                     │
│     ∅     A               ││     ∅     A               │
│    ─── × ───              ││    ─── × ───              │
│     B     C               ││     B     ∅               │
│  Repertoire:              ││  Repertoire:              │
│    ┌──────────────┐       ││    ┌─────────────┐        │
│    │ S     Pr(S)  │       ││    │ S    Pr(S)  │        │
│    │ ------------ │       ││    │ ----------- │        │
│    │ 00    0      │       ││    │ 0    1/2    │        │
│    │ 10    1/3    │       ││    │ 1    1/2    │        │
│    │ 01    1/3    │       ││    └─────────────┘        │
│    │ 11    1/3    │       ││                           │
│    └──────────────┘       ││  Partitioned repertoire:  │
│                           ││    ┌─────────────┐        │
│  Partitioned repertoire:  ││    │ S    Pr(S)  │        │
│    ┌──────────────┐       ││    │ ----------- │        │
│    │ S     Pr(S)  │       ││    │ 0    3/4    │        │
│    │ ------------ │       ││    │ 1    1/4    │        │
│    │ 00    1/6    │       ││    └─────────────┘        │
│    │ 10    1/6    │       │└───────────────────────────┘
│    │ 01    1/3    │       │
│    │ 11    1/3    │       │
│    └──────────────┘       │
└───────────────────────────┘
```

Note that in PyPhi, the repertoires are distributions over purview states, rather than system states. Occasionally it is more convenient to represent repertoires as distributions over the entire system. This can be done with the `expand_cause_repertoire()` and `expand_effect_repertoire()` methods of the `Concept` object, which assume the unconstrained (maximum-entropy) distribution over the states of non-purview nodes:

```
>>> full_cr = concept.expand_cause_repertoire()
>>> print(pyphi.distribution.flatten(full_cr))
[ 0.          0.          0.16666667  0.16666667  0.16666667  0.16666667
  0.16666667  0.16666667]
>>> full_er = concept.expand_effect_repertoire()
>>> print(pyphi.distribution.flatten(full_er))
[ 0.0625  0.1875  0.0625  0.1875  0.0625  0.1875  0.0625  0.1875]
```

Also note that `Subsystem.concept()` will return a `Concept` object when $\varphi = 0$ even though these are not concepts, strictly speaking. For convenience, `bool(concept)` evaluates to **`True`** if $\varphi > 0$ and **`False`** otherwise.

Cause-effect structures (system-level information)

The next step is to compute the CES, the set of all concepts specified by the subsystem. The CES characterizes all of the causal constraints that are intrinsic to a physical system. This is implemented by the `pyphi.compute.ces()` function, which simply calls `Subsystem.concept()` for every mechanism $M \in \mathcal{P}(S)$, where $\mathcal{P}(S)$ is the power set of subsystem nodes. It returns a `CauseEffectStructure` object containing those `Concepts` for which $\varphi > 0$.

We see that every mechanism in $\mathcal{P}(S)$ except for AC specifies a concept, as described in Figure 10 of (Oizumi et al., 2014):

```
>>> ces = pyphi.compute.ces(subsystem)
>>> print(ces.labeled_mechanisms)
(['A'], ['B'], ['C'], ['A', 'B'], ['B', 'C'], ['A', 'B', 'C'])
```

Irreducible cause-effect structures (system-level integration)

At this point, the irreducibility of the subsystem's CES is evaluated by applying the integration postulate at the system level. As with integration at the mechanism level, the idea is to measure the difference made by each partition and then take the minimal value as the irreducibility of the subsystem.

We begin by performing a *system cut*. Graphically, the subsystem is partitioned into two parts and the edges going from one part to the other are *cut*, rendering them causally ineffective. This is implemented as an operation on the TPM as follows: Let E_{cut} denote the set of directed edges in the subsystem that are to be cut, where each edge $e \in E_{\mathrm{cut}}$ has a source node a and a target node b. For each edge, we modify the individual TPM of node b (Figure 2.2) by marginalizing over the states of a at t. The resulting TPM specifies the function implemented by b with the causal influence of a removed. We then combine the modified node TPMs to recover the full TPM of the partitioned subsystem. Finally, we recalculate the CES of the subsystem with this modified TPM (the *partitioned CES*).

The irreducibility of a CES with respect to a partition is the distance between the unpartitioned and partitioned CESs (calculated with `pyphi.compute.ces_distance()`; several distances are supported; see § 2.3, Configuration). This distance is evaluated for every partition, and the minimum value across all partitions is the subsystem's integrated information Φ, which measures the extent to which the CES specified by the subsystem is irreducible to the CES under the minimal partition.

This procedure is implemented by the `pyphi.compute.sia()` function, which returns a `SystemIrreducibilityAnalysis` object (Figure 2.1). We can verify that the Φ value of the example system in (Oizumi et al., 2014) is 1.92 and the minimal partition is that which removes the causal connections from AB to C:

```
>>> sia = pyphi.compute.sia(subsystem)
>>> print(sia.phi)
1.916665
>>> print(sia.cut)
Cut [A, B] ━━/ /━━> [C]
```

Complexes (system-level exclusion)

The final step in unfolding the CES of the system is to apply the postulate of exclusion at the system level. We compute the CES of each subset of the network, considered as a subsystem (that is, *fixing* the external nodes as background conditions), and find the CES with maximal Φ, called the *maximally-irreducible cause-effect structure* (MICS) of the system. The subsystem giving rise to it is called the *major complex*; any overlapping subsets with lower Φ are excluded. Non-overlapping subsets may be further analyzed to find additional complexes within the system.

In this example, we find that the whole system ABC is the system's major complex, and all proper subsets are excluded:

```
>>> state = (1, 0, 0)
>>> major_complex = pyphi.compute.major_complex(network, state)
>>> print(major_complex.subsystem)
Subsystem(A, B, C)
```

Note that since `pyphi.compute.major_complex()` is a function of the `Network`, rather than a particular `Subsystem`, it is necessary to specify the state in which the system should be analyzed.

2.3 Design and implementation

PyPhi was designed to be easy to use in interactive, exploratory research settings while nonetheless remaining suitable for use in large-scale simulations or as a component in larger applications. It was also designed to be efficient, given the high computational complexity of the algorithms in IIT. Here we describe some implementation details and optimizations used in the software.

Representation of the TPM and probability distributions

PyPhi supports three different TPM representations: 2-*dimensional state-by-node*, *multidimensional state-by-node*, and *state-by-state*. The state-by-node form is the canonical representation in PyPhi, with the 2-dimensional form used for input and visualization and the multidimensional form used for internal computation. The state-by-state representation is given as an input option for those accustomed to this more general form. If the TPM is given in state-by-state form, PyPhi will raise an error if it does not satisfy (2.1) (conditional independence).

2-dimensional state-by-node form

A TPM in state-by-node form is a matrix where the entry (i, j) gives the probability that the j^{th} node will be ON at $t+1$ if the system is in the i^{th} state at t. This representation has the advantage of being more compact than the state-by-state form, with $2^n \times n$ entries instead of $2^n \times 2^n$, where n is the number of nodes. Note that the TPM admits this representation because in PyPhi the nodes are binary; both $\Pr(N_{t+1} = \text{ON})$ and $\Pr(N_{t+1} = \text{OFF})$ can be specified by a single entry, in our case $\Pr(N_{t+1} = \text{ON})$, since the two probabilities must sum to 1.

Because the possible system states at t are represented implicitly as row indices in 2-dimensional TPMs, there must be an implicit mapping from states to indices. In PyPhi this mapping is achieved by listing the state tuples in lexicographical order and then interpreting them as binary numbers where the state of the first node corresponds to the least-significant bit, so that *e.g.* the state `(0, 0, 0, 1)` is mapped to the row with index 8 (the ninth row, since Python uses zero-based indexing (Dijkstra, 1982)). Designating the first node's state as the least-significant bit is analogous to choosing the little-endian convention in organizing computer memory. This convention is preferable because the mapping is stable under the inclusion of new nodes: including another node in a subsystem only requires concatenating new rows and a new column to its TPM rather than interleaving them. Note that this is opposite convention to that used in writing numbers in positional notation; care must be taken when converting between states and indices and between different TPM representations (the `pyphi.convert` module provides convenience functions for these purposes).

Multidimensional state-by-node form

When a state-by-state TPM is provided to PyPhi by the user, it is converted to state-by-node form and the conditional independence property ((2.1)) is checked. Note that any TPM in state-by-node form necessarily satisfies (2.1). For internal computations, the TPM is then reshaped so that it has $n + 1$ dimensions rather than two: the first n dimensions correspond to the states of each of the n nodes at t, while the last dimension corresponds to the probabilities of each node being ON at $t + 1$. In other words, the indices of the rows (current states) in the 2-dimensional TPM are "unraveled" into n dimensions, with the i^{th} dimension indexed by the i^{th} bit of the 2-dimensional row index according to the little-endian convention. Because the TPM is stored in a NumPy array, this multidimensional form allows us to take advantage of NumPy indexing (Walt et al., 2011) and use a state tuple as an index directly, without converting it to an integer index:

```
>>> state = (0, 1, 1)
>>> print(tpm[state])
[ 1.    0.25  0.75]
```

The first entry of this array signifies that if the state of the system is `(0, 1, 1)` at t, then the probability of the first node N_0 being ON at $t + 1$ is $\Pr(N_{0,\,t+1} = \text{ON}) = 1$. Similarly, the second entry means $\Pr(N_{1,\,t+1} = \text{ON}) = 0.25$ and the third entry means $\Pr(N_{2,\,t+1} = \text{ON}) = 0.75$.

Most importantly, the multidimensional representation simplifies the calculation of marginal and conditional distributions and cause/effect repertoires, because it allows efficient "broadcasting" (Walt et al., 2011) of probabilities when multiplying distributions. Specifically, the Python multiplication operator '$*$' acts as the tensor product when the operands are NumPy arrays `A` and `B` of equal dimensionality such that for each dimension `d`, either `A.shape[d] == 1` or `B.shape[d] == 1`.

Calculation of cause/effect repertoires from the TPM

The cause and effect repertoires of a mechanism over a purview describe how the mechanism nodes in a particular state at t constrain the possible states of the purview nodes at $t-1$ and $t+1$, respectively. Here we describe how they are derived from the TPM in PyPhi.

The effect repertoire

We begin with the simplest case: calculating the effect repertoire of a mechanism $M \subseteq S$ over a purview consisting of a single element $P_i \in S$. This is defined as a conditional probability distribution over states of the purview element at $t+1$ given the current state of the mechanism,

$$\texttt{effect\_repertoire}(M,\, P_i) \;:=\; \Pr(P_{i,t+1} \mid M_t = m_t). \tag{2.2}$$

It is derived from the TPM by conditioning on the state of the mechanism elements, marginalizing over the states of non-purview elements $P' = S \setminus P_i$ (these states correspond to columns in the state-by-state TPM), and marginalizing over the states of non-mechanism elements $M' = S \setminus M$ (these correspond to rows):

$$\Pr(P_{i,t+1} \mid M_t = m_t) \;=\; \frac{1}{|\Omega_{M'}|} \sum_{m'_t \,\in\, \Omega_{M'}} \frac{1}{|\Omega_{P'}|} \sum_{p'_{t+1} \,\in\, \Omega_{P'}} \Pr(P_{i,t+1},\, p'_{t+1} \mid M = m_t,\, M' = m'_t).$$

This operation is implemented in PyPhi by several subroutines. First, in a pre-processing step performed when the `Subsystem` object is created, a `Node` object is created for each element in the subsystem. Each `Node` contains its own individual TPM, extracted from the subsystem's TPM; this is a $2^s \times 2$ matrix where s is the number of the node's parents and the entry (i, j) gives the probability that the node is in state j (`0` or `1`) at $t+1$ given that its parents are in state i at t. This node TPM is represented internally in multidimensional state-by-node form as usual, with singleton dimensions for those subsystem elements that are not parents of the node. The effect repertoire is then calculated by conditioning the purview node's TPM on the state of the mechanism nodes that are also parents of the purview node, via the `pyphi.tpm.condition_tpm()` function, and marginalizing out non-mechanism nodes, with `pyphi.tpm.marginalize_out()`.

In cases where there are mechanism nodes that are not parents of the purview node, the resulting array is multiplied by an array of ones that has the desired shape (dimensions of size two for each mechanism node, and singleton dimensions for each non-mechanism node). Because of NumPy's broadcasting feature, this step is equivalent to taking the tensor product of the array with the maximum-entropy distribution over mechanism nodes that are not parents, so that the final result is a distribution over all mechanism nodes, as desired.

The effect repertoire over a purview of more than one element is given by the tensor product of the effect repertories over each individual purview element,

$$\texttt{effect\_repertoire}(M,\, P) \;:=\; \bigotimes_{P_i \in P} \texttt{effect\_repertoire}(M,\, P_i). \tag{2.3}$$

Again, because PyPhi TPMs and repertoires are represented as tensors (multidimensional arrays), with each dimension corresponding to a node, the NumPy multiplication operator between distributions over different nodes is equivalent to the tensor product. Thus the effect repertoire over an arbitrary purview is trivially implemented by taking the product of the effect repertoires over each purview node with `numpy.multiply()`.

The cause repertoire

The cause repertoire of a single-element mechanism $M_i \in S$ over a purview $P \subseteq S$ is defined as a conditional probability distribution over the states of the purview at $t-1$ given the current state of the mechanism,

$$\texttt{cause\_repertoire}(M_i,\, P) \;:=\; \Pr(P_{t-1} \mid M_{i,t} = m_{i,t}). \tag{2.4}$$

As with the effect repertoire, it is obtained by conditioning and marginalizing the TPM. However, because the TPM gives conditional probabilities of states at $t+1$ given the state at t, Bayes' rule is first applied to express the cause repertoire in terms of a conditional distribution over states at $t-1$ given the state at t,

$$\Pr(P_{t-1} \mid M_{i,t} = m_{i,t}) \;=\; \frac{\Pr(m_{i,t} \mid P_{t-1}) \Pr(P_{t-1})}{\Pr(m_{i,t})}.$$

where the marginal distribution $\Pr(P_{t-1})$ over previous states is the uniform distribution. In this way, the analysis captures how a mechanism in a state constrains a purview without being biased by whether certain states arise more frequently than others in the dynamical evolution of the system (Ay & Polani, 2008; Balduzzi & Tononi, 2008; Hoel et al., 2013; Oizumi et al., 2014). Then the cause repertoire can be calculated by marginalizing over the states of non-mechanism elements $M' = S \setminus M_i$ (now corresponding to columns in the state-by-state TPM) and non-purview elements $P' = S \setminus P$ (now corresponding to rows),

$$\begin{aligned}
&\frac{\Pr(m_{i,t} \mid P_{t-1}) \Pr(P_{t-1})}{\Pr(m_{i,t})} \\
&= \frac{\left(\frac{1}{|\Omega_{P'}|} \sum\limits_{p'_{t-1} \in \Omega_{P'}} \frac{1}{|\Omega_{M'}|} \sum\limits_{m'_t \in \Omega_{M'}} \Pr(m_{i,t}, m'_t \mid P_{t-1}, P'_{t-1} = p'_{t-1}) \right) \Pr(P_{t-1})}{\frac{1}{|\Omega_{M'}|} \sum\limits_{m'_t \in \Omega_{M'}} \Pr(m_{i,t}, m'_t)} \\
&= \frac{\left(\frac{1}{|\Omega_{P'}|} \sum\limits_{p'_{t-1} \in \Omega_{P'}} \sum\limits_{m'_t \in \Omega_{M'}} \Pr(m_t, m'_t \mid P_{t-1}, P'_{t-1} = p'_{t-1}) \right) \Pr(P_{t-1})}{\sum\limits_{m'_t \in \Omega_{M'}} \Pr(m_{i,t}, m'_t)}.
\end{aligned}$$

In PyPhi, the "backward" conditional probabilities $\Pr(m_{i,t} \mid P_{t-1})$ for a single mechanism node are obtained by indexing into the last dimension of the node's TPM with the state $m_{i,t}$ and then marginalizing out non-purview nodes via `pyphi.tpm.marginalize_out()`. As with the effect repertoire, the resulting array is then multiplied by an array of ones with the desired shape in order to obtain a distribution over the entire purview. Finally, because in this case the probabilities were obtained from columns of the TPM, which do not necessarily sum to 1, the distribution is normalized with `pyphi.distribution.normalize()`.

The cause repertoire of a mechanism with multiple elements is the normalized tensor product of the cause repertoires of each individual mechanism element,

$$\texttt{cause\_repertoire}(M, P) \;=\; \frac{1}{K} \bigotimes_{M_i \,\in\, M} \texttt{cause\_repertoire}(M_i, P), \tag{2.5}$$

where

$$K \;=\; \sum_{p_{t-1} \,\in\, \Omega_P} \;\prod_{m_{i,t} \,\in\, \Omega_M} \Pr(P_{t-1} = p_{t-1} \mid M_{i,t} = m_{i,t})$$

is a normalization factor that ensures that the distribution sums to 1. This is implemented in PyPhi via `numpy.multiply()` and `pyphi.distribution.normalize()`. For a more complete illustration of these procedures, see Calculating *Φ*.

Code organization and interface design

The postulates of IIT induce a natural hierarchy of computations Tononi et al., 2016, Supplementary Information S2, and PyPhi's implementation mirrors this hierarchy by using object-oriented programming (Table 2.1) and factoring the computations into compositions of separate functions where possible. One advantage of this approach is that each level of the computation can be performed independently of the higher levels; for example, if one were interested only in the MIE of certain mechanisms rather than the full MICS, then one could simply call `Subsystem.effect_mip()` on those mechanisms instead of calling `pyphi.compute.sia()` and extracting them from the resulting `SystemIrreducibilityAnalysis` object (this is especially important in the case of large systems where the full calculation is infeasible). Separating the calculation into many subroutines and exposing them to the user also has the advantage that they can be easily composed to implement functionality that is not already built-in.

Configuration

Many aspects of PyPhi's behavior may be configured via the `pyphi.config` object. The configuration can be specified in a YAML file (Ben-Kiki et al., 2009); an example is available in the GitHub repository. When PyPhi is imported, it checks the current directory for a file named `pyphi_config.yml` and automatically loads it if it exists. Configuration settings can also be loaded on the fly from an arbitrary file with the `pyphi.config.load_config_file()` function.

Alternatively, `pyphi.config.load_config_dict()` can load configuration settings from a Python

Table 2.1. Correspondence between theoretical objects and PyPhi objects.

Theoretical object	PyPhi object
Discrete dynamical system	`Network`
Candidate system	`Subsystem`
System element	`Node` in `Subsystem.nodes`
System state	Python `tuple` containing a `0` or `1` for each node
Mechanism	Python `tuple` of node indices
Purview	Python `tuple` of node indices
Repertoire over a purview P	NumPy array with $\lvert P \rvert$ dimensions, each of size 2
MIP	The `partition` attribute of the `RepertoireIrreducibilityAnalysis` returned by `Subsystem.cause_mip()` or `Subsystem.effect_mip()`
MIC and MIE	`MaximallyIrreducibleCause` and `MaximallyIrreducibleEffect`
Concept	`Concept`
φ	The `phi` attribute of a `Concept`
CES	`CauseEffectStructure`
Φ	The `phi` attribute of a `CauseEffectStructure`
MICS	The `ces` attribute of the `SystemIrreducibilityAnalysis` returned by `pyphi.compute.major_complex()`
Complex	The `subsystem` attribute of the `SystemIrreducibilityAnalysis` returned by `pyphi.compute.major_complex()`

dictionary. Many settings can also be changed by directly assigning them a new value.

Default settings are used if no configuration is provided. A full description of the available settings and their default values is available in the “Configuration” section of the online documentation.

Optimizations and approximations

Here we describe various optimizations and approximations used by the software to reduce the complexity of the calculations (see § 2.3, Limitations). Memoization and caching optimizations are described in Memoization and caching optimizations.

Connectivity optimizations

As mentioned in § 2.2, Input, providing connectivity information explicitly with a CM can greatly reduce the time complexity of the computations, because in certain cases missing connections imply reducibility *a priori*.

For example, at the system level, if the subsystem is not strongly connected then Φ is necessarily zero. This is because a unidirectional cut between one system part and the rest can always be found that will not actually remove any edges, so the CESs with and without the cut will be identical (see Proof of the strong connectivity optimization for proof). Accordingly, PyPhi immediately excludes these subsystems when finding the major complex of a system.

Similarly, at the mechanism level, PyPhi uses the CM to exclude certain purviews from consideration when computing a MIC or MIE by efficiently determining that repertoires over those purviews are reducible without needing to explicitly compute them. Suppose there are two sets of nodes X and Y for which there exist partitions $X = (X_1,\ X_2)$ and $Y = (Y_1,\ Y_2)$ such that there are no edges from X_1 to Y_2 and no edges from X_2 to Y_1. Then the effect repertoire of mechanism X over purview Y can be factored as

$$\begin{aligned}\texttt{effect\_repertoire}(X, Y) \ &= \\ &\texttt{effect\_repertoire}(X_1, Y_1) \ \otimes \ \texttt{effect\_repertoire}(X_2, Y_2),\end{aligned}$$

and the cause repertoire of mechanism Y over purview X can be factored as

$$\begin{aligned}\texttt{cause\_repertoire}(Y, X) \ &= \\ &\texttt{cause\_repertoire}(Y_1, X_1) \ \otimes \ \texttt{cause\_repertoire}(Y_2, X_2).\end{aligned}$$

Thus in these cases the mechanism is reducible for that purview and $\varphi = 0$ (see Proof of the block-factorable optimization. for proof).

Analytical solution to the earth mover's distance

One of the repertoire distances available in PyPhi is the earth mover's distance (EMD), with the Hamming distance as the ground metric. Computing the EMD between repertoires is a costly operation, with time complexity $O(n2^{3n})$ where n is the number of nodes in the purview (Pele & Werman, 2009). However, when comparing effect repertoires, PyPhi exploits a theorem that states that the EMD between two distributions p and q over multiple nodes is the sum of the EMDs between the marginal distributions over each individual node, if p and q are independent. This analytical solution has time complexity $O(n)$, a significant improvement over the general EMD algorithm (note that this estimate does not include the cost of computing the marginals, which already have been computed to obtain the repertoires). By the conditional independence property (2.1), the conditions of the theorem hold for EMD calculations between effect repertoires, and thus the analytical solution can be used for half of all repertoire calculations performed in the analysis. The theorem is formally stated and proved in Proof of an analytical solution to the EMD between effect repertoires.

Approximations

Currently, two approximate methods of computing Φ are available. These can be used via settings in the PyPhi configuration file (they are disabled by default):

1. `pyphi.config.CUT_ONE_APPROXIMATION` (the "cut one" approximation), and
2. `pyphi.config.ASSUME_CUTS_CANNOT_CREATE_NEW_CONCEPTS` (the "no new concepts" approximation).

In both cases, the complexity of the calculation is greatly reduced by replacing the optimal partitioned CES by an approximate solution. The system's Φ value is then computed as usual as the difference between the unpartitioned CES and the approximate partitioned CES.

Cut one. The "cut one" approximation reduces the scope of the search for the MIP over possible system cuts. Instead of evaluating the partitioned CES for each of the 2^n unidirectional bipartitions of the system, only those $2n$ bipartitions are evaluated that sever the edges from a single node to the rest of the network or vice versa. Since the goal is to find the minimal Φ value across all possible partitions, the "cut one" approximation provides an upper bound on the exact Φ value of the system.

No new concepts. For most choices of mechanism partitioning schemes and distance measures, it is possible that the CES of the partitioned system contains concepts that are reducible in the unpartitioned system and thus not part of the unpartitioned CES. For this reason, PyPhi by default computes the partitioned CES from scratch from the partitioned TPM. Under the "no new concepts" approximation, such new concepts are ignored. Instead of repeating the entire CES computation for each system partition, which requires reevaluating all possible candidate mechanisms for irreducibility, only those mechanisms are taken into account that already specify concepts in the unpartitioned CES. In many types of systems, new concepts due to the partition are rare. Approximations using the "no new concepts" option are thus often accurate. Note, however, that this approximation provides neither a theoretical upper nor lower bound on the exact Φ value of the system.

Limitations

PyPhi's main limitation is that the algorithm is exponential time in the number of nodes, $O(n103^n)$ (see Appendix A.8 for a derivation and comparison to the incorrect calculation in Hanson and Walker, 2023).[1] This is because the number of states, subsystems, mechanisms, purviews, and partitions that must be considered each grows exponentially in the size of the system. This limits the size of systems that can be practically analyzed to ~10–12 nodes. For example, calculating the major complex of systems of three, five, and seven stochastic majority gates, connected in a circular chain of bidirectional edges, takes ~1 s, ~16 s, and ~2.75 h respectively (parallel evaluation of system cuts, 32 × 3.1GHz CPU cores). Using the "cut one" approximation, these calculations take ~1 s, ~12 s, and ~0.63 h. In practice, actual execution times are substantially variable and depend on the specific network under analysis, because network structure determines the effectiveness of the optimizations discussed above.

Another limitation is that the analysis can only be meaningfully applied to a system that is Markovian and satisfies the conditional independence property. These are reasonable assumptions for the intended use case of the software: analyzing a causal TPM derived using the calculus of perturbations (Pearl, 2009). However, there is no guarantee that these assumptions will be valid in other circumstances, such as TPMs derived from observed time series (*e.g.*, EEG recordings). Whether a system has the Markov property and conditional independence property should be carefully checked before applying the software in novel contexts.

2.4 Availability and future directions

PyPhi can be installed with Python's package manager via the command '`pip install pyphi`' on Linux and macOS systems equipped with Python 3.4 or higher. It is open-source and licensed under the GNU General Public License v3.0. The source code is version-controlled with `git` and hosted in a public repository on GitHub at `https://github.com/wmayner/pyphi`. Comprehensive and continually-updated

[1] Note that this expression has been corrected; the $O(n53^n)$ expression published previously in Mayner et al. (2018) resulted from an error in counting the number of mechanisms.

documentation is available online at `https://pyphi.readthedocs.io`. The `pyphi-users` mailing list can be joined at `https://groups.google.com/forum/#!forum/pyphi-users`. A web-based graphical interface to the software is available at `http://integratedinformationtheory.org/calculate.html`.

Several additional features are in development and will be released in future versions. These include a module for calculating Φ over multiple spatial and temporal scales, as theoretically required by the exclusion postulate (in the current version, the `Network` is assumed to represent the system at the spatiotemporal timescale at which Φ is maximized; see Hoel et al., 2016; Marshall et al., 2018), and a module implementing a calculus for "actual causation" as formulated in Albantakis et al. (2019) (preliminary versions of these modules are available in the current release). The software will also be updated to reflect developments in IIT and further optimizations in the algorithm.

CHAPTER 3

Integrated information theory 4.0

Formulating the properties of phenomenal existence in physical terms

Larissa Albantakis,[1☯], Leonardo Barbosa[1,2,☯], Graham Findlay[1,3,☯], Matteo Grasso[1,☯], Andrew M. Haun[1,☯], William Marshall[1,4,☯], William G. P. Mayner[1,3,☯], Alireza Zaeemzadeh[1,☯], Melanie Boly[1,5], Bjørn E. Juel[1,6], Shuntaro Sasai[1,7], Keiko Fujii[1], Isaac David[1], Jeremiah Hendren[1,8], Jonathan P. Lang[1], Giulio Tononi[1]

1 Department of Psychiatry, University of Wisconsin, Madison, WI 53719, USA

2 Fralin Biomedical Research Institute at VTC, Virginia Tech, Roanoke, VA 24016, USA

3 Neuroscience Training Program, University of Wisconsin, Madison, WI 53705, USA

4 Department of Mathematics and Statistics, Brock University, St. Catharines, ON L2S 3A1, Canada

5 Department of Neurology, University of Wisconsin, Madison, WI 53719, USA

6 Institute of Basic Medical Sciences, University of Oslo, Oslo, 0372, Norway

7 Araya Inc., Tokyo, 107-0052, Japan

8 Graduate School Language & Literature, Ludwig Maximilian University of Munich, Munich, 80799, Germany

☯ These authors contributed equally to this work.

Acknowledgments

This project was made possible through the support of a grant from Templeton World Charity Foundation (TWCF0216). In addition, this research was supported by the David P White Chair in Sleep Medicine at the University of Wisconsin-Madison, by the Tiny Blue Dot Foundation (UW 133AAG3451; GT), and by the Natural Science and Engineering Research Council of Canada (NSERC; RGPIN-2019-05418; WM). LA also acknowledges the support of a grant from the Templeton World Charity Foundation (TWCF-2020-20526).

3.1 Introduction

A scientific theory of consciousness should account for experience, which is subjective, in objective terms (Ellia et al., 2021). Being conscious—having an experience—is understood to mean that "there is something it is like to be" (Nagel, 1974): something it is like to see a blue sky, hear the ocean roar, dream of a friend's face, imagine a melody flow, contemplate a choice, or reflect on the experience one is having.

IIT aims to account for phenomenal properties—the properties of experience—in physical terms. IIT's starting point is experience itself rather than its behavioral, functional, or neural correlates (Ellia et al., 2021). Furthermore, in IIT "physical" is meant in a strictly operational sense—in terms of what can be observed and manipulated.

The starting point of IIT is the existence of an experience, which is immediate and irrefutable (Tononi, forthcoming; Tononi, 2015). From this "zeroth" axiom, IIT sets out to identify the essential properties of consciousness—those that are immediate and irrefutably true of every conceivable experience. These are IIT's five axioms of phenomenal existence: every experience is for the experiencer (intrinsicality), specific (information), unitary (integration), definite (exclusion), and structured (composition).

Unlike phenomenal existence, which is immediate and irrefutable (an axiom), physical existence is an explanatory construct (a postulate) and it is assessed operationally (from within consciousness): in physical terms, to be is to have cause-effect power (see Box 2, Principle of being). In other words, something can be said to exist physically if it can "take and make a difference"—bear a cause and produce an effect—as judged by a conscious observer/manipulator.

The next step of IIT is to formulate the essential phenomenal properties (the axioms) in terms of corresponding physical properties (the postulates). This formulation is an "inference to a good explanation" and rests on basic assumptions such as realism, physicalism, and atomism (see Box 1, Methodological guidelines of IIT). If IIT is correct, the substrate[1] of consciousness, beyond having cause-effect power (existence), must satisfy all five essential phenomenal properties in physical terms: its cause-effect power must be for itself (intrinsicality), specific (information), unitary (integration), definite (exclusion), and structured (composition).

On this basis, IIT proposes a fundamental explanatory identity: an experience is identical to the

[1] A substrate should be understood as a set of units that can be observed and manipulated.

cause-effect structure unfolded from a maximal substrate (defined below). Accordingly, all the specific phenomenal properties of any experience must have a good explanation in terms of the specific physical properties of the corresponding cause-effect structure, with no additional ingredients.

Based again on "inferences to a good explanation" (see Box 1), IIT formulates the postulates in a mathematical framework that is in principle applicable to general models of interacting units.[2] A mathematical framework is needed (1) to evaluate whether the theory is self-consistent and compatible with our overall knowledge about the world,, (2) to make specific predictions regarding the quality and quantity of our experiences and their substrate within the brain, and, and (3) to extrapolate from our own consciousness to infer the presence (or absence) and nature of consciousness in beings different from ourselves.

Ultimately, the theory should account for why our consciousness depends on certain portions of the world and their state, such as certain regions of the brain and not others, and for why it fades during dreamless sleep, even though the brain remains active. It should also account for why an experience feels the way it does—why the sky feels extended, why a melody feels flowing in time, and so on. Moreover, the theory makes several predictions concerning both the presence and the quality of experience, some of which have been and are being tested empirically (Tononi et al., 2016).

While the main tenets of the theory have remained the same, its formal framework has been progressively refined and extended (Balduzzi & Tononi, 2008; Oizumi et al., 2014; Tononi, 2004; Tononi & Sporns, 2003). Compared to IIT 1.0 (Tononi, 2004; Tononi & Sporns, 2003), 2.0 (Balduzzi & Tononi, 2008, 2009), and 3.0 (Oizumi et al., 2014), IIT 4.0 presents a more complete, self-consistent formulation and incorporates several recent advances (Albantakis et al., 2019; Barbosa et al., 2021; Haun & Tononi, 2019; Marshall et al., 2023). Chief among them are a more accurate translation of the axioms into postulates and mathematical expressions, the introduction of an Intrinsic Difference (ID) measure (Barbosa et al., 2020, 2021) that is uniquely consistent with IIT's postulates, and the explicit assessment of causal relations (Haun & Tononi, 2019).

In what follows, after introducing IIT's axioms and postulates, we provide its updated mathematical formalism. In the "Results and discussion" section, we apply the mathematical framework of IIT to representative examples and discuss some of their implications. The article is meant as a reference

[2] As mentioned in § Determining maximal unit grains, a substrate unit must be maximally irreducible within, which is likely the case for "real" neurons in the brain, but is not the case for "virtual," simulated neurons in a computer program.

for the theory's mathematical formalism, a concise demonstration of its internal consistency, and an illustration of how a substrate's cause-effect structure is unfolded computationally. A discussion of the theory's motivation, its axioms and postulates, and its assumptions and implications can be found in a forthcoming book (Tononi, forthcoming) and wiki (Intrinsic Ontology Wiki, in preparation) as well as in several publications (Albantakis, 2020a; Albantakis et al., 2023; Ellia et al., 2021; Findlay et al., 2019, in preparation; Grasso et al., 2021; Tononi, Albantakis, et al., 2022; Tononi & Koch, 2015). A survey of the explanatory power and experimental predictions of IIT can be found in Tononi et al. (2016). The way IIT's analysis of cause-effect power can be applied to actual causation, or "what caused what," is presented in Albantakis et al. (2019).

3.2 From phenomenal axioms to physical postulates

Axioms of phenomenal existence

That experience exists—that "there is something it is like to be"—is immediate and irrefutable, as everybody can confirm, say, upon awakening from dreamless sleep. Phenomenal existence is immediate in the sense that my experience is simply there, directly rather than indirectly: I do not need to infer its existence from something else. It is irrefutable because the very doubting that my experience exists is itself an experience that exists—the experience of doubting (Ellia et al., 2021; Tononi & Koch, 2015). Thus, to claim that my experience does not exist is self-contradictory or absurd. The existence of experience is IIT's zeroth axiom.

Existence Experience *exists*: there is *something*.

Traditionally, an axiom is a statement that is assumed to be true, cannot be inferred from any other statement, and can serve as a starting point for inferences. The existence of experience is the ultimate axiom—the starting point for everything, including logic and physics.

On this basis, IIT proceeds by considering whether experience—phenomenal existence—has some axiomatic or essential properties, properties that are immediate and irrefutably true of every conceivable experience. Drawing on introspection and reason, IIT identifies the following five:

Intrinsicality Experience is *intrinsic*: it exists *for itself*.

Information Experience is *specific*: it is *this one*.

Integration Experience is *unitary*: it is *a whole*, irreducible to separate experiences.

Exclusion Experience is *definite*: it is *this whole*.

Composition Experience is *structured*: it is composed of *distinctions* and the *relations* that bind them together, yielding a *phenomenal structure* that feels *the way it feels*.

To exemplify, if I awaken from dreamless sleep, and experience the white wall of my room, my bed, and my body, the experience not only exists, immediately and irrefutably, but (1) it exists for me, not for something else, (2) it is specific (this one experience, not a generic one), (3) it is unitary (the left side is not experienced separately from the right side, and vice versa), (4) it is definite (it includes the visual scene in front of me—neither less, say, its left side only, nor more, say, the wall behind my head), and (5) it is structured by distinctions (the wall, the bed, the body) and relations (the body is on the bed, the bed in the room), which make it feel the way it does and not some other way.

The axioms are not only immediately given, but they are irrefutably true of every conceivable experience. For example, once properly understood, the unity of experience cannot be refuted. Trying to conceive of an experience that were not unitary leads to conceiving of two separate experiences, each of which is unitary, which reaffirms the validity of the axiom. Even though each of the axioms spells out an essential property in its own right, the axioms must be considered together to properly characterize phenomenal existence.

IIT takes the above set of axioms to be complete: there are no further properties of experience that are essential. Other properties that might be considered as candidates for axiomatic status include space (experience typically takes place in some spatial frame), time (an experience usually feels like it flows from a past to a future), change (an experience usually transitions or flows into another), subject-object distinction (an experience seems to involve both a subject and an object), intentionality (experiences usually refer to something in the world, or at least to something other than the subject), a sense of self (many experiences include a reference to one's body or even to one's narrative self), figure-ground segregation (an experience usually includes some object and some background), situatedness (an experience is often bound to a time and a place), will (experience offers the opportunity for action), and affect (experience is often colored by some mood), among others. However, experiences lacking each of these candidate properties are conceivable—that is, conceiving of them does not lead to self-contradiction or absurdity. They are also achievable, as revealed by altered states of consciousness reached through dreaming, meditative practices, or drugs.

Postulates of physical existence

To account for the many regularities of experience (Box 1), it is a good inference to assume the existence of a world that persists independently of one's experience (*realism*). From within consciousness, we can probe the physical existence of things outside of our experience operationally—through observations and manipulations. To be granted physical existence, something should have the power to "take a difference" (be affected) and "make a difference" (produce effects) in a reliable way (*physicalism*). IIT also assumes "operational reductionism", which means that, ideally, to establish what exists in physical terms, one would start from the smallest units that can take and make a difference, so that nothing is left out (*atomism*).

By characterizing physical existence operationally as cause-effect power, IIT can proceed to translate the axioms of phenomenal existence into postulates of physical existence. This establishes the requirements for the *substrate of consciousness*, where "substrate" is meant operationally as a set of units that can be observed and manipulated.

Existence The substrate of consciousness can be characterized operationally by *cause–effect power*: its units must *take and make a difference*.

Building from this "zeroth" postulate, IIT formulates the five axioms in terms of postulates of physical existence that must be satisfied by the substrate of consciousness:

Intrinsicality Its cause-effect power must be *intrinsic*: it must take and make a difference *within itself.*

Information Its cause-effect power must be *specific*: it must be in *this state* and select *this cause-effect state*.

This state is the one with maximal *intrinsic information* (ii), a measure of the difference a system takes or makes over itself for a given cause state and effect state.

Integration Its cause-effect power must be *unitary*: it must specify its cause-effect state as *a whole set* of units, irreducible to separate subsets of units.

Irreducibility is measured by *integrated information* (φ) over the substrate's minimum partition.

Exclusion Its cause-effect power must be *definite*: it must specify its cause-effect state as *this whole set* of units.

This is the set of units that is maximally irreducible, as measured by maximum φ (φ^*). This set is called a *maximal substrate*, also known as *complex* (Marshall et al., 2023; Oizumi et al., 2014).

Composition Its cause-effect power must be *structured*: subsets of its units must specify cause-effect states over subsets of units (*distinctions*) that can overlap with one another (*relations*), yielding a *cause–effect structure* or *Φ-structure* ("Phi-structure") that is *the way it is*.

Distinctions and relations, in turn, must also satisfy the postulates of physical existence: they must have cause-effect power, within the substrate of consciousness, in a specific, unitary, and definite way (they do not have components, being components themselves). They thus have an associated φ value. The *Φ*-structure unfolded from a complex corresponds to the quality of consciousness. The sum total of the φ values of the distinctions and relations that compose the *Φ*-structure measures its *structure integrated information Φ* ("big Phi", "structure Phi") and corresponds to the quantity of consciousness.

According to IIT, the physical properties characterized by the postulates are necessary and sufficient for an entity to be conscious. They are necessary because they are needed to account for the properties of experience that are essential, in the sense that it is inconceivable for an experience to lack any one of them. They are also sufficient because no additional property of experience is essential, in the sense that it is conceivable for an experience to lack that property. Thus, no additional physical property is a necessary requirement for being a substrate of consciousness.

The postulates of IIT have been and are being applied to account for the location of the substrate of consciousness in the brain (Tononi et al., 2016) and for its loss and recovery in physiological and pathological conditions (Casarotto et al., 2016; Massimini et al., 2005).

The explanatory identity between experiences and *Φ*-structures

Having determined the necessary and sufficient conditions for a substrate to support consciousness, IIT proposes an explanatory identity: every property of an experience is accounted for in full by the physical properties of the *Φ*-structure unfolded from a maximal substrate (a complex) in its current state, with no further or “ad hoc” ingredients. That is, there must be a one-to-one correspondence between the way the experience feels and the way distinctions and relations are structured. Importantly, the identity is not meant as a correspondence between the properties of two separate things. Instead, the identity should be understood in an explanatory sense: the intrinsic (subjective) feeling of the experience can be explained extrinsically (objectively, *i.e.*, operationally or physically) in terms of cause-effect power.[3]

The explanatory identity has been applied to account for how space feels (spatial extendedness) and which neural substrates may account for it(Haun & Tononi, 2019). Ongoing work is applying the identity to provide a basic account of the feeling of temporal flow (Comolatti & Grasso, in preparation) and that of objects (Grasso, in preparation).

Box 1. Methodological guidelines of IIT

Inference to a good explanation

We should generally assume that an explanation is good if it can account for a broad set of facts (*scope*), does so in a unified manner (*synthesis*), can explain facts precisely (*specificity*), is internally coherent (*self-consistency*), is coherent with our overall understanding of things (*system consistency*), is simpler than alternatives (*simplicity*), and can make testable predictions (*scientific validation*). For example, IIT 4.0 aims at expressing the postulates of intrinsicality, information, integration, and exclusion in a self-consistent manner when applied to systems, causal distinctions, and relations (see formulas).

Realism

We should assume that something exists (and persists) independently of our own experience. This is a much better hypothesis than solipsism, which explains nothing and predicts nothing. Although IIT starts from our own phenomenology, it aims to account for the many regularities of experience in a way that is fully consistent with realism.

[3] Strictly speaking, distinctions and relations that can be singled out phenomenally, such as a spot, a book, and so on, correspond, in physical terms, to bundles of distinctions and relations (compound distinctions and relations)—that is, to sub-structures of a Φ-structure (Φ-folds) (Ellia et al., 2021; Haun & Tononi, 2019) This can be understood in neural terms because attentional mechanisms can only highlight subsets of units, and thereby all the associated distinctions and relations, rather than individual distinctions and relations. In other words, introspection is the starting point for an explanation of experience in physical terms, but it can only go so far. A full explanation can only be provided through a back-and-forth

Operational physicalism

To assess what exists independently of our own experience, we should employ an operational criterion: we should systematically observe and manipulate a substrate's units and determine that they can indeed take and make a difference in a way that is reliable. Doing so demonstrates a substrate's cause-effect power—the signature of physical existence. Ideally, cause-effect power is fully captured by a substrate's transition probability matrix (TPM) (3.1). This assumption is embedded in IIT's zeroth postulate.

Operational reductionism ("atomism")

Ideally, we should account for what exists physically in terms of the smallest units we can observe and manipulate, as captured by unit TPMs. Doing so would leave nothing unaccounted for. IIT assumes that in principle it should be possible to account for everything purely in terms of cause-effect power—cause-effect power "all the way down" to conditional probabilities between atomic units.[a] Eventually, this would leave neither room nor need to assume intrinsic properties or laws.

[a] Whether these assumptions are ultimately compatible with fundamental physics remains to be determined. However, it is only consistent to assume that the TPM should include all causally relevant aspects of a system and causation may still be discrete even if the system's evolution may be described in terms of continuous fields.

between the properties of a substrate, which can be explored in great detail, and the properties of experience, which can only be characterized crudely through introspection.

Intrinsic perspective

When accounting for experience itself in physical terms, existence should be evaluated from the intrinsic perspective of an entity—what exists for the entity itself, not from the perspective of an external observer. This assumption is embedded in IIT's postulate of intrinsicality and has several consequences. One is that, from the intrinsic perspective, the quality and quantity of existence must be observer-independent and cannot be arbitrary. For instance, information in IIT must be relative to the specific state the entity is in, rather than an average of states as assessed by an external observer. Similarly, it should be evaluated based on the uniform distribution of possible states, as captured by the entity's TPM ((3.1)), rather than on an observed probability distribution. By the same token, units outside the entity should be treated as background conditions that do not contribute directly to what the system is. The intrinsic perspective also imposes a tension between expansion and dilution (see below and Barbosa et al., 2020, 2021): from the intrinsic perspective of a system (or a mechanism within the system), having more units may increase its informativeness (cause-effect power measured as deviation from chance), while at the same time diluting its selectivity (ability to concentrate cause-effect power over a specific state).

3.3 Overview of IIT's framework

IIT 4.0 aims at providing a formal framework to characterize the cause-effect structure of a substrate in a given state by expressing IIT's postulates in mathematical terms. In line with operational physicalism (Box 1), we characterize a substrate by the transition probability function of its constituting units.

On this basis, the IIT formalism first identifies sets of units that fulfill all required properties of a substrate of consciousness according to the postulates of physical existence. First, for a candidate system, we determine a maximal cause-effect state based on the intrinsic information (ii) that the system in its current state specifies over its possible cause states and effect states. We then determine the maximal substrate based on the integrated information (φ_s, "system phi") of the maximal cause-effect state. To qualify as a substrate of consciousness, a candidate system must specify a maximum of integrated information (φ_s^*) compared to all competing candidate systems with overlapping units.

The second part of the IIT formalism *unfolds* the cause-effect structure specified by a maximal substrate in its current state, its *Φ-structure*. To that end, we determine the distinctions and relations specified by the substrate's subsets according to the postulates of physical existence. Distinctions are cause-effect states specified over subsets of substrate units (*purviews*) by subsets of substrate units (*mechanisms*). Relations are congruent overlaps among distinctions' cause and/or effect states. Distinctions and relations are also characterized by their integrated information (φ_d, φ_r). The *Φ*-structure they compose corresponds to the quality of the experience specified by the substrate; the sum of their $\varphi_{d/r}$ values corresponds to its quantity (*Φ*).

While IIT must still be considered as work in progress, having undergone successive refinements, IIT 4.0 is the first formulation of IIT that strives to characterize *Φ*-structures completely and to do so based on measures that satisfy the postulates uniquely. For a comparison of the updated framework with IIT 1.0, 2.0, and 3.0, see Appendix B.2.

Substrates, transition probabilities, and cause-effect power

IIT takes physical existence as synonymous with having cause-effect power, the ability to take and make a difference. Consequently, a substrate U with state space Ω_U is operationally defined by its potential interactions, assessed in terms of conditional probabilities (physicalism, Box 1). We denote the complete transition probability function of a substrate U over a system update $u \to \bar{u}$ as

$$\mathcal{T}_U \equiv p(\bar{u} \mid u), \quad u, \bar{u} \in \Omega_U. \tag{3.1}$$

A substrate in IIT can be described as a stochastic system $U = \{U_1, U_2, \ldots, U_n\}$ of n interacting units with state space $\Omega_U = \prod_i \Omega_{U_i}$ and current state $u \in \Omega_U$. We define units in state u as a set of tuples, where each tuple contains the unit and the state of the unit, *i.e.*, $u = \{(U_i, \text{state}(U_i)) : U_i \in U\}$. This allows us to define set operations over u that consider both the units and their states. Ω_U is the set of all possible such tuple sets, corresponding to all the possible states of U. We assume that the system updates in discrete steps, that the state space Ω_U is finite, and that the individual random variables $U_i \in U$ are conditionally independent from each other given the preceding state of U:

$$p(\bar{u} \mid u) = \prod_{i=1}^{n} p(\bar{u}_i \mid u). \tag{3.2}$$

Finally, we assume a complete description of the substrate, which means that we can determine the conditional probabilities in (3.2) for every system state,

$$\exists\, p(\bar{u} \mid u) \quad \forall\, u, \bar{u} \in \Omega_U, \tag{3.3}$$

with $p(\bar{u} \mid u) = p(\bar{u} \mid \mathrm{do}(u))$ (Albantakis et al., 2019; Ay & Polani, 2008; Janzing et al., 2013; Pearl, 2000), where the "do-operator" $\mathrm{do}(u)$ indicates that u is imposed by intervention. This implies that U must correspond to a causal network (Albantakis et al., 2019), and $\mathcal{T}_U$ is a transition probability matrix (TPM) of size $|\Omega_U|$.[4]

The TPM $\mathcal{T}_U$, which forms the starting point of IIT's analysis, serves as an overall description of a system's causal evolution under all possible interventions: what is the probability that the system will transition into each of its possible states upon being initialized into every possible state (Figure 3.1)? (Notably, there is no additional role for intrinsic physical properties or laws of nature.) In practice, a causal model will be neither complete nor atomic (capturing the smallest units that can be observed and manipulated), but will capture the relevant features of what we are trying to explain and predict.[5]

In the "Results and discussion" section, the IIT formalism will be applied to extremely simple, simulated networks, rather than causal models of actual substrates. The cause-effect structures derived from these simple networks only serve as convenient illustrations of how a hypothetical substrate's cause-effect power can be unfolded.

Implementing the postulates

In what follows, our goal is to evaluate whether a hypothetical substrate (also called "system") satisfies all the postulates of IIT. To that end, we must verify whether the system has cause-effect power that is intrinsic, specific, integrated, definite, and structured.

[4] While the IIT formalism can be applied to hypothetical or simulated systems (as we do for the example systems in the "Results and discussion" section), for the resulting quantities to capture existence in physical terms they must be applied to substrate units that can actually be observed and manipulated in physical terms.

[5] As demonstrated in Albantakis et al. (2023), it is possible to extend IIT's causal framework to finite quantum systems under unitary evolution, where the conditional independence assumption (3.2) applies to non-entangled subsystems.

Existence

According to IIT, existence understood as cause-effect power requires the capacity to both take *and* make a difference (see Box 2, Principle of being). On the basis of a complete description of the system in terms of interventional conditional probabilities ($\mathcal{T}_U$) (3.1), cause-effect power can be quantified as causal *informativeness*. Cause informativeness measures how much a potential cause increases the probability of the current state, and effect informativeness how much the current state increases the probability of a potential effect (as compared to chance).

Intrinsicality

Building upon the existence postulate, the intrinsicality postulate further requires that a system exerts cause-effect power *within itself*. In general, the systems we want to evaluate are open systems $S \subseteq U$ that are part of a larger "universe" U. From the intrinsic perspective of a system S (see Box 1), the set of the remaining units $W = U \setminus S$ merely act as background conditions that do not contribute directly to cause-effect power. To enforce this, we causally marginalize the background units, conditional on the current state of the universe, rendering them causally inert (see Identifying substrates of consciousness for details).

Information

The information postulate requires that a system's cause-effect power be specific: the system in its current state must select a specific cause-effect state for its units. Based on the *principle of maximal existence* (Box 2), this is the state for which intrinsic information is maximal—the *maximal cause-effect state*. *Intrinsic information* (ii) measures the difference a system takes or makes over itself for a given cause and effect state as the product of informativeness and selectivity. As we have seen (existence), *informativeness* quantifies the causal power of a system in its current state as a reduction of uncertainty with respect to chance. *Selectivity* measures how much cause-effect power is concentrated over that specific cause or effect state. Selectivity is reduced by uncertainty in the cause or effect state with respect to other potential cause and effect states.

From the intrinsic perspective of the system, the product of informativeness and selectivity leads to a tension between *expansion* and *dilution*, whereby a system comprising more units may show increased deviation from chance but decreased concentration of cause-effect power over a specific state (Barbosa et al., 2020, 2021).

Integration

By the integration postulate, it is not sufficient for a system to have cause-effect power within itself and select a specific cause-effect state: it must also specify its maximal cause-effect state in a way that is irreducible. This can be assessed by *partitioning* the set of units that constitute the system into separate parts. The system integrated information (φ_s) then quantifies how much the intrinsic information specified by the maximal state is reduced due to the partition.[6] Integrated information is evaluated over the partition that makes the least difference, the *minimum partition* (MIP), in accordance with the *principle of minimal existence* (see Box 2).

Integrated information is highly sensitive to the presence of *fault lines*—partitions that separate parts of a system that interact weakly or directionally (Marshall et al., 2023).

Exclusion

Many overlapping sets of units may have a positive value of integrated information (φ_s). However, the exclusion postulate requires that the substrate of consciousness must be constituted of a definite set of units, neither less nor more. Moreover, units, updates, and states must have a definite grain. Operationally, the exclusion postulate is enforced by selecting the set of units that maximizes integrated information over itself (φ_s^*), based again on the principle of maximal existence (see Box 2). That set of units is called a *maximal substrate*, or *complex*. Over a universal substrate, sets of units for which integrated information is maximal compared to all competing candidate systems with overlapping units can be assessed recursively (by identifying the first complex, then the second complex, and so on).

Composition

Once a complex has been identified, composition requires that we characterize its *cause-effect structure* by considering all its subsets and fully *unfolding* its cause-effect power.

Usually, causal models are conceived in holistic terms, as state transitions of the system as a whole

[6] Note that this notion of irreducibility based on set-partitions differs from typical information-theoretic notions such as redundancy or synergy (Albantakis & Tononi, 2019; Beer & Williams, 2015; Mediano et al., 2019).

(3.1), or in reductionist terms, as a description of the individual units of the system and their interactions (3.2) (Albantakis & Tononi, 2019). However, to account for the structure of experience, considering only the cause-effect power of the individual units or of the system as a whole would be insufficient (Albantakis & Tononi, 2019; Grasso et al., 2021). Instead, by the composition postulate, we have to evaluate the system's cause-effect structure by considering the cause-effect power of its subsets as well as their causal relations.

To contribute to the cause-effect structure of a complex, a system subset must both take *and* make a difference (as required by existence) *within* the system (as required by intrinsicality). A subset $M \subseteq S$ in state $m \in \Omega_M$ is called a *mechanism* if it *links* a cause and effect state over subsets of units $Z_{c/e} \subseteq S$, called *purviews*. A mechanism together with the cause and effect state it specifies is called a *causal distinction*. Distinctions are evaluated based on whether they satisfy all the postulates of IIT (except for composition). For every mechanism, the cause-effect state is the one having maximal intrinsic information (ii), and the cause and effect purviews are those yielding the maximum value of integrated information (φ_d) within the complex—that is, those that are maximally irreducible. By the information postulate, the cause-effect power of a complex must be specific, which means that it selects a specific cause-effect state at the system level. Consequently, the distinctions that exist for the complex are only those whose cause-effect state is congruent with the cause-effect state of the complex as a whole (incongruent distinctions are not components of the complex and its specific cause-effect power because they would violate the specificity postulate, according to which the experience can only be "this one").

Distinctions whose cause or effect states overlap congruently within the system (over the same subset of units in the same state) are *bound* together by *causal relations*. Relations also have an associated value of integrated information (φ_r), corresponding to their irreducibility.

Together, these distinctions and relations compose the *cause-effect structure* of the complex in its current state. The cause-effect structure specified by a complex is called a *Φ-structure*. The sum of its distinction and relation integrated information amounts to the structure integrated information (Φ) of the complex.

In the following, we will provide a formal account of the IIT analysis. The first part demonstrates how to identify complexes. This requires that we (1) determine the cause-effect state of a system in its current state,, (2) evaluate the system integrated information (φ_s) over that cause-effect state, and, and (3) search iteratively for maxima of integrated information (φ_s^*) within a universe. The second part describes how the postulates of IIT are applied to unfold the cause-effect structure of a complex. This requires that we identify the causal distinctions specified by subsets of units within the complex and the causal relations determined by the way distinctions overlap, yielding the system's Φ-structure and its structure integrated information Φ.

Box 2. Ontological principles of IIT

Principle of being

The *principle of being* states that *to be is to have cause-effect power*. In other words, in physical, operational terms, to exist requires being able to take and make a difference. The principle is closely related to the so-called Eleatic principle, as found in Plato's Sophist dialogue (Cooper, 1997): "I say that everything possessing any kind of power, either to do anything to something else, or to be affected to the smallest extent by the slightest cause, even on a single occasion, has real existence: for I claim that entities are nothing else but power." A similar principle can be found in the work of the Buddhist philosopher Dharmakīrti: "Whatever has causal powers, that really exists." (Tillemans, 2021). Note that the Eleatic principle is enunciated as a disjunction (either to do something... *or* to be affected...), whereas IIT's principle of being is presented as a conjunction (take *and* make a difference).

Principle of maximal existence

The *principle of maximal existence* states that, when it comes to a requirement for existence, *what exists is what exists the most*. The principle is offered by IIT as a good explanation for why the system state specified by the complex and the cause-effect states specified by its mechanisms are what they are. It also provides a criterion for determining the set of units constituting a complex—the one with maximally irreducible cause-effect power, for determining the subsets of units constituting the distinctions and relations that compose its cause-effect structure, and for determining the units' grain. To exemplify, consider a set of candidate complexes overlapping over the same substrate. By the postulates of integration and exclusion, a complex must be both unitary and definite. By the maximal existence principle, the complex should be the one that lays the greatest claim to existence as *one* entity, as measured by system integrated information (φ_s). For the same reason, candidate complexes that overlap over the same substrate but have a lower value of φ_s are excluded from existence. In other words, if having maximal φ_s is the reason for assigning existence as a unitary complex to a set of units, it is also the reason to exclude from existence any overlapping set not having maximal φ_s.

Principle of minimal existence

Another key principle of IIT is the *principle of minimal existence*, which complements that of maximal existence. The principle states that, when it comes to a requirement for existence, *nothing exists more than the least it exists*. The principle is offered by IIT as a good explanation for why, given that a system can only exist as one system if it is irreducible, its degree of irreducibility should be assessed over the partition across which it is least irreducible (the minimum partition). Similarly, a distinction within a system can only exist as one distinction to the extent that it is irreducible, and its degree of irreducibility should be assessed over the partition across which it is least irreducible. Moreover, a set of units can only exist as a system, or as a distinction within the system, if it specifies both an irreducible cause and an irreducible effect, so its degree of irreducibility should be the minimum between the irreducibility on the cause side and on the effect side.[a]

[a] A principle of IIT not discussed here is the Principle of becoming, which states that powers become what powers do. That is, conditional probabilities in the TPM update depending on what happens. The principle and some of its implications—examined in Tononi (forthcoming) and other forthcoming work.

3.4 Identifying substrates of consciousness

Our starting point is a substrate U in current state u with TPM $\mathcal{T}_U$ (3.1). We consider any subset $s \subseteq u$ as a possible complex and refer to a set of units $S \subseteq U$ as a candidate system. (Note that s and u are sets of tuples containing both the units and their states.)

By the intrinsicality postulate, the units $W = U \setminus S$ are background conditions, and do not contribute directly to the cause-effect power of the system. To discount the contribution of background units, they are *causally marginalized*, conditional on the current state of the universe. This means that the background units are marginalized based on a uniform marginal distribution, updated by conditioning on u. The process is repeated separately for each unit in the system, and they are then combined using a product (in line with conditional independence), which eliminates any residual correlations due to the background units. Accordingly, we obtain two TPMs $\mathcal{T}_e$ and $\mathcal{T}_c$ (for evaluating effects and causes, respectively) for the candidate system S. For evaluating effects, the state of the background units is fully determined by the current state of the universe. The corresponding TPM, $\mathcal{T}_e$, is used to identify the effect of the current state:

$$\mathcal{T}_e = \mathcal{T}_e(\mathcal{T}_U, u, w) \equiv p_e(\bar{s} \mid s) = p(\bar{s} \mid s, w), \quad s, \bar{s} \in \Omega_S, \tag{3.4}$$

where $w = u \setminus s$. For evaluating causes, knowledge of the current state is used to compute the probability distribution over potential prior states of the background units, which is not necessarily uniform or deterministic. The corresponding TPM, $\mathcal{T}_c$, is used to evaluate the cause of the current state:

$$\mathcal{T}_c = \mathcal{T}_c(\mathcal{T}_U, u, w) \equiv p_c(s \mid \bar{s}) = \prod_{i=1}^{|S|} \sum_{\bar{w}} p(s_i \mid \bar{s}, \bar{w}) \left(\frac{\sum_{\hat{s}} p(u \mid \hat{s}, \bar{w})}{\sum_{\hat{u}} p(u \mid \hat{u})} \right), \quad s, \bar{s} \in \Omega_S. \tag{3.5}$$

In both TPMs, the background units W are rendered causally inert, so that causes and effects are evaluated from the intrinsic perspective of the system.

The intrinsic information $\text{ii}_{c/e}$ is a measure of the intrinsic cause or effect power exerted by a system S in its current state s over itself by selecting a specific cause or effect state $\bar{s}$. The cause-effect state for which intrinsic information (ii_c and ii_e) is maximal is called the maximal cause-effect state $s' = \{s'_c, s'_e\}$. The integrated information φ_s is a measure of the irreducibility of a cause-effect state, compared to the directional system partition θ' that affects the maximal cause-effect state the least (minimum partition, or MIP). Systems for which integrated information is maximal (φ_s^*) compared to any competing candidate system with overlapping units are called maximal substrates, or complexes.

The IIT 4.0 formalism to measure a system's integrated information φ_s and to identify maximal substrates was first presented in Marshall et al. (2023). An example of how to identify complexes in a simple system is given in Figure 3.1, while a comparison with prior accounts (IIT 1.0, IIT 2.0, and IIT 3.0) can be found in Appendix B.2. An outline of the IIT algorithm is included in Appendix B.4.

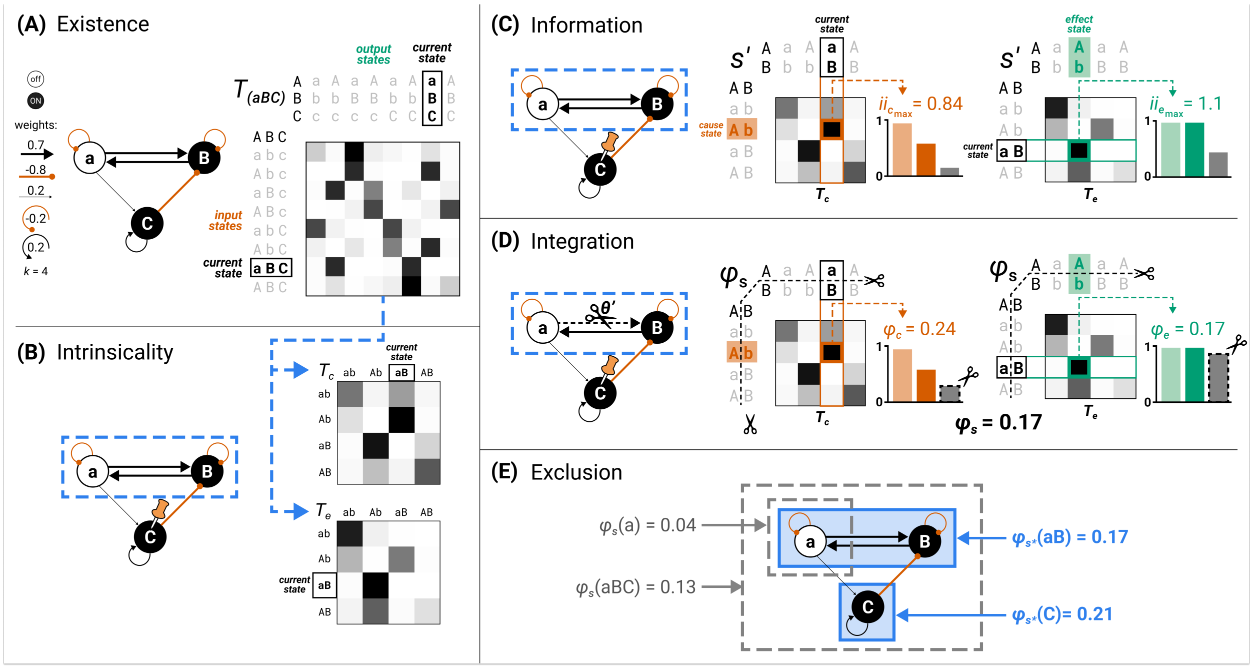


Figure 3.1. Identifying substrates of consciousness through the postulates of existence, intrinsicality, information, integration, and exclusion. **(A)** The substrate $S = aBC$ in state $(-1, 1, 1)$ (lowercase letters for units indicated state "-1", uppercase letters state "$+1$") is the starting point for applying the postulates. The substrate updates its state according to the depicted transition probability matrix (TPM) (gray shading indicates probability value from white ($p = 0$) to black ($p = 1$); each unit follows a logistic equation (see § 3.6 for definition) with $k = 4.0$ and connection weights as indicated in the causal model). Existence requires that the substrate must have cause-effect power, meaning that the TPM among substrate states must differ from chance. **(B)** Intrinsicality requires that a candidate substrate, for example, units aB, has cause-effect power over itself. Units outside the candidate substrate (in this case, unit C) are treated as background conditions. **(C)** Information requires that the candidate substrate aB selects a specific cause-effect state (s'). This is the cause state (red) and effect state (green) for which intrinsic information (ii) is maximal. Bar plots on the right indicate the three probability terms relevant for computing ii_c (3.8) and ii_e (3.6): the selectivity (light colored bar), as well as the constrained (dark colored bar) and unconstrained (gray bar) effect probabilities in the informativeness term. **(D)** Integration requires that the substrate specifies its cause-effect state irreducibly ("as one"). This is established by identifying the minimum partition (MIP; θ') and measuring the integrated information of the system (φ_s)—the minimum between cause integrated information (φ_c) and effect integrated information (φ_e). Here, gray bars represent the partitioned probability required for computing φ_c (3.21) and φ_e (3.20). **(E)** Exclusion requires that the substrate of consciousness is definite, including some units and excluding others. This is established by identifying the candidate substrate with the maximum value of system integrated information (φ_s^*)—the maximal substrate, or complex. In this case, aB is a complex since its system integrated information ($\varphi_s = 0.17$) is higher than the one of all other overlapping systems (for example, subset a with $\varphi_s = 0.04$ and superset aBC with $\varphi_s = 0.13$).

Existence, intrinsicality, and information: Determining the maximal cause-effect state of a candidate system

Given a causal model with corresponding TPMs $\mathcal{T}_e$ (3.4) and $\mathcal{T}_c$ (3.5), we wish to identify the maximal cause-effect state specified by a system in its current state over itself and to quantify the causal power with which it does so. In this way, we quantify the cause-effect power of a system from its intrinsic perspective, rather than from the perspective of an outside observer (see Box 1).

System intrinsic information ii

Intrinsic information $\text{ii}(s, \bar{s})$ measures the causal power of a system S over itself, for its current state s, over a specific cause or effect state $\bar{s}$. Intrinsic information depends on interventional conditional probabilities and unconstrained probabilities of cause or effect states and is the product of selectivity and informativeness.

On the effect side, intrinsic effect information ii_e of the current state s over a possible effect state $\bar{s}$ is defined as:

$$\underset{e}{\text{ii}}(s, \bar{s}) = p_e(\bar{s} \mid s) \log \left(\frac{p_e(\bar{s} \mid s)}{p_e(\bar{s})} \right), \tag{3.6}$$

where $p_e(\bar{s} \mid s)$ (3.4) is the interventional conditional probability that the current state s produces the effect state $\bar{s}$, as indicated by $\mathcal{T}_e$.

The interventional unconstrained probability $p_e(\bar{s})$

$$p_e(\bar{s}) = |\Omega_S|^{-1} \sum_{s \in \Omega_S} p_e(\bar{s} \mid s), \tag{3.7}$$

is defined as the marginal probability of $\bar{s}$, averaged across all possible current states of S with equal probability (where $|\Omega_S|$ denotes the cardinality of the state space Ω_S).

On the cause side, intrinsic cause information ii_c of the current state s over a possible cause state $\bar{s}$ is defined as:

$$\underset{c}{\text{ii}}(s, \bar{s}) = p_c^{\leftarrow}(\bar{s} \mid s) \log \left(\frac{p_c(s \mid \bar{s})}{p_c(s)} \right), \tag{3.8}$$

where $p_c(s, \bar{s})$ (3.5) is the interventional conditional probability that the cause state $\bar{s}$ produces the current state s, as indicated by $\mathcal{T}_c$, and the interventional unconstrained probability is again defined as the marginal probability of s, averaged across all possible cause states of S with equal probability,

$$p_c(s) = |\Omega_S|^{-1} \sum_{\bar{s} \in \Omega_S} p_c(s \mid \bar{s}). \tag{3.9}$$

Moreover, $p_c^{\leftarrow}(\bar{s} \mid s)$ (3.5) is the interventional conditional probability that the current state $s \in \Omega_S$ was produced by $\bar{s}$; it is derived from $\mathcal{T}_c$ using Bayes' rule, where we again assign a uniform prior to the possible cause states $\bar{s}$,

$$p_c^{\leftarrow}(\bar{s} \mid s) = \frac{p_c(s \mid \bar{s}) \cdot |\Omega_S|^{-1}}{p_c(s)} = \frac{p_c(s \mid \bar{s})}{\sum\limits_{\hat{s} \in \Omega_S} p_c(s \mid \hat{s})}. \tag{3.10}$$

Informativeness (over chance)

In (3.6) and (3.8), the logarithmic term (in base 2 throughout) is called *informativeness*. Note that informativeness is expressed in terms of 'forward' probabilities (probability of a subsequent state given the current state) for both ii_e (3.6) and ii_c (3.8). However, ii_e (3.6) evaluates the increase in probability of the effect state due to the current state based on $\mathcal{T}_e$, while ii_c (3.8) evaluates the increase in probability of the current state due to the cause state based on $\mathcal{T}_c$.

In line with the existence postulate, a system S in state s has cause-effect power (it takes and makes a difference) if it raises the probability of a possible effect state compared to chance, which is to say compared to its unconstrained probability,

$$\log\left(\frac{p_e(\bar{s} \mid s)}{p_e(\bar{s})}\right) > 0, \tag{3.11}$$

and if the probability of the current state is raised above chance by a possible cause state,

$$\log\left(\frac{p_c(s \mid \bar{s})}{p_c(s)}\right) > 0. \tag{3.12}$$

Informativeness is additive over the number of units: if a system specifies a cause or effect state with probability $p = 1$, its causal power increases additively with the number of units whose states it fully specifies (*expansion*), given that the chance probability of all states decreases exponentially.

Selectivity (over states)

From the intrinsic perspective of a system, cause-effect power over a specific cause or effect state depends not only on the deviation from chance it produces, but also on how its probability is concentrated on that state, rather than being diluted over other states. This is measured by the *selectivity* term in front of the logarithmic term in (3.6) and (3.8), corresponding to the conditional probability $p_c^{\leftarrow}(\bar{s} \mid s)$ or $p_e(\bar{s} \mid s)$ of that specific cause or effect state. (Note that here, on the cause side, we use the 'backward' probability (probability of a prior state given the current state) obtained through Bayes' rule, while we use the 'forward' probability of the effect state $\bar{s}$ given s on the effect side.) Selectivity means that if $p < 1$, the system's causal power becomes subadditive (*dilution*) (see Barbosa et al. (2020) for details). For example, as shown in Barbosa et al. (2021), if an unconstrained unit is added to a fully specified unit, intrinsic information does not just stay the same, but decreases exponentially. From the intrinsic perspective of the system, the informativeness of a specific cause or effect state is diluted because it is spread over multiple possible states, yet the system must select only one state.

Altogether, taking the product of informativeness and selectivity leads to a tension between expansion and dilution: a larger system will tend to have higher informativeness than a smaller system because it will deviate more from chance, but it will also tend to have lower selectivity because it will have a larger repertoire of states to select from.

Because of the selectivity term, intrinsic information is reduced by indeterminism and degeneracy. As shown in Marshall et al. (2023), indeterminism decreases the probability of the selected effect state because it implies that the same state can lead to multiple states. In turn, degeneracy decreases the probability of the selected cause state because it implies that multiple states can lead to the same state, even in a deterministic system.

The intrinsic information ii is quantified in units of *intrinsic bits*, or *ibits*, to distinguish it from standard information-theoretic measures (which are typically additive). Formally, the *ibit* corresponds to a point-wise information value (measured in bits) weighted by a probability.

The maximal cause-effect state

Taking the product of informativeness and selectivity on the system's cause and effect sides captures the postulates of existence (taking and making a difference) and intrinsicality (taking and making a difference over itself) for each possible cause or effect state, as measured by intrinsic information. However, the information postulate further requires that the system selects a specific cause or effect state. The selection is determined by the principle of maximal existence (Box 1): the cause or effect specified by the system should be the one that maximizes intrinsic information. On the effect side (and similarly for the cause side, see Appendix B.4, IIT Algorithm),

$$\begin{aligned} s'_e(\mathcal{T}_e, s) &= \operatorname*{argmax}_{\bar{s}\in\Omega_S} \underset{e}{\text{ii}}(s, \bar{s}) \\ &= \operatorname*{argmax}_{\bar{s}\in\Omega_S} p_e(\bar{s} \mid s) \log\left(\frac{p_e(\bar{s} \mid s)}{p_e(\bar{s})}\right). \end{aligned} \tag{3.13}$$

The system's intrinsic effect information is the value of ii_e (3.6) for its maximal effect state:

$$\underset{e}{\text{ii}}(\mathcal{T}_e, s) := \underset{e}{\text{ii}}(s, s'_e) = \max_{\bar{s}\in\Omega_S} p_e(\bar{s} \mid s) \log\left(\frac{p_e(\bar{s} \mid s)}{p_e(\bar{s})}\right). \tag{3.14}$$

We have made the dependency of s' and ii_e on $\mathcal{T}_e$ explicit in (3.13) and (3.14) to highlight that, for intrinsic information to properly assess cause-effect power, all probabilities must be derived from the system's interventional transition probability function, while imposing a uniform prior distribution over all possible system states. If $\text{ii}_e(\mathcal{T}_e, s) = 0$, the system S in state s has no causal power. This is the case if and only if $p_e(\bar{s} \mid s) = p_e(\bar{s})$ for every $\bar{s}$ (Barbosa et al., 2020; and likewise, it can be shown that $\text{ii}_c(\mathcal{T}_c, s) = 0$ if and only if $p_c(s \mid \bar{s}) = p_c(s)$ for every $\bar{s}$.) It is worthwhile to mention that when $\text{ii}_e(\mathcal{T}_e, s) \neq 0$, the system state s always increases the probability of the intrinsic effect state compared to chance. Similarly, when $\text{ii}_c(\mathcal{T}_c, s) \neq 0$ the intrinsic cause state increases the probability of the system state, satisfying (3.12). Note also that a system's intrinsic cause-effect state does not necessarily correspond to the actual cause and effect states (what actually happened before / will happen after) in the dynamical evolution of the system, which typically also depends on extrinsic influences. (For an account of actual causation according to the causal principles of IIT, see Albantakis et al., 2019.)

Intrinsic Difference

Because consciousness is the way it is, the translation of its properties in physical, operational terms should be unique and based on quantities that uniquely satisfy the postulates (Barbosa et al., 2021; A. B. Barrett & Mediano, 2019). Intrinsic information is formulated as a product of selectivity and informativeness based on the notion of intrinsic difference (ID)(Barbosa et al., 2020). This is a measure of the difference between two probability distributions which uniquely satisfies three properties (causality, specificity, and intrinsicality) that align with the postulates of IIT (but also have independent justification):

Causality (Existence): the measure is zero if and only if the system does not make a difference

Intrinsicality (Intrinsicality): the measure increases if the system is expanded without noise (expansion) and decreases if the system is expanded without signal (dilution)

Specificity (Information): the measure reflects the cause-effect power of a specific state over a specific cause and effect state.

The properties uniquely satisfied by the ID are described in a general mathematical context in Barbosa et al. (2020), as well as some additional discussion in Appendix B.2.

Note that, on the effect side, ii_e is formally equivalent to the ID between the constrained effect repertoire $p_e(\bar{s} \mid s)$ and the unconstrained effect repertoire $p_e(\bar{s})$. On the cause side, the application of Bayes rule to compute $p_c^{\leftarrow}(\bar{s} \mid s)$ as the selectivity term means that ii_c is not strictly equivalent to the ID between two probability distributions. However, analogously to the effect formulation, it is defined as the product of selectivity and informativeness of causes.

Integration: Determining the irreducibility of a candidate system

Having identified the maximal cause-effect state $s' = \{s'_c, s'_e\}$ of a candidate system S in its current state s, the next step is to evaluate whether the system specifies the cause-effect state of its units in a way that is *irreducible*, as required by the integration postulate: a candidate system can only be a substrate of consciousness if it is *one* system—that is, if it cannot be subdivided into subsets of units that exist separately from one another.

Directional system partitions

To that end, we define a set of *directional* system partitions $\Theta(S)$ that divide S into $k \geq 2$ parts $\{S^{(i)}\}_{i=1}^{k}$, such that

$$S^{(i)} \neq \varnothing,\ S^{(i)} \cap S^{(j)} = \varnothing, \text{ and } \bigcup_{i=1}^{k} S^{(i)} = S. \tag{3.15}$$

In words, each part $S^{(i)}$ must contain at least one unit, there must be no overlap between any two parts $S^{(i)}$ and $S^{(j)}$, and every unit of the system must appear in exactly one part. For each part $S^{(i)}$, the partition removes the causal connections of that part with the rest of the system in a directional manner: either the part's inputs, outputs, or both are replaced by independent "noise" (they are "cut" by the partition in the sense that their causal powers are substituted by chance). Directional partitions are necessary because, from the intrinsic perspective of a system, a subset of units that cannot affect the rest of the system, or cannot be affected by it, cannot truly be a part of the system. In other words, to be a part of a system, a subset of units must be able to interact with the rest of the system in both directions (cause *and* effect).

A partition $\theta \in \Theta(S)$ thus has the form

$$\theta = \{S^{(1)}_{\delta_1}, S^{(2)}_{\delta_2}, \ldots, S^{(k)}_{\delta_k}\}, \tag{3.16}$$

where $\delta_i \in \{\leftarrow, \rightarrow, \leftrightarrow\}$ indicates whether the inputs ($\leftarrow$), outputs ($\rightarrow$), or both ($\leftrightarrow$) are cut for a given part. For each part $S^{(i)}$, we can then identify a set of units $X^{(i)} \subseteq S$ whose inputs to $S^{(i)}$ have been cut by the partition, and the complementary set $Y^{(i)} = S \setminus X^{(i)}$ whose inputs to $S^{(i)}$ are left intact. Specifically,

$$X^{(i)} = \begin{cases} S \setminus S^{(i)} & \text{if } \delta_i \in \{\leftarrow, \leftrightarrow\} \\ \bigcup\limits_{\substack{j \neq i: \\ \delta_j \in \{\rightarrow, \leftrightarrow\}}} S^{(j)} & \text{if } \delta_i \in \{\rightarrow\}. \end{cases} \tag{3.17}$$

In the first case, if $\delta_i \in \{\leftarrow, \leftrightarrow\}$, all inputs to $S^{(i)}$ from $S \setminus S^{(i)}$ are cut. In the second case, if $\delta_i \in \{\rightarrow\}$, there may still be inputs to $S^{(i)}$ that are cut, which correspond to the outputs of all $S^{(j)}$ with $\delta_j \in \{\rightarrow, \leftrightarrow\}$.

Given a partition $\theta \in \Theta(S)$, we define partitioned transition probability matrices $\mathcal{T}^{\theta}_{e}$ and $\mathcal{T}^{\theta}_{c}$ in which all connections affected by the partition are "noised." This is done by combining the independent contributions of each unit $S_j \in S$ in line with the conditional independence assumption (3.2). For the effect TPM (and analogously for the cause TPM)

$$\mathcal{T}^{\theta}_{e} \equiv p^{\theta}_{e}(\bar{s} \mid s) = \prod_{j=1}^{n} p^{\theta}_{e}(\bar{s}_j \mid s),\ \bar{s}, s \in \Omega_S, \tag{3.18}$$

where the partitioned probability of a unit $S_j \in S^{(i)}$ is defined as

$$p^{\theta}_{e}(\bar{s}_j \mid s) = |\Omega_{X^{(i)}}|^{-1} \sum_{x^{(i)} \in \Omega_{X^{(i)}}} p_e(\bar{s}_j \mid x^{(i)}, y^{(i)}), \tag{3.19}$$

and $y^{(i)} = s \setminus x^{(i)}$. This means that all connections to unit S_j that are affected by the partition are *causally* *marginalized* (replaced by independent noise).

System integrated information φ_s

The integrated effect information φ_e measures how much the partition $\theta \in \Theta_S$ reduces the probability with which a system S in state $s \in \Omega_S$ specifies its effect state s'_e (3.13),

$$\varphi_e(\mathcal{T}_e, s, \theta) = p_e(s'_e \mid s) \left| \log\left(\frac{p_e(s'_e \mid s)}{p_e^{\theta}(s'_e \mid s)} \right) \right|_{+}. \tag{3.20}$$

Note that φ_e has the same form as the intrinsic information $\mathrm{ii}_e(s, \bar{s})$ (3.6), with the partitioned effect probability taking the place of the unconstrained (marginal) probability. Here, $|.|_+$ represents the positive part operator, which sets the negative values to 0. This ensures that the system as a whole raises the probability of the effect state compared to the partitioned probability. Likewise, the integrated cause information φ_c is defined as

$$\varphi_c(\mathcal{T}_c, s, \theta) = p_c^{\leftarrow}(s'_c \mid s) \left| \log\left(\frac{p_c(s \mid s'_c)}{p_c^{\theta}(s \mid s'_c)} \right) \right|_{+}. \tag{3.21}$$

(By the principle of maximal existence, if two or more cause-effect states are tied for maximal intrinsic information, the system specifies the one that maximizes $\varphi_{c/e}$.)

By the zeroth postulate, existence requires cause *and* effect power, and the integration postulate requires that its cause-effect power be irreducible. By the principle of minimal existence (Box 2), then, system integrated information for a given partition is the minimum of its irreducibility on the cause *and* effect sides:

$$\varphi_s(\mathcal{T}_e, \mathcal{T}_c, s, \theta) = \min\{\varphi_c(\mathcal{T}_c, s, \theta), \varphi_e(\mathcal{T}_e, s, \theta)\}. \tag{3.22}$$

Moreover, again by the principle of minimal existence, the integrated information of a system is given by its irreducibility over its minimum partition (MIP) $\theta' \in \Theta_S$, such that

$$\varphi_s(\mathcal{T}_e, \mathcal{T}_c, s) := \varphi_s(\mathcal{T}_e, \mathcal{T}_c, s, \theta'). \tag{3.23}$$

The MIP is defined as the partition $\theta \in \Theta_S$ that minimizes the system's integrated information, relative to the maximum possible value it could take for arbitrary TPMs $\mathcal{T}'_e, \mathcal{T}'_c$ over the units of system

S:

$$\theta' = \underset{\theta\in\Theta(S)}{\operatorname{argmin}} \frac{\varphi_s(\mathcal{T}_e, \mathcal{T}_c, s, \theta)}{\max\limits_{\mathcal{T}_e', \mathcal{T}_c'} \varphi_s(\mathcal{T}_e', \mathcal{T}_c', s, \theta)}. \tag{3.24}$$

Accordingly, the system is reducible if at least one partition $\theta \in \Theta_S$ makes no difference to the cause or effect probability. The normalization term in the denominator of (3.24) ensures that $\varphi_s(\mathcal{T}_e, \mathcal{T}_c, s)$ is evaluated fairly over a system's fault lines by assessing integration relative to its maximum possible value over a given partition. Using the *relative* integrated information quantifies the strength of the interactions between parts in a way that does not depend on the number of parts and their size. As proven in Marshall et al. (2023), the maximal value of $\varphi_s(\mathcal{T}_e, \mathcal{T}_c, s, \theta)$ for a given partition θ is the normalization factor $\max\limits_{\mathcal{T}_e', \mathcal{T}_c'} \varphi_s(\mathcal{T}_e', \mathcal{T}_c', s, \theta) = \sum\limits_{i=1}^{k} |S^{(i)}||X^{(i)}|$, which corresponds to the maximal possible number of "connections" (pairwise interactions) affected by θ. For example, as shown in Marshall et al. (2023), the MIP will correctly identify the fault line dividing a system into two large subsets of units linked through a few interconnected units (a "bridge"), rather than defaulting to partitions between individual units and the rest of the system. Once the minimum partition has been identified, the integrated information across it is an *absolute* quantity, quantifying the loss of intrinsic information due to cutting the minimum partition of the system. (If two or more partitions $\theta \in \Theta(S)$ minimize Eq. (3.24), we select the partition with the largest unnormalized φ_s value as θ', applying the principle of maximal existence.) Defining θ' as in (3.24), moreover, ensures that $\varphi_s(\mathcal{T}_e, \mathcal{T}_c, s) = 0$ if the system is not *strongly connected* in graph-theoretic terms.[7]

In summary, the system integrated information ($\varphi_s(\mathcal{T}_e, \mathcal{T}_c, s)$, also called "small phi," quantifies the extent to which system S in state s has cause-effect power over itself as *one* system (*i.e.*, irreducibly). $\varphi_s(\mathcal{T}_e, \mathcal{T}_c, s)$ is thus a quantifier of irreducible existence.

Exclusion: Determining maximal substrates (complexes)

In general, multiple candidate systems with overlapping units may have positive values of $\varphi_s(\mathcal{T}_e, \mathcal{T}_c, s)$. By the exclusion postulate, the substrate of consciousness must be definite; that is, it must comprise a definite set of units. But which one? Once again, we employ the principle of maximal existence (Box 2): among candidate systems competing over the same substrate with respect to an essential requirement

[7] *Strongly connected* means that every node can be reached from every other node in a directed graph.

for existence, in this case irreducibility, the one that exists is the one that exists the most. Accordingly, the maximal substrate, or complex, is the candidate substrate with the maximum value of system integrated information (φ_s^*), and overlapping substrates with lower φ_s are thus excluded from existence.

Determining maximal substrates recursively

Within a universal substrate U_0 in state u_0, subsets of units that specify maxima of irreducible cause-effect power (complexes) can be identified iteratively: the substrate with maximum φ_s^* is identified as a complex, the corresponding units are excluded from further consideration, the remaining units are searched for the next maximal substrate. Formally, an iterative search is performed to find a sequence of systems $S_k^* \subseteq U_k$ with

$$\varphi_s^*(\mathcal{T}_e, \mathcal{T}_c, u_k) = \max_{S \subseteq U_k} \varphi_s(\mathcal{T}_e, \mathcal{T}_c, s), \tag{3.25}$$

such that

$$S_k^* = \operatorname*{argmax}_{S \subseteq U_k} \varphi_s(\mathcal{T}_e, \mathcal{T}_c, s), \tag{3.26}$$

and $U_{k+1} = U_k \setminus S_k^*$ until $U_{k+1} = \varnothing$ or $U_{k+1} = U_k$ (the units in $U_0 \setminus U_{k+1}$ still serve as background conditions, for details see Marshall et al. (2023)). If the maximal substrate S_k^* is not unique, and all tied systems overlap, the next best system that is unique is chosen instead (see Appendix B.1, Resolving ties in the IIT algorithm).

For any complex S^* in its corresponding state $s^* \in \Omega_{S^*}$, overlapping substrates that specify less integrated information ($\varphi_s < \varphi_s(\mathcal{T}_e, \mathcal{T}_c, s^*)$) are excluded. Consequently, specifying a maximum of integrated information φ_s^* compared to all overlapping systems

$$S \cap \tilde{S} \neq \varnothing \Rightarrow \varphi_s(s) > \varphi_s(\tilde{s}), \ \ \forall \tilde{S} \neq S \subseteq U \tag{3.27}$$

is a sufficient requirement for a system $S \subseteq U$ to be a complex.

As described in Marshall et al. (2023), this recursive search for maximal substrates "condenses" the universe U_0 in state $u_0 \in \Omega_{U_0}$ into a disjoint (non-overlapping) and exhaustive set of complexes—the first complex, second complex, and so on.

Determining maximal unit grains

Above, we presented how to determine the borders of a complex within a larger system U, assuming a particular grain for the units $U_i \in U$. In principle, however, all possible grains should be considered (Hoel et al., 2016; Marshall et al., 2018). In the brain, for example, the grain of units could be brain regions, groups of neurons, individual neurons, sub-cellular structures, molecules, atoms, quarks, or anything finer, down to hypothetical atomic units of cause-effect power (Tononi & Koch, 2015; Tononi et al., 2016). For any unit grain—neurons, for example—the grain of updates could be minutes, seconds, milliseconds, micro-seconds, and so on. And the grain of the states associated with these updates could be 2 states ("no spikes / any number of spikes per neuron over one hundred milliseconds"), 4 states ("no spikes, 1 spike, 2–5 spikes, bursts of $>$ 5 spikes"), or 256 states (from 0 to 255 spikes in intervals of 1 spike), and so on. However, by the exclusion postulate, the units that constitute a system S must also be definite, in the sense of having a definite grain.

Once again, the grain is defined by the principle of maximal existence: across the possible micro- and macroscopic levels, the "winning" grain is the one that ensures maximally irreducible existence (φ_s^*) for the entity to which the units belong (Hoel et al., 2016; Marshall et al., 2018).

To evaluate integrated information across grains requires a mathematical framework for defining coarser (macro) units from finer (micro) units. Such a framework has been developed in previous work (Hoel et al., 2013, 2016; Marshall et al., 2018), and is updated here to fully align with the postulates.

Supposing that $U = u$ is a universe of micro units in a state, a macro unit $J = j$ is a combination of a set of micro units $\hat{S} \subseteq U$, and a mapping g from the state $\hat{S}$ to the state of J,

$$j = g(\hat{s}),$$

where

$$g : \Omega_{\hat{S}} \to \Omega_J.$$

As constituents of a complex upon which its cause-effect power rests, the units themselves should comply with the postulates of IIT (Tononi, forthcoming). Otherwise it would be possible to "make something out of nothing." Accordingly, units themselves must also be maximally irreducible, as measured by the integrated information of the units when they are treated as candidate systems (φ_s);

otherwise, they would not be units but “disintegrate” into their constituents. However, in contrast to systems, units only need to be maximally irreducible *within*, because they do not exist as complexes in their own right: a unit J with substrate $\hat{S}$ qualifies as a candidate unit of a larger system S if its integrated information when treated as a candidate system (φ_s) is higher than that of any system of units (including potential macro units) that can be defined using a subset of $\hat{S}$. Out of all possible sets of such candidate units, the set of (macro) units that define a complex is the one that maximizes the existence of the complex to which the units belong, rather than their own existence.

In practice, the search for the maximal grain should be an iterative process, starting from micro units: identify potential substrates for macro units ($\hat{S}$) that are maximally irreducible within, identify mappings g that maximize the integrated information of systems of macro units, then consider additional potential substrates for macro units, and so on iteratively, until a global maximum is found. The iterative approach is necessary for establishing that a substrate is maximally irreducible within, as this criterion requires consideration not only of micro units, but also of all finer grains (potential meso units defined from subsets of $\hat{S}$).

Here we outlined an overall framework for identifying macro units consistent with the postulates. Additional details about the nature of the mapping g, and how to derive the transition probabilities for a system of macro units are also be informed by the postulates and presented in Marshall et al. (2023).

3.5 Unfolding the cause-effect structure of a complex

Once a maximal substrate and the associated maximal cause-effect state have been identified, we must unfold its cause-effect power to reveal its cause-effect structure of distinctions and relations, in line with the composition postulate. As components of the cause-effect structure, distinctions and relations must also satisfy the postulates of IIT (save for composition).

Composition and causal distinctions

Causal distinctions capture how the cause-effect power of a substrate is structured by subsets of units that specify irreducible causes and effects over subsets of its units. A candidate distinction $d(m)$ consists of (1) a mechanism $M \subseteq S$ in state $m \in \Omega_M$ inherited from the system state $s \in \Omega_S$;, (2) a maximal cause-effect state $z^* = \{z^*_c, z^*_e\}$ over the cause and effect purviews ($Z_c, Z_e \subseteq S$) linked by the mechanism;

and, and (3) an associated value of irreducibility ($\varphi_d > 0$). A distinction $d(m)$ is thus represented by the tuple

$$d(m) = \Big(m, z^*, \varphi_d\Big). \tag{3.28}$$

For a given mechanism m, our goal is to identify its maximal cause Z_c^* in state $z_c^* \in \Omega_{Z_c^*}$ and its maximal effect Z_e^* in state $z_e^* \in \Omega_{Z_e^*}$ within the system, where $Z_c^*, Z_e^* \subseteq S$.

As above, in line with existence, intrinsicality, and information, we determine the maximal cause or effect state specified by the mechanism over a candidate purview within the system based on the value of intrinsic information $\text{ii}(m, z)$. Next, in line with integration, we determine the value of integrated information $\varphi_d(m, Z, \theta)$ over the minimum partition θ'. In line with exclusion, we determine the maximal cause-effect purviews for that mechanism over all possible purviews $Z \subseteq S$ based on the associated value of irreducibility $\varphi_d(m, Z, \theta')$. Finally, we determine whether the maximal cause-effect state specified by the mechanism is congruent with the system's overall cause-effect state ($z_c^* \subseteq s_c^*$, $z_e^* \subseteq s_e^*$), in which case we conclude that it contributes a distinction to the overall cause-effect structure.

The updated formalism to identify causal distinctions within a system S in state s was first presented in Barbosa et al. (2021). Here we provide a summary with minor adjustments on selecting z_c^* and z_e^*, the cause integrated information $\varphi_c(m, Z)$, and the requirement that causal distinctions must be congruent with the system's maximal cause-effect state (see Appendix B.2, Comparison to IIT 1.0–3.0 and subsequent publications).

Existence, intrinsicality, and information: Determining the cause and effect state specified by a mechanism over candidate purviews

Like the system as a whole, its subsets must comply with existence, intrinsicality, and information. As for the system, we begin by quantifying, in probabilistic terms, the difference a subset of units $M \subseteq S$ in its current state $m \subseteq s$ takes and makes from and to subsets of units $Z \subseteq S$ (cause and effect purview). As above, we start by establishing the interventional conditional probabilities and unconstrained probabilities from the TPMs $\mathcal{T}_c$ and $\mathcal{T}_e$.

When dealing with a mechanism constituted by a subset of system units, it is important to capture the constraints on a purview state z that are exclusively due to the mechanism in its state (m), removing any potential contribution from other system units. This is done by causally marginalizing all variables in $X = S \setminus M$, which corresponds to imposing a uniform distribution as $p(X)$ (Albantakis et al., 2019; Barbosa et al., 2021; Oizumi et al., 2014).[8] The effect probability of a single unit $Z_i \in Z$ conditioned on the current state m is thus defined as

$$p_e(z_i \mid m) = |\Omega_X|^{-1} \sum_{x \in \Omega_X} p\left(z_i \mid m, x\right), \quad z_i \in \Omega_{Z_i}. \tag{3.29}$$

In addition, product probabilities $\pi(z \mid m)$ are used instead of conditional probabilities $p_e(z \mid m)$ to discount correlations from units in $X = S \setminus M$ with divergent outputs to multiple units in $Z \subseteq S$ (Albantakis et al., 2019; S. Krohn & Ostwald, 2017; Oizumi et al., 2014). Otherwise, X might introduce correlations in Z that would be wrongly considered as effects of M. Based on the appropriate TPM, the probability over a set Z of $|Z|$ units is thus defined as the product of the probabilities over individual units

$$\pi_e(z \mid m) = \prod_{i=1}^{|Z|} p_e(z_i \mid m), \quad z \in \Omega_Z, \tag{3.30}$$

and

$$\pi_c(m \mid z) = \prod_{i=1}^{|M|} p_c(m_i \mid z), \quad m \in \Omega_M. \tag{3.31}$$

Note that for a single unit purview $\pi_e(z \mid m) = p_e(z \mid m)$, and for a single unit mechanism $\pi_c(m \mid z) = p_c(m \mid z)$. By using product probabilities, causal marginalization maintains the conditional independence between units (3.2) because independent noise is applied to individual connections. The assumption of conditional independence distinguishes IIT's causal powers analysis from standard information-theoretic analyses of information flow (Albantakis et al., 2019; Ay & Polani, 2008) and corresponds to an assumption that variables are "physical" units in the sense that they are irreducible within and can be observed and manipulated independently.

From Eq. (3.30) and Eq. (3.31) we can also define unconstrained probabilities

$$\pi_e(z; M) = |\Omega_M|^{-1} \sum_{m \in \Omega_M} \pi_e(z \mid m), \quad z \in \Omega_Z, \tag{3.32}$$

[8] Units within the candidate system are causally marginalized based on a uniform distribution to discount their causal contribution if they are not part of the mechanism or purview under consideration. By contrast, units outside the candidate system (background conditions) are causally marginalized conditional on the current state of the universe, potentially leading to a non-uniform distribution.

and

$$\pi_c(m; Z) = |\Omega_Z|^{-1} \sum_{z \in \Omega_Z} \pi_c(m \mid z), \quad m \in \Omega_M. \tag{3.33}$$

Given the set $Y = S \setminus Z$, the backward cause probability (selectivity) for a mechanism m with $|M|$ units is computed using Bayes' rule over the product distributions

$$\pi_c^{\leftarrow}(z \mid m) = \frac{\pi_c(m \mid z) \cdot |\Omega_Z|^{-1}}{\pi_c(m; Z)} = \frac{\prod_{i=1}^{|M|} p_c(m_i \mid z)}{\sum_{\hat{z} \in \Omega_Z} \prod_{i=1}^{|M|} p_c(m_i \mid \hat{z})}, \quad z \in \Omega_Z, \tag{3.34}$$

where $p_c(m_i \mid z) = |\Omega_Y|^{-1} \sum_{y \in \Omega_Y} p_c(m_i \mid z, y)$ in line with (3.29).

To correctly quantify intrinsic causal constraints, the marginal probability of possible cause states (for computing $\pi_c^{\leftarrow}(z \mid m)$ or $\pi_c(m; Z)$) is again set to the uniform distribution. As above, all probabilities are obtained from the TPMs $\mathcal{T}_e$ (3.4) and $\mathcal{T}_c$ (3.5) and thus correspond to *interventional* probabilities throughout.

Having defined cause and effect probabilities, we can now evaluate the intrinsic information of a mechanism m over a purview state $z \in \Omega_Z$ analogously to the system intrinsic information (3.6) and (3.8). The intrinsic effect information that a mechanism in a state m specifies about a purview state z is

$$\underset{e}{\text{ii}}(m, z) = \pi_e(z \mid m) \log\left(\frac{\pi_e(z \mid m)}{\pi_e(z; M)}\right). \tag{3.35}$$

The intrinsic cause information that a mechanism in a state m specifies about a purview state z is

$$\underset{c}{\text{ii}}(m, z) = \pi_c^{\leftarrow}(z \mid m) \log\left(\frac{\pi_c(m \mid z)}{\pi_c(m; Z)}\right). \tag{3.36}$$

As with system intrinsic information, the logarithmic term is the informativeness, which captures how much causal power is exerted by the mechanism m on its potential effect z (how much it increases the probability of that state above chance), or by the potential cause z on the mechanism m. The term in front of the logarithm corresponds to the mechanism's selectivity, which captures how much the causal power of the mechanism m is concentrated on a specific state of its purview (as opposed to other states). In the following we will again focus on the effect side, but an equivalent procedure applies on the cause side (see Appendix B.4, IIT Algorithm).

Based on the principle of maximal existence, the maximal effect state of m within the purview Z is defined as

$$z_e'(m, Z) = \underset{z \in \Omega_Z}{\operatorname{argmax}} \underset{e}{\mathrm{ii}}(m, z), \tag{3.37}$$

which corresponds to the specific effect of m on Z. Note that z_e' is not always unique (see Appendix B.1, Resolving ties in the IIT algorithm). The maximal intrinsic information of mechanism m over a purview Z is then

$$\underset{e}{\mathrm{ii}}(m, Z) := \underset{e}{\mathrm{ii}}(m, z_e') = \max_{z \in \Omega_Z} \underset{e}{\mathrm{ii}}(m, z). \tag{3.38}$$

Note that, by this definition, if $\mathrm{ii}_e(m, Z) \neq 0$, mechanism m always raises the probability of its maximal effect state compared to the unconstrained probability. This is because there is at least one state $z \in \Omega_Z$ such that $\pi_e(z \mid m) > \pi_e(z; M)$.

The intrinsic information of a candidate distinction, like that of the system as a whole, is sensitive to indeterminism (the same state leading to multiple states) and degeneracy (multiple states leading to the same state) because both factors decrease the probability of the selected state. Moreover, the product of selectivity and informativeness leads to a tension between expansion and dilution: larger purviews tend to increase informativeness because conditional probabilities will deviate more from chance, but they also tend to decrease selectivity because of the larger repertoire of states.

Integration: Determining the irreducibility of a candidate distinction

To comply with integration, we must next ask whether the specific effect of m on Z is irreducible. As for the system, we do so by evaluating the integrated information $\varphi_e(m, Z)$. To that end, we define a set of "disintegrating" partitions $\Theta(M, Z)$ as

$$\begin{aligned}\Theta(M, Z) = \Bigg\{ \left\{ \left(M^{(i)}, Z^{(i)} \right) \right\}_{i=1}^{k} \;:\; k \in \{2, 3, 4, \ldots\},\ M^{(i)} \in \mathbb{P}(M),\ Z^{(i)} \in \mathbb{P}(Z),\ \bigcup M^{(i)} = M,\\ \bigcup Z^{(i)} = Z,\ Z^{(i)} \cap Z^{(j)} = M^{(i)} \cap M^{(j)} = \varnothing\ \forall\, i \neq j,\ M^{(i)} = M \Longrightarrow Z^{(i)} = \varnothing \Bigg\},\end{aligned} \tag{3.39}$$

where $\{M^{(i)}\}$ is a partition of M and $\{Z^{(i)}\}$ is a partition of Z, but the empty set may also be used as a part ($\mathbb{P}$ denotes the power set). As introduced in Albantakis et al. (2019) and Barbosa et al. (2021), a disintegrating partition $\theta \in \Theta(M, Z)$ either "cuts" the mechanism into at least two independent parts if $|M| > 1$, or it severs all connections between M and Z, which is always the case if $|M| = 1$ (we refer to (Albantakis et al., 2019; Barbosa et al., 2021) for details). Note that disintegrating partitions differ from system partitions (3.24), which divide the system into two or more parts in a directed manner to evaluate whether and to what extent the system is integrated in terms of its cause-effect power. Instead, disintegrating partitions apply to mechanism-purview pairs within the system, which are already directed, to evaluate the cause or effect power specified by the mechanism over its purview.

Given a partition $\theta \in \Theta(M, Z)$, we can define the partitioned effect probability

$$\pi_e^\theta(z_e' \mid m) = \prod_{i=1}^{k} \pi_e(z_e'^{(i)} \mid m^{(i)}), \tag{3.40}$$

with $\pi(\varnothing | m^{(i)}) = \pi(\varnothing) = 1$. In the case of $m^{(i)} = \varnothing$, $\pi_e(z_e'^{(i)} | \varnothing)$ corresponds to the fully partitioned effect probability

$$\pi_e(z \mid \varnothing) = \prod_{i=1}^{|Z|} \sum_{s \in \Omega_S} p_e(z_i \mid s) |\Omega_S|^{-1}. \tag{3.41}$$

The integrated effect information of mechanism m over a purview $Z \subseteq S$ with effect state z_e' for a particular partition $\theta \in \Theta(M, Z)$ is then defined as

$$\varphi_e(m, Z, \theta) = \pi_e(z_e' \mid m) \left| \log \left(\frac{\pi_e(z_e' \mid m)}{\pi_e^\theta(z_e' \mid m)} \right) \right|_+ . \tag{3.42}$$

The effect of m on z_e' is reducible if at least one partition $\theta \in \Theta(M, Z)$ makes no difference to the effect probability or increases it compared to the unpartitioned probability. In line with the principle of minimal existence, the total integrated effect information $\varphi_e(m, Z)$ again has to be evaluated over θ', the minimum partition (MIP)

$$\varphi_e(m, Z) := \varphi_e(m, Z, \theta'), \tag{3.43}$$

which requires a search over all possible partitions $\theta \in \Theta(M, Z)$:

$$\theta' = \underset{\theta \in \Theta(M,Z)}{\operatorname{argmin}} \frac{\varphi(m, Z, \theta)}{\max_{\mathcal{T}'} \varphi(m, Z, \theta)}. \tag{3.44}$$

As in (3.24), the minimum partition is evaluated against its maximum possible value across all possible systems TPMs $\mathcal{T}'$, which again corresponds to the number of possible pairwise interactions affected by the partition.

The integrated cause information is defined analogously, as

$$\varphi_c(m, Z) := \varphi_c(m, Z, \theta') = \pi_c^{\leftarrow}(z'_c \mid m) \left| \log\left(\frac{\pi_c(m \mid z'_c)}{\pi_c^{\theta'}(m \mid z'_c)}\right)\right|_+, \tag{3.45}$$

where the partitioned probability $\pi_c^{\theta}(m \mid z)$ is again a product distribution over the parts in the partition, as in (3.40).

Taken together, the intrinsic information (3.38) determines what cause or effect state the mechanism m specifies. Its integrated information quantifies to what extent m specifies its cause or effect in an irreducible manner. Again, $\varphi(m, Z)$ is a quantifier of irreducible existence.

Exclusion: Determining causal distinctions

Finally, to comply with exclusion, a mechanism must select a definite effect purview, as well as a cause purview, out of a set of candidate purviews. Resorting again to the principle of maximal existence, the mechanism's effect purview and associated effect is the one having the maximum value of integrated information across all possible purviews $Z \subseteq S$ in state $z'_e(m, Z)$ (3.37)

$$z_e^*(m) = \operatorname*{argmax}_{Z \subseteq S} \varphi_e(m, z'_e(m, Z)). \tag{3.46}$$

The integrated effect information of a mechanism m within S is then

$$\varphi_e(m) := \varphi_e(m, z_e^*(m)) = \max_{Z \subseteq S} \varphi_e(m, z'_e(m, Z)). \tag{3.47}$$

The integrated cause information $\varphi_c(m)$ and the maximally irreducible cause $z_c^*(m)$ are defined in the same way (see Appendix B.4, IIT Algorithm). Based again on the principle of minimal existence, the irreducibility of the distinction specified by a mechanism is given by the minimum between its integrated cause and effect information

$$\varphi_d(m) = \min\big(\varphi_c(m), \varphi_e(m)\big). \tag{3.48}$$

Determining the set of causal distinctions that are congruent with the system cause-effect state

As required by composition, unfolding the full cause-effect structure of the system S in state s requires assessing the irreducible cause-effect power of every subset of units within S (Figure 3.2). Any $m \subseteq s$ with $\varphi_d > 0$ specifies a candidate distinction $d(m) = (m, z^*, \varphi_d)$ (3.28) within the system S in state s. However, in order to contribute to the cause-effect structure of a system, distinctions must also comply with intrinsicality and information at the system level. Thus, the fact that the system must select a specific cause-effect state implies that the cause-effect state they specify over subsets of the system ($z^* = \{z_c^*, z_e^*\}$) must be congruent with the cause-effect state specified over itself by the system as a whole s'.

We thus define the set of all causal distinctions within S in state s as

$$D(\mathcal{T}_e, \mathcal{T}_c, s) = \{d(m) \ : \ m \subseteq s,\ \varphi_d(m) > 0,\ z_c^*(m) \subseteq s_c',\ z_e^*(m) \subseteq s_e'\}. \tag{3.49}$$

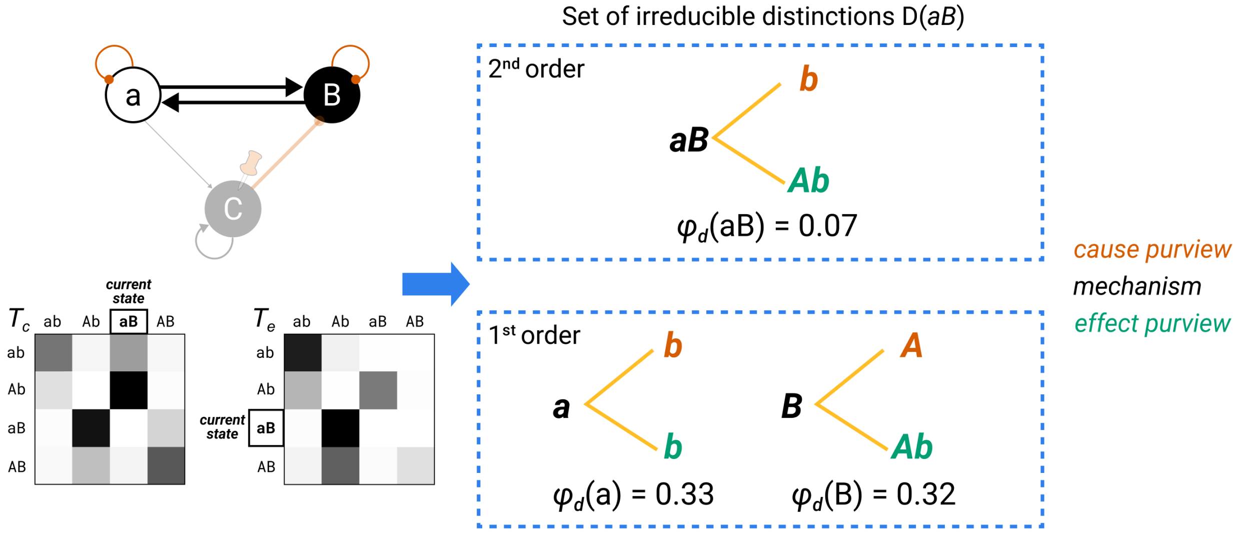


Figure 3.2. Composition and causal distinctions. Identifying the irreducible causal distinctions specified by a substrate in a state requires evaluating the specific causes and effects of every system subset. The candidate substrate is constituted of two interacting units $S = aB$ (see Figure 3.1) with TPMs $\mathcal{T}_e$ and $\mathcal{T}_c$ as shown. In addition to the two first-order mechanisms a and B, the second-order mechanism aB specifies its own irreducible cause and effect, as indicated by $\varphi_d > 0$.

Altogether, distinctions can be thought of as irreducible "handles" through which the system can take and make a difference to itself by linking an intrinsic cause to an intrinsic effect over subsets of itself. As components within the system, causal distinctions have no inherent structure themselves.

Whatever structure there may be between the units that make up a distinction is not a property of the distinction but due to the structure of the system, and thus captured already by its compositional set of distinctions. Similarly, from an extrinsic perspective, one may uncover additional causes and effects, both within the system and across its borders, at either macro or micro grains. However, from the intrinsic perspective of the system causes and effects that are excluded from its cause-effect structure do not exist (Albantakis & Tononi, 2019; Grasso et al., 2021).

For example, as shown in Figure 3.3A, a system may have a mechanism through which it specifies, in a maximally irreducible manner, the effect state of a triplet of units (*e.g.*, $z_e^* = abc$, a third-order purview; again lowercase letters for units indicate state "-1", uppercase letters state "$+1$"). However, if the system lacks a mechanism through which it can specify the effect state of single units, each taken individually (say, unit a, a first-order effect purview), then, from its intrinsic perspective, that unit does not exist as a single unit. By the same token, if the system can specify individually the state of unit a, b, and c, but lacks a way to specify irreducibly the state of abc together, then, from its intrinsic perspective, the triplet abc does not exist as a triplet (see Figure 3.3B). Finally, even if the system can distinguish the single units a, b, and c, as well as the triplet abc, if it lacks handles to distinguish pairs of units such as ab and bc, it cannot order units in a sequence.

Composition and causal relations

Causal relations capture how the causes and/or effects of a set of distinctions within a complex overlap with each other. Just as a distinction specifies which units/states constitute a cause purview and the linked effect purview, a relation specifies which units/states correspond to which units/states among the purviews of a set of distinctions. Relations thus reflect how the cause-effect power of its distinctions is "bound together" within a complex. The irreducibility due to this binding of cause-effect power is measured by the relations' irreducibility ($\varphi_r > 0$). Relations between distinctions were first described in Haun and Tononi (2019) (for differences with the initial presentation see Appendix B.2).

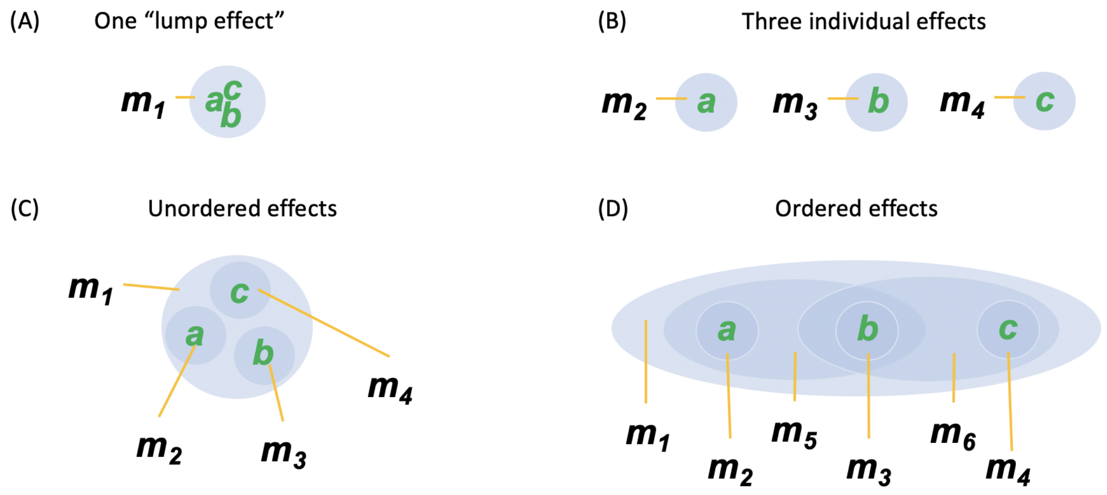


Figure 3.3. Composition of intrinsic effects. From the intrinsic perspective of the system, a specific cause or effect is only available to the system if it is selected by a causal distinction $d \in D(s)$. In **(A)**, only the top-order effect is specified. From the intrinsic perspective, the system cannot distinguish the individual units. In **(B)**, only first-order effects are specified. The system has no "handle" to select all three units together. **(C)** If both first- and third-order effects are specified, but no second-order effects, the system can distinguish individual units and select them together, but has no way of ordering them sequentially. **(D)** The system can distinguish individual units, select them altogether, as well as order them sequentially, in the sense that it has a handle for *ab* and *bc*, but not *ac*. The ordering becomes apparent once the relations among the distinctions are considered (see below, Figure 3.5).

A set of distinctions $\boldsymbol{d} \subseteq D(s)$ is related if the cause-effect state of each distinction $d \in \boldsymbol{d}$ overlaps congruently over a set of shared units, which may be part of the cause, the effect, or both the cause and the effect of each distinction. Below we will denote the cause of a distinction d as $z_c^*(d)$ and its effect as $z_e^*(d)$. For a given set of distinctions $\boldsymbol{d} \subseteq D(s)$, there are potentially many "relating" sets of causes and/or effects $\boldsymbol{z}$ such that

$$\boldsymbol{z} \; : \; \boldsymbol{z} \cap \{z_c^*(d), z_e^*(d)\} \neq \varnothing \;\; \forall d \in \boldsymbol{d}, \; \bigcap_{z \in \boldsymbol{z}} z \neq \varnothing, \; |\boldsymbol{z}| > 1 \tag{3.50}$$

with maximal overlap

$$o^*(\boldsymbol{z}) = \bigcap_{z \in \boldsymbol{z}} z \neq \varnothing. \tag{3.51}$$

Since $z_c^*(m) \subseteq s_c'$ and $z_e^*(m) \subseteq s_e'$ are sets of tuples containing both the units and their states, the intersection operation considers both the units and the state of the units.

All possible sets $\boldsymbol{z}$ specify unique aspects about a relation $r(\boldsymbol{d})$ and constitute the various "faces" of the relation (Figure 3.4). The maximal overlap $o^*(\boldsymbol{z})$ (3.51) is also called the "face purview." The set of faces associated with a relation thus specifies which type of relation it is (e.g., a single-faceted relation that only relates the causes of the set of distinctions, or a multi-faceted relation, which requires some of the distinctions to overlap on both the cause and effect side). Note that (3.50) includes the case $\boldsymbol{z} = \{z_c^*(d), z_e^*(d)\}$, which indicates a "self-relation" over the cause and effect of a single distinction $d \in D(s)$.

A relation $r(\boldsymbol{d})$ thus consists of a set of distinctions $\boldsymbol{d} \subseteq D(s)$, with an associated set of faces $\boldsymbol{f}(\boldsymbol{d}) = \{f(\boldsymbol{z})\}_{\boldsymbol{d}}$ and irreducibility $\varphi_r > 0$,

$$r(\boldsymbol{d}) = \Big(\boldsymbol{d}, \boldsymbol{f}(\boldsymbol{d}), \varphi_r\Big). \tag{3.52}$$

A relation that binds together $h = |\boldsymbol{d}|$ distinctions is a h-degree relation. A relation face $f(\boldsymbol{z}) \in \boldsymbol{f}(\boldsymbol{d})$ consists of a set of causes and effects $\boldsymbol{z}$ (as in (3.50)), with associated face purview $o^*(\boldsymbol{z})$(3.51)

$$f(\boldsymbol{z}) = \Big(\boldsymbol{z}, o^*(\boldsymbol{z})\Big). \tag{3.53}$$

A relation face over $k = |\boldsymbol{z}|$ purviews is a k-degree face. The set of faces includes all the ways in which the set of distinctions $\boldsymbol{d}$ counts as related according to (3.50). Because $\boldsymbol{z}$ may include either the cause, or the effect, or both the cause and effect of a distinction $d \in \boldsymbol{d}$, a relation $r(\boldsymbol{d})$ with $|\boldsymbol{d}| > 1$ may comprise up to $3^{|\boldsymbol{d}|}$ faces. If a set of distinctions $\boldsymbol{d} \in D(s)$ does not overlap congruently, it is not related (in that case $o^*(\boldsymbol{z}) = \varnothing$ for all possible $f(\boldsymbol{z}) \in \boldsymbol{f}(\boldsymbol{d})$) (Figure 3.5).

Causal relations inherit existence from the cause-effect power of the distinctions that compose them. They inherit intrinsicality because the causes and effects that compose their faces are specified within the substrate. Moreover, relations are specific because the joint purviews of their faces must be congruent for all causes and effects $z^* \in \boldsymbol{z}$. Note that relation purviews are necessarily congruent with the overall cause and effect state specified by the system as a whole, because the causes and effects of the distinctions composing a relation must themselves be congruent.

The irreducibility of a causal relation is measured by "unbinding" distinctions from their joint purviews, taking into account all faces of the relation. Distinctions $d \in D(s)$ are already established as maximally irreducible components, characterized by their value of integrated information φ_d. To assess the irreducibility of a relation, we thus assume that the integrated information φ_d of a distinction is distributed uniformly across unique cause and effect purview units, such that

$$\frac{\varphi_d}{|z_c^*(d) \cup z_e^*(d)|} \tag{3.54}$$

is the average irreducible information φ_d per unique purview unit for an individual distinction $d \in \boldsymbol{d}$ with cause-effect state $z^*(d) = \{z_c^*(d), z_e^*(d)\}$. Since the union operator takes the states of the units into account, incongruent units are counted separately, while congruent units on the cause and effect side count as one.

Since distinctions are related by specifying common units into common states, the effect of "unbinding" a distinction must be proportional to the number of units jointly specified in the relation, *i.e.* the number of distinct units over the joint purviews of all faces in the relation:

$$\left| \bigcup_{f \in \boldsymbol{f}(\boldsymbol{d})} o_f^* \right|. \tag{3.55}$$

This union of the face purviews o_f^* is also called the "relation purview" or the "joint purview" of the relation. While any partition of one or more distinctions from the relation will "unbind" the set of distinctions $\boldsymbol{d}$, by the principle of minimal existence, a relation can only be as irreducible as the minimal amount of integrated information specified by any one distinction in the relation. Therefore, the relation integrated information $\varphi_r(\boldsymbol{d})$ is defined as

$$\varphi_r(\boldsymbol{d}) = \min_{d \in \boldsymbol{d}} \left| \bigcup_{f \in \boldsymbol{f}(\boldsymbol{d})} o_f^* \right| \frac{\varphi_d}{|z_c^*(d) \cup z_e^*(d)|}. \tag{3.56}$$

In words, for each distinction, we take the average integrated information per distinct purview element (3.54), multiply it by the number of units across all faces of the relation (3.55), and then find the distinction that contributes the least integrated information per overlap unit as the minimum partition of the relation (with corresponding integrated information φ_r). Defining φ_r in this way guarantees that the integrated information of a relation cannot exceed the integrated information of its weakest distinction. For a given set of distinctions, the maximum value of φ_r occurs for a relation in which the cause and effect of each distinction is fully overlapped by all other distinctions in the relation (in that case, $\varphi_r = \min_{d \in \boldsymbol{d}} \varphi_d$). Note also that a relation satisfies exclusion (distinctions overlap on *this whole*

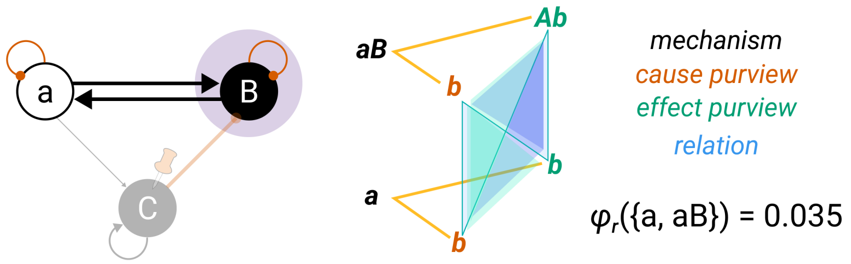


Figure 3.4. Composition and causal relations. Relations between distinctions specify joint causes and/or effects. The two distinctions $d(a)$ and $d(aB)$ each specify their own cause and effect. In this example, their cause and effect purviews overlap over the unit b and are congruent, which means that they all specify b to be in state "-1". The relation $r(\{a, aB\})$ thus binds the two distinctions together over the same unit. Relation faces are indicated by the blue lines and surfaces between the distinctions' causes and/or effects (different shades are used to individuate the faces). Because all four purviews overlap over the same unit, all nine possible faces exist. Note that the fact that the two distinctions overlap irreducibly can only be captured by a relation and not by a high-order distinction.

set of units) in that its integrated information is naturally maximized (per the principle of maximal existence) over the maximal congruent overlap o_f^* for each relation face (3.51) (taking subsets of these overlaps could only reduce the integrated information of the relation).

In summary, just as distinctions link a cause with an effect, relations bind various combinations of causes and effects that are congruent over the same units (Figure 3.4). And just as a distinction captures the irreducibility of an individual cause-effect linked by a mechanism, a relation captures the irreducibility of a set of distinctions bound by the joint purviews of their causes and/or effects.

For a set of distinctions D, we define the set of all relations among them as

$$R(D) = \{r(\boldsymbol{d}) : \varphi_r(\boldsymbol{d}) > 0\}, \; \forall \boldsymbol{d} \subseteq D. \tag{3.57}$$

In practice, the total number of relations and their $\sum_{\boldsymbol{R(D)}} \varphi_r$ can be determined analytically for a given set of distinctions D, which greatly reduces the necessary computations (see Appendix B.3, Analytical solution for $\sum \varphi_r$ and the number of causal relations). Together, a set of distinctions D and its associated set of relations $R(D)$ compose a cause-effect structure.

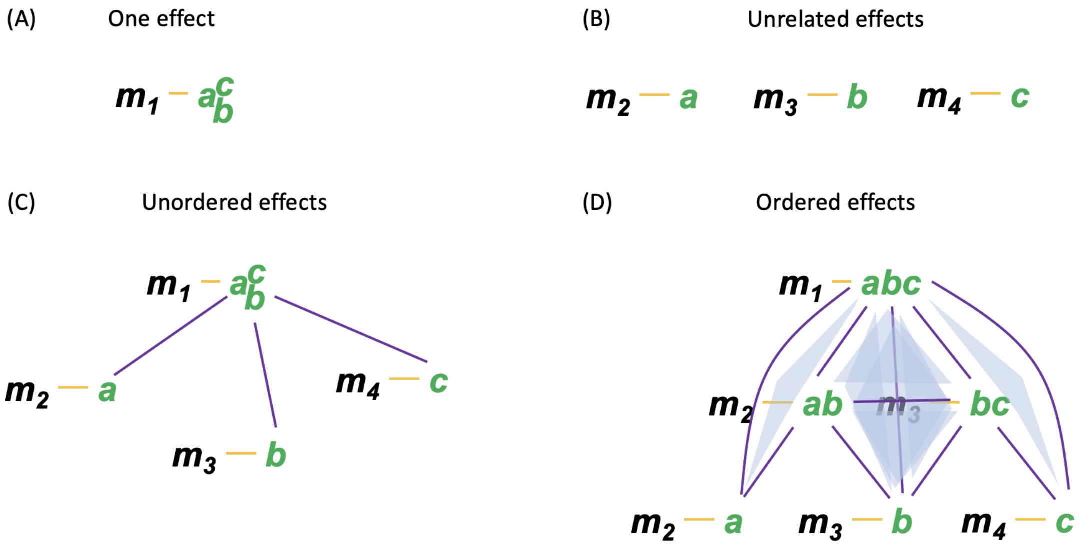


Figure 3.5. Structuring of intrinsic effects by relations. (A) A single undifferentiated effect has no relations. **(B)** Likewise, there are no relations among multiple non-overlapping effects. **(C)** The set of three first-order effects and one third-order effect supports three relations, which bind the effects together. **(D)** The set of first, second, and third-order effects supports a large number of relations (ten 2-relations (between two effects), six 3-relations, and one 4-relation), which bind the effects in a structure that is ordered sequentially.

Cause-effect structures and Φ-structures

A cause-effect structure is defined as the union of the distinctions specified by a substrate and the relations binding them together:

$$C(D) = D \cup R(D). \tag{3.58}$$

The cause-effect structure specified by a maximal substrate—a complex—is also called a Φ-structure:

$$C(\mathcal{T}_e, \mathcal{T}_c, s^*) = \left\{ \left\{ d(m) = \{m, z^*, \varphi_d\} \in D(\mathcal{T}_e, \mathcal{T}_c, s^*) \right\} \right. \\ \left. \bigcup \left\{ r(\boldsymbol{d}) = \{\boldsymbol{d}, \boldsymbol{f}(d), \varphi_r\} \in R\big(D(\mathcal{T}_e, \mathcal{T}_c, s^*)\big) \right\} \right\}. \tag{3.59}$$

The sum of the values of integrated information of a substrate's distinctions and relations, called Φ ("big Phi". "structure Phi") corresponds to the *structure integrated information* of the Φ-structure,

$$\Phi(\mathcal{T}_e, \mathcal{T}_c, s^*) = \sum_{C(\mathcal{T}_e, \mathcal{T}_c, s^*)} \varphi. \tag{3.60}$$

Note that Φ is not computed based on a partition (as system phi), but rather a sum of the integrated information within the structure (where each term of the sum was computed by partitioning). Within a Φ-structure, various types of meaningful sub-structures can be specified, which we term *Φ-folds*. A Φ-fold is composed of a subset of the distinctions and relations that compose the overall cause-effect structure. A special case is the *distinction Φ-fold*, denoted $C(\{d\})$, a sub-structure composed of a single distinction and the relations bound to it, which form its *context* (Haun & Tononi, 2019).[9] A *compound Φ-fold* is a sub-structure composed of the distinction Φ-folds specified by a subset of units. A *compound Φ-fold* is a relevant part of a Φ-structure because it can be accessed or manipulated by changing the state, connections, or functioning of a part of the substrate. Finally, a *content Φ-fold*, or simply *content*, is composed of a subset of distinctions that are highly interrelated (regardless of the mechanisms and units that specify them).

In conclusion, a maximal substrate or complex is a set of units $S^* = s^*$ that satisfies all of IIT's postulates: its cause-effect power is intrinsic, specific, irreducible, definite, and structured. By IIT, a complex S^* does not exist as such, but exists "unfolded" into its Φ-structure, with all the causal distinctions and relations that compose it. In other words, a substrate is what can be observed and manipulated "operationally" from the extrinsic perspective. From the intrinsic perspective, what truly exists is a complex with all its causal powers unfolded—an *intrinsic entity* that exists for itself, absolutely, rather than relative to an external observer.

According to the explanatory identity of IIT, an experience is identical to the Φ-structure of an intrinsic entity: every property of the experience should be accounted for by a corresponding property of the Φ-structure, with no additional ingredients. If a system S in state s is a complex, then its Φ-structure corresponds to the quality of the experience of S in state s, while its Φ value corresponds to its quantity—in other words, to the nature and amount of intrinsic content.

3.6 Results and discussion

In this section, we apply the mathematical framework of IIT 4.0 to several example systems. The goal is to illustrate three critical implications of IIT's postulates:

[9] It is useful to note that we can partition a cause-effect structure into distinction Φ-folds as follows. To do so, we assume that each distinction contributes equally to the existence of a relation $r(\boldsymbol{d})$, because removing any distinction $d \in \boldsymbol{d}$ will "unrelate" the set $\boldsymbol{d}$. Thus, we assign the proportion of $\varphi_r(\boldsymbol{d})$ that each individual distinction $d \in \boldsymbol{d}$ contributes to the full quantity to be

(1) **Consciousness and connectivity:** how the way units interact determines whether a substrate can support a $\mathit{\Phi}$-structure of high $\mathit{\Phi}$.

(2) **Consciousness and activity:** how changes in the state of a substrate's units change $\mathit{\Phi}$-structures.

(3) **Consciousness and functional equivalence:** how substrates that are functionally equivalent may not be equivalent in terms of their $\mathit{\Phi}$-structures, and thus in terms of consciousness.

The following examples will feature very simple networks constituted of binary units $U_i \in U$ with $\Omega_{U_i} = \{-1, 1\}$ for all U_i and a logistic (sigmoidal) activation function

$$p(U_{i,t} = 1 \mid u_{t-1}) = \frac{1}{1 + \exp\left(-k \sum_{j=1}^{n} w_{j,i} u_{j,t-1}\right)}, \tag{3.61}$$

where $k > 0$ and

$$\sum_{j=1}^{n} w_{j,i} = 1 \quad \forall\, i. \tag{3.62}$$

In (3.61), the parameter k defines the slope of the logistic function and allows one to adjust the amount of noise or determinism in the activation function (higher values signify a steeper slope and thus more determinism). The units U_i can thus be viewed as noisy linear threshold units with weighted connections among them, where k determines the connection strength.

As in Figures 3.1 and 3.2, units denoted by uppercase letters are in state '1' (ON, depicted in black), units denoted by lowercase letters are in state '-1' (OFF, depicted in white). cause-effect structures are illustrated as geometrical shapes projected into 3D space (Figure 3.6). Distinctions are depicted as mechanisms (black labels) tying a cause (red labels) and an effect (green labels) through a link (orange edges, thickness indicating φ_d). Relation faces of second- and third-degree relations are depicted as edges or triangular surfaces between the causes and effects of the related distinctions. While edges always bind pairs of distinctions (a second-degree relation), triangular surfaces may bind the causes and effects of two or three distinctions (second- or third-degree relation). Relations of higher degrees are not depicted.

All examples were computed using the "`feature/iit-4.0`" branch of PyPhi (Mayner et al., 2018).

$\varphi_r(d) = \varphi_r(\boldsymbol{d})/|\boldsymbol{d}|$. We can then define the $\mathit{\Phi}_d$ of a distinction $\mathit{\Phi}$-fold $C(\{d\})$ as the sum of all $\varphi_r(d)$ values of each relation in $C(\{d\})$. The $\mathit{\Phi}_d$ values of all distinction $\mathit{\Phi}$-folds with $d \in D$ then sum to the $\mathit{\Phi}$ value of the entire cause-effect structure $C(D)$.

This branch will be available in the next official release of the software. An example notebook is available here which recreates the analysis of Figure 3.1 (identifying complexes), Figure 3.2 (computing distinctions), and Figure 3.4 (computing relations).

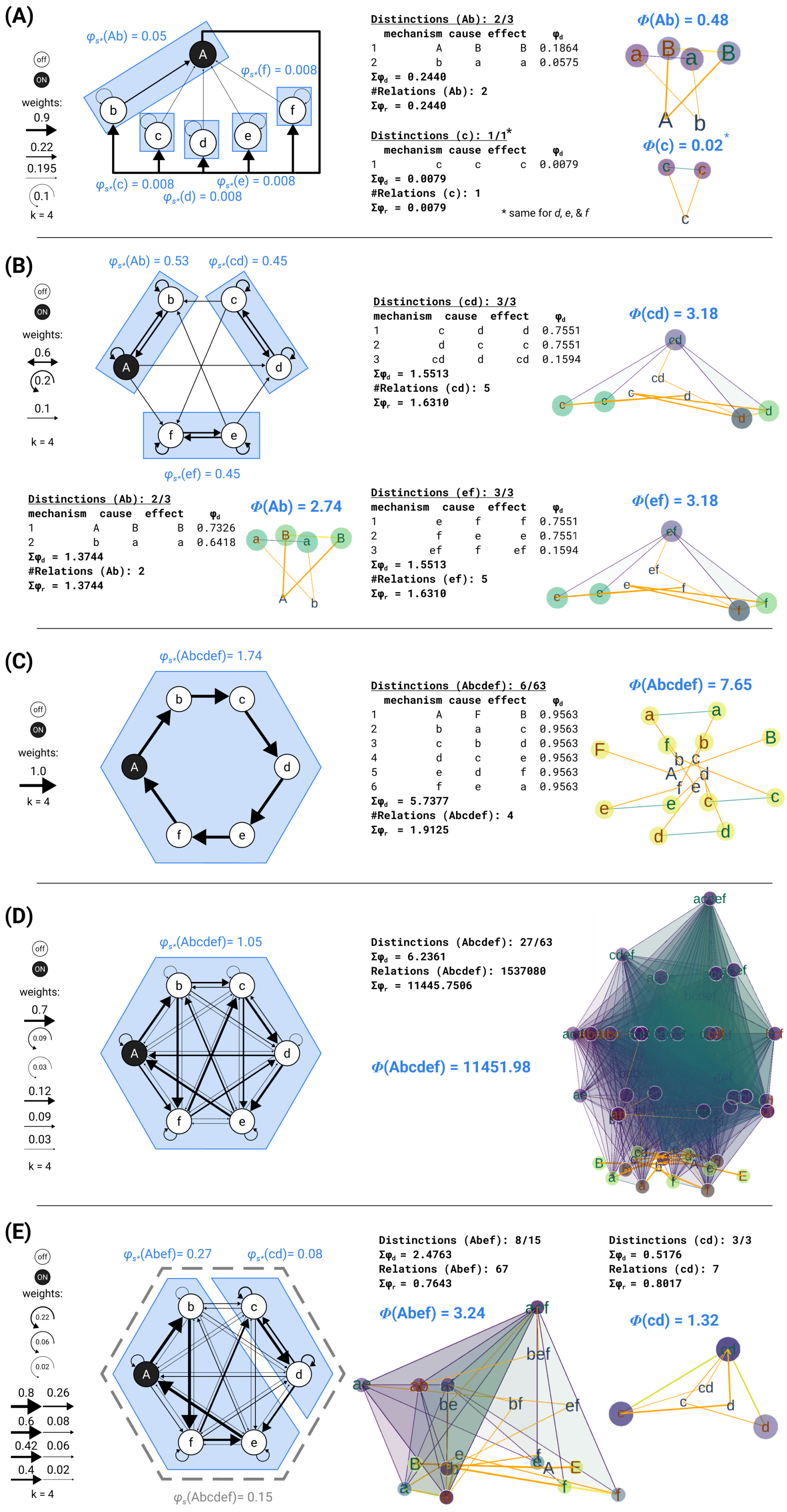
Consciousness and connectivity

The first set of examples highlights how the organization of connections among units impacts the ability of a substrate to support a cause-effect structure with high structure integrated information (high Φ). Figure 3.6 shows five systems, all in the same state $s = Abcdef$ with the same number of units, but with different connectivity among the units.

Degenerate systems, indeterminism, and specificity

Figure 3.6A shows a network with medium indeterminism ($k = 4$) and high degeneracy, due to the fact that unit A forms a “bottleneck” with inputs and outputs to and from the remaining units. The network condenses into one complex of two units Ab and four complexes corresponding to the individual units c, d, e, and f (also called “monads”).

The causes and effects of the causal distinctions for the two types of complexes are shown in the middle, and the corresponding cause-effect structures are illustrated on the right. In this case, degeneracy (coupled with indeterminism) undermines the ability of the maximal substrate to grow in size, which in turn limits the richness of the Φ-structure that can be supported. Because of the bottleneck architecture, the current state of candidate system $Abcdef$ has many possible causes and effects, leading to an exponential decrease in selectivity (the conditional probabilities of cause and effect states). This dilutes the value of intrinsic information (ii) for larger subsets of units, which in turn reduces their value of system integrated information φ_s. Consequently, the maximal substrates are small, and their Φ values are necessarily low.



```
Distinctions (Ab): 2/3
  mechanism cause effect     φd
1          A     B      B  0.1864
2          b     a      a  0.0575
Σφd = 0.2440
#Relations (Ab): 2
Σφr = 0.2440

Distinctions (c): 1/1*
  mechanism cause effect     φd
1          c     c      c  0.0079
Σφd = 0.0079
#Relations (c): 1
Σφr = 0.0079
```

* same for *d*, *e*, & *f*

```
Distinctions (cd): 3/3
mechanism  cause  effect    φd
1            c      d      d  0.7551
2            d      c      c  0.7551
3           cd      d     cd  0.1594
Σφd = 1.5513
#Relations (cd): 5
Σφr = 1.6310
```

```
Distinctions (Ab): 2/3
mechanism  cause  effect    φd
1           A      B      B  0.7326
2           b      a      a  0.6418
Σφd = 1.3744
#Relations (Ab): 2
Σφr = 1.3744
```

```
Distinctions (ef): 3/3
mechanism  cause  effect    φd
1            e      f      f  0.7551
2            f      e      e  0.7551
3           ef      f     ef  0.1594
Σφd = 1.5513
#Relations (ef): 5
Σφr = 1.6310
```

```
Distinctions (Abcdef): 6/63
  mechanism cause effect     φd
1          A     F      B  0.9563
2          b     a      c  0.9563
3          c     b      d  0.9563
4          d     c      e  0.9563
5          e     d      f  0.9563
6          f     e      a  0.9563
Σφd = 5.7377
#Relations (Abcdef): 4
Σφr = 1.9125
```

```
Distinctions (Abcdef): 27/63
Σφd = 6.2361
Relations (Abcdef): 1537080
Σφr = 11445.7506
```

```
Distinctions (Abef): 8/15
Σφd = 2.4763
Relations (Abef): 67
Σφr = 0.7643
```

```
Distinctions (cd): 3/3
Σφd = 0.5176
Relations (cd): 7
Σφr = 0.8017
```

Figure 3.6. Causal powers analysis of various network architectures. Each panel shows the network's causal model and weights on the left. Blue regions indicate complexes with their respective φ_s values. In all networks, $k = 4$ and the state is $Abcdef$. The Φ-structure(s) specified by the network's complexes are illustrated to the right (with only second- and third-degree relation faces depicted) with a list of their distinctions for smaller systems and their $\sum \varphi$ values for those systems with many distinctions and relations. All integrated information values are in ibits. **(A)** A degenerate network in which unit A forms a bottleneck with redundant inputs from and outputs to the remaining units. The first-maximal complex is Ab, which excludes all other subsets with $\varphi_s > 0$ except for the individual units c, d, e, and f. **(B)** The modular network condenses into three complexes along its fault lines (which exclude all subsets and supersets), each with a maximal φ_s value, but low Φ, as the modules each specify only two or three distinctions and at most five relations. **(C)** A directed cycle of six units forms a six-unit complex with $\varphi_s = 1.74$ ibits, as no other subset is integrated. However, the Φ-structure of the directed cycle is composed of only first-order distinctions and few relations. **(D)** A specialized lattice also forms a complex (which excludes all subsets), but specifies 27 first- and high-order distinctions, with many relations ($> 1.5 \times 10^6$) among them. Its Φ value is 11452 ibits. **(E)** A slightly modified version of the specialized lattice in which the first-maximal complex is $Abef$. The full system is not maximally irreducible and is excluded as a complex, despite its positive φ_s value (indicated in gray).

This example suggests that to grow and achieve high values of Φ, substrates must be constituted of units that are specialized (low degeneracy) and interact very effectively (low indeterminism).

Notably, the organization of the cerebral cortex, widely considered as the likely substrate of human consciousness, is characterized by extraordinary specialization of neural units at all levels (Kanwisher, 2010; Khosla & Wehbe, 2022; Ponce et al., 2019). Moreover, if the background conditions are well controlled, neurons are thought to interact in a highly reliable, nearly deterministic manner (Hires et al., 2015; Mainen & Sejnowski, 1995; Nolte et al., 2019).

Modular systems, fault lines, and irreducibility

Figure 3.6B shows a network comprising three weakly interconnected modules, each having two strongly connected units ($k = 4$). In this case, the weak inter-module connections are clear fault lines. Properly normalized, partitions along these fault lines separating modules yield values of φ_s that are much smaller than those yielded by partitions that cut across modules. As a consequence, the 6-unit system condenses into three complexes (Ab, cd, and ef), as determined by their maximal φ_s values. Again, because the modules are small, their Φ values are low. Intriguingly, a brain region such as the cerebellum, whose anatomical organization is highly modular, does not contribute to consciousness (Lemon & Edgley, 2010; Yu et al., 2015), even though it contains several times more neurons than the cerebral cortex (and is indirectly connected to it).

Note that fault lines can be due not just to neuroanatomy but also to neurophysiological factors. For example, during early slow-wave sleep, the dense interconnections among neuronal groups in cerebral cortical areas may break down, becoming causally ineffective due to the bistability of neuronal excitability. This bistability, brought about by neuromodulatory changes (Steriade et al., 1993), is associated with the loss of consciousness (Pigorini et al., 2015).

Directed cycles, structural sparseness, and composition

Figure 3.6C shows a directed cycle in which six units are unidirectionally connected with weight $w = 1.0$ and $k = 4$. Each unit copies the state of the unit before it, and its state is copied by the unit after it, with some indeterminism. The copy cycle constitutes a 6-unit complex with a maximal $\varphi_s = 1.74$ ibits. However, despite the "large" substrate, the Φ-structure it specifies has low structure integrated information ($\Phi = 7.65$). This is because the system's Φ-structure is composed exclusively of first-order distinctions, and consequently of a small number of relations.

Highly deterministic directed cycles can easily be extended to constitute large complexes, being more irreducible than any of their subsets. However, the lack of cross-connections ("chords" in graph-theoretic terms) greatly limits the number of components of the Φ-structures specified by the complexes, and thus their structure integrated information (Φ). (Note also that increasing the number of units that constitute the directed cycle would not change the amount of φ_s specified by the network as a whole.)

The brain is rich in partially segregated, directed cycles, such as those originating in cortical areas, sequentially reaching stations in the basal ganglia and thalamus, and cycling back to cortex (Foster et al., 2021; Middleton & Strick, 2000). These cycles are critical for carrying out many cognitive and other functions, but they do not appear to contribute directly to experience (Tononi et al., 2016).

Specialized lattices and Φ-structures with high structure integrated information

Figure 3.6D shows a network consisting of six heterogeneously connected units—a "specialized" lattice, again with $k = 4$. While many subsystems within the specialized network have positive values of system integrated information φ_s (not shown), the full 6-unit system is the maximal substrate (excluding all its subsets from being maximal substrates). Out of 63 possible distinctions, the Φ-structure comprises 27 distinctions with causes and effects congruent with the system's maximal cause-effect state. Consequently, the full 6-unit system also specifies a much larger number of causal relations compared to the copy cycle system.

Preliminary work (not shown) indicates that lattices of specialized units, implementing different input-output functions, but partially overlapping in their inputs (receptive field) and outputs (projective fields), are particularly well suited to constituting large substrates that unfold into extraordinarily rich Φ-structures. The number of distinctions specified by an optimally connected, specialized system is bounded above by $2^n - 1$, and that of the relations among as many distinctions is bounded by $2^{(2^n-1)} - 1$.The structure integrated information of such structures is correspondingly large (Zaeemzadeh & Tononi, 2023).

In the brain, a large part of the cerebral cortex, especially its posterior regions, is organized as a dense, divergent-convergent hierarchical 3D lattice of specialized units, which makes it a plausible candidate for the substrate of human consciousness (Boly et al., 2017; Haun & Tononi, 2019; Tononi et al., 2016; Watakabe et al., 2023). Note that directed cycles originating and ending in such lattices typically remain excluded from the first-maximal complex because minimal partitions across such cycles yield a much lower value of φ_s compared to minimal partitions across large lattices.

Near-maximal substrates, extrinsic entities, and exclusion

Finally, Figure 3.6E shows a network of six units, four of which ($Abef$) constitute a specialized lattice that corresponds to the first complex. Though integrated, the full set of 6 units happens to be slightly less irreducible ($\varphi_s = 0.15$) than one of its 4-unit subsets ($\varphi_s = 0.27$). From the extrinsic perspective, the 6-unit system undoubtedly behaves as a highly integrated whole (nearly as much as its 4-unit subset), one that could produce complex input-output functions due to its rich internal structure. From the intrinsic perspective of the system, however, only the 4-unit subset satisfies all the postulates of existence, including maximal irreducibility (accounting for the definite nature of experience). In this example, the remaining units form a second complex with low φ_s and serve as background conditions for the first complex.

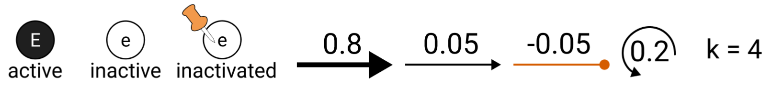
A similar situation may occur in the brain. The brain as a whole is undoubtedly integrated (not to mention that it is integrated with the body as a whole), and neural “traffic” is heavy throughout. However, its anatomical organization may be such that a subset of brain regions, arranged in a dense 3D lattice primarily located in posterior cortex, may achieve a much higher value of integrated information than any other subset. Those regions would then constitute the first complex (the “main complex,” Tononi et al., 2016), and the remaining regions might condense into a large number of much smaller complexes.

Taken together, the examples in Figure 3.6 demonstrate that the connectivity among the units of a system has a strong impact on what set of units can constitute a complex and thereby on the structure integrated information it can specify. The examples also demonstrate the role played by the various requirements that must be satisfied by a substrate of consciousness: existence (causal power), intrinsicality, specificity, maximal irreducibility (integration and exclusion), and composition (structure).

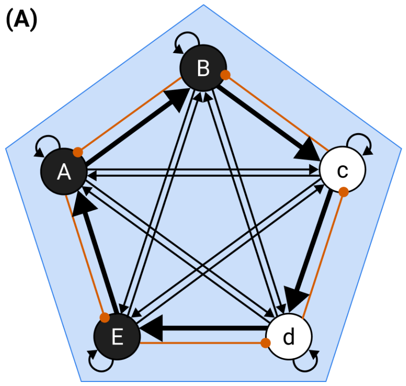
Consciousness and activity: active, inactive, and inactivated units

A substrate exerts cause-effect power in its current state. For the same substrate, changing the state of even one unit may have major consequences on the distinctions and relations that compose its Φ-structure: many may be lost, or gained, and many may change their value of irreducibility (φ_d and φ_r).

Figure 3.7 shows a network of five binary units that interact through excitatory and inhibitory connections (weights indicated in the figure). The system is initially in state $s = ABcdE$ (Figure 3.7A) and is a maximal substrate with $\varphi_s = 1.1$ ibits and a Φ-structure composed of 23 distinctions and their 13740 relations.



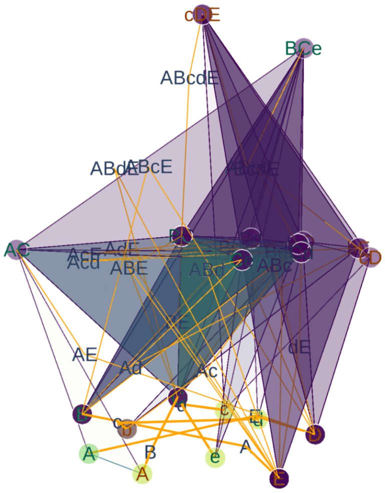
(A)



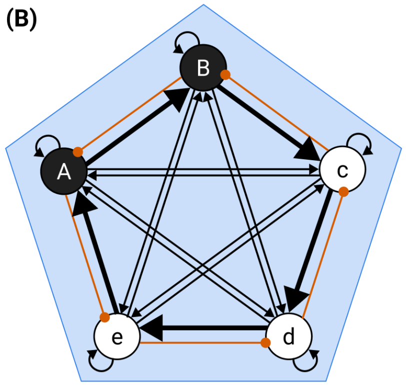
φ_{s*}(ABcdE) = 1.1

```
Distinctions (ABcdE): 23/31
    mechanism cause effect      φd
1           A     E      B  0.783749
2           B     A      C  0.898722
3           c     b      d  0.898722
4           d     c      e  0.898722
5           E     D      A  0.783749
6          Ac     b     Be  0.017030
7          Ad    bE     Ce  0.017030
8          AE     D     AC  0.032622
9          BE     E     BC  0.017030
10         dE     E     Cd  0.011872
11        ABc     D      B  0.009392
12        ABd     E     Ce  0.011014
13        ABE     E     BC  0.018266
14        Acd    bE     Be  0.011014
15        AcE    bD     Bd  0.011014
16        AdE     E      C  0.023537
17        BcE     D      B  0.023537
18        BdE     E      C  0.009392
19       ABcd   cDE    BCe  0.004834
20       ABcE     D      B  0.007660
21       ABdE     E      C  0.007660
22       BcdE    cD     Cd  0.003332
23      ABcdE   cDE     BC  0.002781
#Relations (ABcdE): 13740
```

Φ(ABcdE) = 22.26



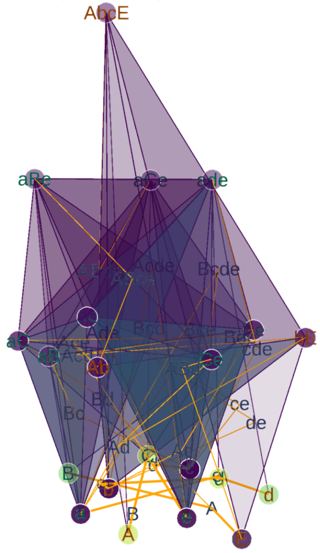
(B)



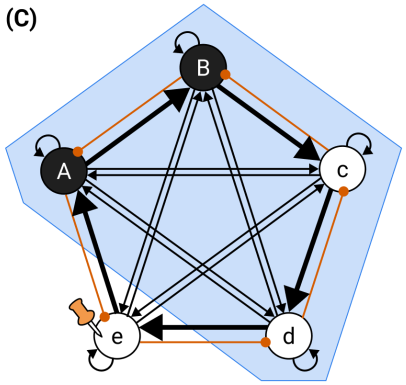
φ_{s*}(ABcde) = 1.31

```
Distinctions (ABcde): 23/31
    mechanism cause effect      φd
1           A     E      B  0.783749
2           B     A      C  0.898722
3           c     b      d  0.898722
4           d     c      e  0.898722
5           e     d      a  0.783749
6          Ac     b     Be  0.017030
7          Ad    bE     Ce  0.017030
8          Bc     c     aB  0.023376
9          Bd     E     aC  0.023376
10         ce     c     ae  0.023376
11         de     b     ae  0.023376
12        ABd     E     Ce  0.011014
13        Acd    bE     Be  0.011014
14        Ace    Ab      e  0.023537
15        Ade     b      e  0.009392
16        Bcd     c      a  0.043528
17        Bce     c      a  0.018799
18        Bde     c     aC  0.004390
19        cde     c    ade  0.013217
20       ABce    Ab    aBe  0.001563
21       ABde  AbcE    aCe  0.002215
22       Acde     b      e  0.007660
23       Bcde     c      a  0.001969
#Relations (ABcde): 13111
```

Φ(ABcde) = 18.55



(C)



φ_{s*}(ABcd) = 0.11

```
Distinctions (ABcd): 14/31
    mechanism cause effect      φd
1           A     D      B  0.885623
2           B     A      C  0.901797
3           c     b      d  0.881767
4           d     c      a  0.124119
5          AB    AD     BC  0.025988
6          Ac     b      B  0.013543
7          Ad     b      C  0.011167
8          Bc    cD     aB  0.011863
9          Bd     c     aC  0.015793
10         cd     b     ad  0.015793
11        ABc   AbD      B  0.011127
12        ABd    bD     aC  0.000758
13        Bcd     c      a  0.001969
14       ABcd    bD    aBC  0.001248
#Relations (ABcd): 363
```

Φ(ABcd) = 3.35

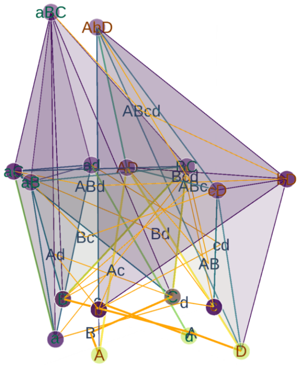


Figure 3.7. Causal powers analysis of the same system with one of its units set to active, inactive, or inactivated. In all panels, the same causal model and weights are shown on the left, but in different states. For all networks $k = 4$. The set of distinctions $D(s)$, their causes and effects, and their φ_d values are shown in the middle. The Φ-structure specified by the network's complex is illustrated on the right (again with only second- and third-degree relation faces depicted). All integrated information values are in ibits. **(A)** The system in state $ABcdE$ is a complex with 23 out of 31 distinctions and $\Phi = 22.26$. **(B)** The same system in state $ABcde$, where unit E is inactive ("OFF") also forms a complex with the same number of distinctions, but a somewhat lower Φ value due to a lower number of relations between distinctions. In addition, the system's Φ-structure differs from that in **(A)**, as the system now specifies a different set of compositional causes and effects. **(C)** If instead of being inactive, unit E is inactivated (fixed into the "OFF" state), the inactivated unit cannot contribute to the complex or Φ-structure anymore. The complex is now constituted of four units ($ABcd$), with only 14 distinctions and markedly reduced structure integrated information ($\Phi = 3.35$).

If we change the state of unit E from ON to OFF (in neural terms, the unit becomes inactive), the distinctions that the unit contributes to when ON, as well as the associated relations, may change (Figure 3.7B). In the case illustrated by the Figure, what changes are the purviews and irreducibility of several distinctions and associated relations, the number of distinctions stays the same, φ_s changes only slightly, but the number of relations is lower, leading to a lower Φ value. In other words, what a single unit contributes to intrinsic existence is not some small "bit" of information. Instead, a unit contributes an entire sub-structure, composed of a very large number of distinctions and relations. The set of distinctions to which a subset of units contributes as a mechanism, either alone or in combination with other units, together with their associated relations, forms a compound Φ-fold. With respect to the neural substrate of consciousness in the brain, this means that even a change in the state of a single unit is typically associated with a change in an entire Φ-fold within the overall Φ-structure, with a corresponding change in the structure of the experience. (Note, however, that in larger systems such changes will typically be less extreme, see also Haun and Tononi, 2019.)

In Figure 3.7C, we see what happens if unit E, instead of just turning inactive (OFF) is *inactivated* (abolishing its cause-effect power because it no longer has any counterfactual states and thus cannot be intervened upon). In this case, all the distinctions and relations to which that unit contributes as a mechanism would cease to exist (its compound Φ-fold collapses). Moreover, all the distinctions and relations to whose purviews that unit contributes—its purview Φ-fold—would also collapse or change. In fact, the complex shrinks because it cannot include that unit. With respect to the neural substrate of consciousness, this means that while an inactive unit contributes to a different experience, an inactivated unit ceases to contribute to experience altogether. The fundamental difference between inactive and inactivated units leads to the following corollary of IIT: unlike a fully inactivated substrate which, as would be suspected, cannot support any experience, an inactive substrate can. If a maximal substrate in an inactive state is in working order and specifies a large Φ-structure, it will support a highly structured experience, such as the experience of empty space (Haun & Tononi, 2019) or the feeling of "pure presence" (Boly, in preparation).

Consciousness and functional equivalence: being is not doing

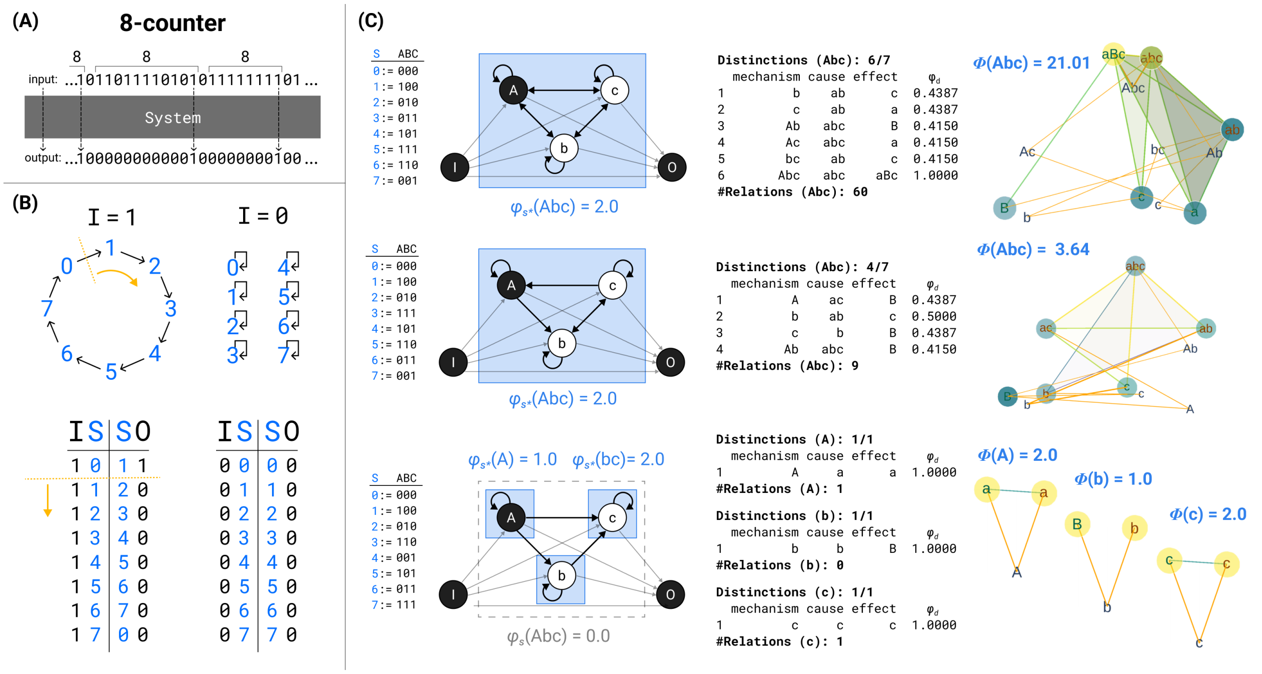
By the intrinsicality postulate, the Φ-structure of a complex depends on the causal interactions between system subsets, not on the system's interaction with its environment (except for the role of the environment in triggering specific system states). In general, different physical systems with different internal causal structure may perform the same input-output functions.

Figure 3.8 shows three simple deterministic systems with binary units (here the "OFF" state is 0, and "ON" is 1) that perform the same input-output function, treating the internal dynamics of the system as a black box. The function could be thought of, for example, as an electronic tollbooth "counting 8 valid coins" (8 times input $I = 1$) before opening the gate (Hanson & Walker, 2021). Each system receives one binary input (I) and has one binary output (O). The output unit switches "ON" on a count of eight positive inputs $I = 1$ (when the global state with label '0' is reached in the cycle), upon which the system resets (Figure 3.8A).

In addition to being functionally equivalent in their outward behavior, the three systems share the same internal global dynamics, as their internal states update according to the same global state-transition diagram (Figure 3.8B). Given an input $I = 1$, the system updates its state, cycling through all its 8 global states (labeled 0–7) over 8 updates. For an input of $I = 0$, the system remains in its present state. Moreover, all three systems are constituted of three binary units whose joint states map one-to-one onto the systems' global state labels (0–7). However, the mapping is different for different systems (Figure 3.8C, left). This is because the internal binary update sequence depends on the interactions among the internal units (Albantakis & Tononi, 2019; Hanson & Walker, 2021), which differ in the three cases, as can easily be determined through manipulations and observations.

For consistency in the causal powers analysis, in all three cases, the global state "0" that activates the output unit if $I = 1$ is selected such that it corresponds to the binary state "all OFF" (000), which is followed by $1 := 100$ and $2 := 010$. Also, the Φ-structure of each system is unfolded in state $1 := 100$ in all three cases.

Despite their functional equivalence and equivalent global dynamics, the systems differ in how they condense into complexes and in the cause-effect structures they specify.



I	S	S	O
1	0	1	1
1	1	2	0
1	2	3	0
1	3	4	0
1	4	5	0
1	5	6	0
1	6	7	0
1	7	0	0

I	S	S	O
0	0	0	0
0	1	1	0
0	2	2	0
0	3	3	0
0	4	4	0
0	5	5	0
0	6	6	0
0	7	7	0

```
S   ABC
0:= 000
1:= 100
2:= 010
3:= 011
4:= 101
5:= 111
6:= 110
7:= 001
```

```
Distinctions (Abc): 6/7
  mechanism cause effect     φd
1         b    ab      c  0.4387
2         c    ab      a  0.4387
3        Ab   abc      B  0.4150
4        Ac   abc      a  0.4150
5        bc    ab      c  0.4150
6       Abc   abc    aBc  1.0000
#Relations (Abc): 60
```

```
S   ABC
0:= 000
1:= 100
2:= 010
3:= 111
4:= 101
5:= 110
6:= 011
7:= 001
```

```
Distinctions (Abc): 4/7
  mechanism cause effect     φd
1         A    ac      B  0.4387
2         b    ab      c  0.5000
3         c     b      B  0.4387
4        Ab   abc      B  0.4150
#Relations (Abc): 9
```

```
S   ABC
0:= 000
1:= 100
2:= 010
3:= 110
4:= 001
5:= 101
6:= 011
7:= 111
```

```
Distinctions (A): 1/1
  mechanism cause effect     φd
1         A     a      a  1.0000
#Relations (A): 1

Distinctions (b): 1/1
  mechanism cause effect     φd
1         b     b      B  1.0000
#Relations (b): 0

Distinctions (c): 1/1
  mechanism cause effect     φd
1         c     c      c  1.0000
#Relations (c): 1
```

Figure 3.8. Functionally equivalent networks with different Φ-structures. (A) The input-output function realized by three different systems (shown in **(C)**): a count of eight instances of input $I = 1$ leads to output $O = 1$. **(B)** The global state-transition diagram is also the same for the three systems: if $I = 0$, the systems will remain in their current global state, labeled as 0-7; if $I = 1$, the systems will move one state forward, cycling through their global states, and activate the output if $S = 0$. **(C)** Three systems constituted of three binary units but differing in how the units are connected and interact. As a consequence, the one-to-one mapping between the 3-bit binary states and the global state labels differ. However, all three systems initially transition from 000 to 100 to 010. Analyzed in state 100, the first system (top) turns out to be a single complex that specifies a Φ-structure with six distinctions and many relations, yielding a high value of Φ. The second system (middle) is also a complex, with the same φ_s value, but it specifies a Φ-structure with fewer distinctions and relations, yielding a lower value of Φ. Finally, the third system (bottom) is reducible ($\varphi_s = 0$) and splits into three smaller complexes (entities) with minimal Φ-structures and low Φ.

As shown in Figure 3.8C, the first system forms a 3-unit complex with a relatively rich Φ-structure ($\Phi = 21.01$ ibits). While the second system also forms a 3-unit complex with the same $\varphi_s = 2$ ibits, it specifies a completely different set of distinctions and has much lower structure integrated information ($\Phi = 3.64$ ibits).

Finally, the third system is reducible ($\varphi_s = 0$ ibits)—in this case, because there are only feed-forward connections from unit A to units B and C—and it condenses into three complexes with small Φ-structures.

These examples illustrate a simple scenario of functional equivalence of three systems characterized by a different architecture. The equivalence is with respect to a simple input-output function, in this case coin counting, which they multiply realize. The systems are also equivalent in terms of their global system dynamics, in the sense that they go through a globally equivalent sequence of internal states. However, because of their different substrates, the three systems specify different cause-effect structures. Therefore, based on the postulates of IIT, they are not phenomenally equivalent. In other words, they are equivalent in what they *do* extrinsically, but not in what they *are* intrinsically.

This dissociation between phenomenal and functional equivalence has important implications. As we have seen, a purely feed-forward system necessarily has $\varphi_s = 0$. Therefore, it cannot support a cause-effect structure and cannot be conscious, whereas systems with a recurrent architecture can. On the other hand, the behavior (input-output function) of any (discrete) recurrent system can also be implemented by a system with a feed-forward architecture (K. Krohn & Rhodes, 1965). This implies that any behavior performed by a conscious system supported by a recurrent architecture can also be performed by an unconscious system, no matter how complex the behavior is. More generally, digital computers implementing programs capable of artificial general intelligence may in principle be able to emulate any function performed by conscious humans and yet, because of the way they are physically organized, they would do so without experiencing anything, or at least anything resembling, in quantity and quality, what each of us experiences (Findlay et al., 2019, in preparation).

The examples also show that the overall system dynamics, while often revealing relevant aspects of a system's architecture, typically do not and cannot exhaust the richness of its current cause-effect structure. For example, a system in a fixed point is dynamically "dead" (and "does" nothing), but it may be phenomenally quite "alive," for example, experiencing "pure presence" (Boly, in preparation). Of course, the system's causal powers can be fully unfolded, and revealed dynamically, by extensive manipulations and observations of subsets of system units because they are implicitly captured by the system's causal model and ultimately by its transition probability matrix (Albantakis & Tononi, 2019).

Conclusions

IIT attempts to account for the presence and quality of consciousness in physical terms. It starts from the existence of experience, and proceeds by characterizing its essential properties—those that are immediate and irrefutably true of every conceivable experience (axioms). These are then formulated as essential properties of physical existence (postulates), the necessary and sufficient conditions that a substrate must satisfy to support an experience—to constitute a complex. Note that "substrate" is meant in purely operational terms—as a set of units that a conscious observer can observe and manipulate. Likewise, "physical" is understood in purely operational terms as cause-effect power—the power to take and make a difference.

The postulates can be assessed based purely on a substrate's transition probability matrix, as was illustrated by a few idealized causal models. Thus, a substrate of consciousness must be able to take and make a difference upon itself (existence and intrinsicality), it must be able to specify a cause and an effect state that are highly informative and selective (information), and it must do so in a way that is both irreducible (integration) and definite (exclusion). Finally, it must specify its cause and effect in a structured manner (composition), where the causal powers of its subsets over its subsets compose a cause-effect structure of distinctions and relations—a Φ-structure. Thus, a complex does not exist as such but only "unfolded" as a Φ-structure—an *intrinsic entity* that exists for itself, absolutely, rather than relative to an external observer.

As shown above, these requirements constrain what substrates can and cannot support consciousness. Substrates that lack in specificity, due to indeterminism and/or degeneracy, cannot grow to be large complexes. Substrates that are weakly integrated, due to architectural or functional fault lines in their interactions, are less integrated than some of their subsets. Because they are not maximally irreducible, they do not qualify as complexes. This is the case even though they may "hang together" well enough from an extrinsic perspective (having a respectable value of φ_s). Furthermore, even substrates that are maximally integrated may support Φ-structures that are extremely sparse, as in the case of directed cycles. Based on the postulates of IIT, a universal substrate ultimately "condenses" into a set of disjoint (non-overlapping) complexes, each constituted of a set of macro or micro units.

The physical account of consciousness provided by IIT should be understood as an explanatory identity: every property of an experience should ultimately be accounted for by a property of the cause-effect structure specified by a substrate that satisfies its postulates, with no additional ingredients. The identity is not between two different substances or realms—the phenomenal and the physical—but between intrinsic (subjective) existence and extrinsic (objective) existence. Intrinsic existence is immediate and irrefutable, while extrinsic existence is defined operationally as cause-effect power discovered through observation and manipulation. The primacy of intrinsic existence (of experience) in IIT contrasts with standard attempts at accounting for consciousness as something "generated by" or "emerging from" a substrate constituted of matter and energy and following physical laws.

The physical correspondent of an experience is not the substrate as such but the Φ-structure specified by the substrate in its current state. Therefore, minor changes in the substrate state can correspond to major changes in the specified Φ-structure. For example, if the state of a single unit changes, an entire Φ-fold within the Φ-structure will change, and if a single inactive unit is inactivated, its associated Φ-fold will collapse, even though the current state of the substrate appears the same (Figure 3.7).

Each experience corresponds to a Φ-structure, not a set of functions. Said otherwise, consciousness is about being, not doing (Albantakis & Tononi, 2015, 2019; Ellia et al., 2021; Tononi, forthcoming). This means that systems with different architectures may be functionally equivalent—both in terms of global input-output functions and global intrinsic dynamics—but they will not be phenomenally equivalent. For example, a feed-forward system can be functionally equivalent to a recurrent system that constitutes a complex, but feed-forward systems cannot constitute complexes because they do not satisfy maximal irreducibility. Accordingly, artificial systems powered by super-intelligent computer programs, but implemented by feed-forward hardware or encompassing critical bottlenecks, would experience nothing (or nearly nothing) because they have the wrong kind of physical architecture, even though they may be behaviorally indistinguishable from human beings (Findlay et al., 2019, in preparation).

Even though the entire framework of IIT is based on just a few axioms and postulates, it is not possible in practice to exhaustively apply the postulates to unfold the cause-effect power of realistic systems (A. B. Barrett & Mediano, 2019; Moyal et al., 2020). It is not feasible to perform all possible observations and manipulations to fully characterize a universal TPM, or to perform all calculations on the TPM that would be necessary to condense it exhaustively into complexes and unfold their cause-effect power in full. The number of possible systems, of system partitions, of candidate distinctions—each with their partitions and relations—is the result of multiple, nested combinatorial explosions. Moreover, these observations, manipulations, and calculations would need to be repeated at many different grains, with many rounds of maximizations. For these reasons, a full analysis of complexes and their cause-effect structure can only be performed on idealized systems of a few units (Mayner et al., 2018; Chapter 2, § 2.3).

On the other hand, we can simplify the computation considerably by using various assumptions and approximations, as with the "cut one" approximation described in Mayner et al. (2018). Also, while the number of relations vastly exceeds the number of units and of distinctions (its upper bound for a system of n units is $2^{(2^n-1)}-1$), it can be determined analytically, and so can $\sum \varphi_r$ for a given set of distinctions Appendix B.3. Developing tight approximations, as well as bounded estimates of a system's integrated information $\Phi(s)$, is one of the main areas of ongoing research related to IIT (Zaeemzadeh & Tononi, 2023).

Despite the infeasibility of an exhaustive calculation of the relevant quantities and structures for a realistic system, IIT already provides considerable explanatory and predictive power in many real-world situations, making it eminently testable (Cogitate Consortium et al., 2023; Melloni et al., 2021; Sarasso et al., 2021; Tononi et al., 2016). A fundamental prediction is that Φ should be high in conscious states, such as wakefulness and dreaming, and low in unconscious states, such as dreamless sleep and anesthesia. This prediction has already found substantial support in human studies that have applied measures of complexity inspired by IIT to successfully classify subjects as conscious vs. unconscious (Casarotto et al., 2016; Massimini et al., 2005; Sarasso et al., 2020; Tononi et al., 2016). IIT can also account mechanistically for the loss of consciousness in deep sleep and anesthesia (Pigorini et al., 2015; Tononi et al., 2016). Furthermore, it can provide a principled account of why certain portions of the brain may constitute an ideal substrate of consciousness and others may not, why the borders of the main complex in the brain should be where they are, and why the units of the complex should have a particular grain (the one that yields a maximum of φ_s). A stringent prediction is that the location of the main complex, as determined by the overall maximum of φ_s within the brain, should correspond to its location as determined through clinical and experimental evidence. Another prediction that follows from first principles is that constituents of the main complex can support conscious contents even if they are mostly inactive, but not if they are inactivated (Haun & Tononi, 2019; Tononi et al., 2016). Yet another prediction is that the complete inactivation of constituents of the main complex should lead to absolute agnosia (unawareness that anything is missing).

IIT further predicts that the quality of experience should be accounted for by the way the Φ-structure is composed, which in turn depends on the architecture of the substrate specifying it. This was demonstrated in a recent paper showing how the fundamental properties of spatial experiences—those that make space feel "extended"—can be accounted for by those of Φ-structures specified by 2D grids of units, such as those found in much of posterior cortex (Haun & Tononi, 2019). This prediction is in line with neurological evidence of their role in supporting the experience of space (Haun & Tononi, 2019). Ongoing work aims at accounting for the quality of experienced time (Comolatti &

Grasso, in preparation) and that of experienced objects (Grasso, in preparation). A related prediction is that changes in the strength of connections within the neural substrate of consciousness should be associated with changes in experience, even if neural activity does not change (Song et al., 2017). Also, similarities and dissimilarities in the structure of experience should be accounted for by similarities and dissimilarities among Φ-structures and Φ-folds specified by the neural substrate of consciousness.

While the listed predictions may appear largely qualitative in nature, many of them rest on specific features of the accompanying quantitative analysis. This is the case for predictions regarding the borders (and grain) of the main complex in the brain, which depend on the relative φ_s values of potential substrates of interest, and even more so for predictions regarding the quality and richness of certain experiences and the predicted features of their underlying substrates. IIT's postulates, and the mathematical framework proposed to evaluate them, rest on "inferences to a good explanation" (Box 1). While we have aimed for maximal consistency, specificity, and simplicity at every junction in formulating IIT's mathematical implementation, some of the algorithmic choices remain open to further evaluation. These include, for example, the proper treatment of background conditions and the resolution of ties given symmetries in the TPMs of specific systems (Appendix B.1). More generally, further validation of IIT will depend on a systematic back-and-forth between phenomenology, theoretical inferences, and neuroscientific evidence (Ellia et al., 2021).

In addition to empirical work aimed at validating the theory, much remains to be done at the theoretical level. According to IIT, the meaning of an experience is its feeling—whether those of spatial extendedness, of temporal flow, or of objects, to name but a few ("the meaning is the feeling"). This means that every meaning is identical to a sub-structure within a current Φ-structure—a content of experience—whether it is triggered by extrinsic inputs or it occurs spontaneously during a dream. Therefore, all meaning is ultimately intrinsic. Ongoing work aims at providing a self-consistent explanation of how intrinsic meanings can capture relevant features of causal processes in the environment (see Chapter 5). It will also be important to explain how intersubjectively validated knowledge can be obtained despite the intrinsic and partially idiosyncratic nature of meaning.

To the extent that the theory is validated through empirical evidence obtained from the human brain, IIT can then offer a plausible inferential basis for addressing several questions that depend on an explicit theory of consciousness. As indicated in the section on phenomenal and functional equivalence, and argued in ongoing work (Findlay et al., 2019, in preparation), one consequence of IIT is that typical computer architectures are not suitable for supporting consciousness, no matter whether their behavior may resemble ours. By the same token, it can be inferred from IIT that animal species that may look and behave quite differently from us may be highly conscious, as long as their brains have a compatible architecture. Other inferences concern our own experience and whether it plays a causal role, or is simply "along for the ride" while our brain performs its functions. As recently argued, IIT implies that we have true free will—that we have true alternatives, make true decisions, and truly cause. Because only what truly exists (intrinsically, for itself) can truly cause, we, rather than our neurons, cause our willed actions and are responsible for their consequences (Tononi, Albantakis, et al., 2022).

Finally, an ontology that is grounded in experience as intrinsic existence—an intrinsic ontology—must not only provide an account of subjective existence in objective, operational terms, but also offer a path toward a unified view of nature—of all that exists and happens. One step in this direction is the application of the same postulates that define causal powers (existence) to the evaluation of actual causes and effects ("what caused what"; Albantakis et al., 2019). Another is to unify classical accounts of information (as communication and storage of signals) with IIT's notion of information as derived from the properties of experience—that is, information as causal, intrinsic, specific, maximally irreducible, and structured (meaningful) (Oizumi et al., 2014; Zaeemzadeh & Tononi, 2023). Yet another is the study of the evolution of a substrate's causal powers as conditional probabilities that update themselves (Albantakis et al., 2014).

Even so, there are many ways in which IIT may turn out to be inadequate or wrong. Are some of its assumptions, including those of a discrete, finite set of "atomic" units of cause-effect power, incompatible with current physics (A. B. Barrett and Mediano, 2019; Carroll, 2021; but see Esteban et al., 2018; Kalita et al., 2019; Kleiner and Tull, 2021; Zanardi et al., 2018)? Are its axiomatic basis and the translation of axioms into postulates sound and unique? And, most critically, can IIT survive the results of empirical investigations assessing the relationship between the quantity and quality of consciousness and its substrate in the brain?

CHAPTER 4

Measuring stimulus-evoked neurophysiological differentiation in distinct populations of neurons in mouse visual cortex

William G. P. Mayner[1,2], William Marshall[2,3], Yazan N. Billeh[4], Saurabh R. Gandhi[4], Shiella Caldejon[4], Andrew Cho[4], Fiona Griffin[4], Nicole Hancock[4], Sophie Lambert[4], Eric K. Lee[4], Jennifer A. Luviano[4], Kyla Mace[4], Chelsea Nayan[4], Thuyanh V. Nguyen[4], Kat North[4], Sam Seid[4], Ali Williford[4], Chiara Cirelli[2], Peter A. Groblewski[4], Jerome Lecoq[4], Giulio Tononi[2], Christof Koch[4], Anton Arkhipov[4]

1 Neuroscience Training Program, University of Wisconsin–Madison, Madison, WI, 53705, United States of America

2 Department of Psychiatry, University of Wisconsin–Madison, Madison, WI, 53719, United States of America

3 Department of Mathematics and Statistics, Brock University, St. Catharines, ON, L2S 3A1, Canada

4 Allen Institute, Seattle, WA, 98109, United States of America

This chapter has been previously published as:

> Mayner, W. G. P., Marshall, W., Billeh, Y. N., Gandhi, S. R., Caldejon, S., Cho, A., Griffin, F., Hancock, N., Lambert, S., Lee, E. K., Luviano, J. A., Mace, K., Nayan, C., Nguyen, T. V., North, K., Seid, S., Williford, A., Cirelli, C., Groblewski, P. A., … Arkhipov, A. (2022). Measuring Stimulus-Evoked Neurophysiological Differentiation in Distinct Populations of Neurons in Mouse Visual Cortex. *eNeuro*, *9*(1). `https://doi.org/10.1523/ENEURO.0280-21.2021`

Acknowledgments

We thank the Allen Institute founder, Paul G. Allen, for his vision, encouragement, and support; Allan Jones for providing the crucial environment that enabled our large-scale team effort; Andrew Haun for helpful comments and assistance with creating the noise stimulus; Jonathan Lang for assistance filming the naturalistic stimuli; and Larissa Albantakis, Leonardo S. Barbosa, Tom Bugnon, Graham Findlay, Marcello Massimini, and Shuntaro Sasai for helpful discussions.

The data presented herein were obtained at the Allen Brain Observatory as part of the *OpenScope* project, which is operated by the Allen Institute. This work was supported by the Allen Institute, the Tiny Blue Dot Foundation, and in part by the Falconwood Foundation. W.M. received support from the Natural Sciences and Engineering Research Council of Canada (RGPIN-2019-05418).

4.1 Introduction

The visual system acts on incoming stimuli to extract meaningful features and guide behavior, a process that transforms physical input into conscious percepts. Since the early experiments of Hubel and Wiesel (1959), neuroscience has yielded considerable insight into the visual system by analyzing neural response properties to uncover which features cells are tuned to and how their activity relates to behavior. Modern decoding approaches have revealed stimulus information in population responses (Quiroga & Panzeri, 2009). However, that a population of neurons represents stimulus information does not imply that this information is used to generate conscious percepts (Brette, 2019). Consequently, despite the success of these "outside-in" methods (Buzsáki, 2019) in understanding neural coding, it remains unclear how the coordinated activity of large neuronal populations relates to what the observer actually sees.

Is there an objective approach that can shed light on this question? *Differentiation analysis*—measuring the extent to which a population of neurons expresses a rich and varied repertoire of states—has been proposed as one such approach (Boly et al., 2015; Mensen et al., 2017, 2018). Differentiation analysis exemplifies "inside-out" methodology in that the spatiotemporal diversity of neural activity (*neurophysiological differentiation* or ND) is quantified without reference to the stimulus or other experimental variables imposed *a priori* by the investigator, in contrast to feature tuning or decoding analyses.

A visual stimulus can be considered meaningful to the observer if it evokes rich and varied perceptual experiences (*phenomenological differentiation*). For example, an engaging movie is meaningful in this sense, as it evokes many distinct percepts with high-level structure; conversely, flickering 'TV noise' essentially evokes a single percept with no high-level structure to a human observer, even though, at the level of pixels, any two frames of noise are likely to be more different from each other than a pair of frames from a movie (*stimulus differentiation*). Since conscious percepts are determined by brain states, physical differences must underlie phenomenological differentiation. Thus, one can expect measures of ND to correlate with subjective perception of the "richness" or "meaningfulness" of stimuli to the extent that such measures capture the relevant physical, *i.e.*, neuronal, differences. This has indeed been shown in human studies using fMRI and EEG (Boly et al., 2015; Mensen et al., 2017, 2018). Moreover, integrated information theory (IIT) posits a fundamental relationship between ND and subjective experience itself (Marshall et al., 2016; Oizumi et al., 2014; Tononi, 2004; Tononi et al., 2016), and several studies showed that loss of ND is implicated in loss of consciousness (Barttfeld et al., 2015; Casali et al., 2013; Hudetz et al., 2014; Wenzel et al., 2019).

Although studies in human subjects suggest that ND can provide a readout of stimulus-evoked phenomenological differentiation (Boly et al., 2015; Mensen et al., 2017, 2018), the low spatial resolution of fMRI and EEG has precluded identifying the cell populations that underlie this correspondence. A longstanding fundamental question is which neuronal populations contribute directly to generating conscious percepts (Koch et al., 2016; Mashour et al., 2020; Tononi et al., 2016). Differentiation analysis may shed light on this question, but to do so it must be applied to signals from specific populations of neurons.

To address this gap, we used *in vivo* two-photon calcium imaging in mice to measure ND evoked by naturalistic and phase-scrambled movie stimuli in excitatory cell populations in cortical layers (L) 2/3, L4, and L5, across five visual cortical areas: primary (V1), lateromedial (LM), anterolateral (AL), posteromedial (PM), and anteromedial (AM). We hypothesized that unscrambled naturalistic stimuli, which presumably elicit meaningful visual percepts, would evoke greater ND than their meaningless phase-scrambled counterparts.

We find that unscrambled stimuli evoke greater ND than scrambled stimuli specifically in L2/3 of areas AL & AM and not in the other neuronal populations. We contrast this layer- and area-specific finding with a decoding analysis that shows that information about the stimulus category, whether meaningful or meaningless, is present in most populations. This highlights a key difference: ND is more plausibly correlated with stimulus meaningfulness than the information measured by decoding, since the latter may not be functionally relevant (Brette, 2019). Furthermore, we find differences in evoked ND among the unscrambled stimuli that suggest that differentiation analysis can probe meaningfulness of individual stimuli.

4.2 Materials and methods

Our experimental design is summarized in Figure 4.1. We collected calcium imaging data from L2/3, L4, and L5 in each of 5 visual areas (V1, LM, AL, PM, and AM) across 45 experimental sessions (9 mice, 3 per layer; L2/3, 2 males; L4, 3 males; L5, 1 male; 15 sessions per transgenic line; 9 sessions per area; 5 ± 1 sessions per mouse; number of cells shown in Table C.1). Areas V1, LM, AL, PM, and AM respectively correspond to areas VISp, VISl, VISal, VISpm, and VISam in the Mouse Brain Common Coordinate Framework v3 (Wang et al., 2020). The two-photon calcium imaging pipeline is described in detail in de Vries et al. (2020) and Groblewski et al. (2020). All animal procedures were performed in accordance with the Allen Institute animal care committee's regulations.

Our sample size was selected on the basis of a pilot study using existing, publicly available calcium imaging data from the Allen Institute Brain Observatory (de Vries et al., 2020). We measured spectral differentiation of responses to a movie stimulus (clips from the film "Touch of Evil") versus artificial stimuli (drifting gratings and locally sparse noise) across 8 experimental sessions, from which we estimated that we required at least 3 sessions per layer/area pair to have statistical power of at least 0.8.

Transgenic mice

We maintained all mice on reverse 12-hour light cycle following surgery and throughout the duration of the experiment and performed all experiments during the dark cycle. We used the transgenic mouse line *Ai93*, in which GCaMP6f expression is dependent on the activity of both Cre recombinase and the tetracycline controlled transactivator protein (tTA) (Madisen et al., 2010). Triple transgenic mice (*Ai93, tTA, Cre*) were generated by first crossing *Ai93* mice with *Camk2a-tTA* mice, which preferentially express tTA in forebrain excitatory neurons.

Cux2-CreERT2;Camk2a-tTA;Ai93(TITL-GCaMP6f) expression is regulated by the tamoxifen-inducible *Cux2* promoter, induction of which results in Cre-mediated expression of GCaMP6f predominantly in superficial cortical L2/3 and L4. *Rorb-IRES2-Cre;Cam2a-tTA;Ai93* exhibit GCaMP6f in excitatory neurons in cortical L4 (dense patches) and L5 & L6 (sparse). Rbp4-Cre;Camk2a-tTA;Ai93 exhibit GCaMP6f in excitatory neurons in cortical L5. Calcium indicator kinetics did not differ between cell populations (Figure C.2).

Surgery

Transgenic mice expressing GCaMP6f were weaned and genotyped at ~P21, and surgery was performed between P37 and P63. The craniotomy was centered at $X = -2.8$mm and $Y = 1.3$mm with respect to lambda (centered over the left mouse visual cortex). A circular piece of skull 5 mm in diameter was removed, and a durotomy was performed. A coverslip stack (two 5 mm and one 7 mm glass coverslip adhered together) was cemented in place with Vetbond. Metabond cement was applied around the cranial window inside the well to secure the glass window.

Intrinsic imaging

To define area boundaries and target *in vivo* two-photon calcium imaging experiments to consistent retinotopic locations, retinotopic maps for each animal were created using intrinsic signal imaging (ISI) while mice were lightly anesthetized with 1.0–1.4% isoflurane. This procedure and data processing pipeline are described in detail in de Vries et al. (2020).

Habituation

Following successful ISI mapping, mice spent two weeks being habituated to head fixation and visual stimulation. During the second week, mice were head-fixed and presented with visual stimuli, starting with 10 minutes and progressing to 50 minutes of visual stimuli by the end of the week. During this week they were exposed to the "mouse montage 2" stimulus (§ 4.2, Stimuli).

Imaging

Calcium imaging was performed using a two-photon-imaging instrument (Nikon A1R MP+). Laser excitation was provided by a Ti:Sapphire laser (Chameleon Vision – Coherent) at 910 nm. Mice were head-fixed on top of a rotating disc and free to run at will. The screen center was positioned 118.6 mm lateral, 86.2 mm anterior and 31.6 mm dorsal to the right eye. The distance between the screen and the eye was 15 cm. Movies were recorded at 30 Hz using resonant scanners over a 400 μm field of view.

Locomotion

Locomotion velocity was recorded from the running wheel and preprocessed as follows. First, artifacts were removed using custom code that iteratively identified large positive or negative peaks (indicative of artifactual discontinuities in the signal) in several passes of `scipy.signal.find_peaks()` (specific parameters were manually chosen for each session). Remaining artifacts were then manually removed by inspecting the resulting timeseries and visually identifying clear discontinuities. The removed samples were filled using linear interpolation (`pandas.Series.interpolate`).

The resulting signal was then low-pass filtered at 1 Hz using a zero-phase 4th-order Butterworth filter (`scipy.signal.butter(2, 1/15, btype='lowpass', output='ba', analog=False)` applied with `scipy.signal.filtfilt`).

For the effect size analysis, the fraction of time spent running was calculated by binarizing the preprocessed velocity timeseries at a threshold of 2.5 cm/s.

Pupillometry

Pupil diameter was extracted from video of the mouse' s ipsilateral eye (relative to the stimulus presentation monitor) using the AllenSDK (`https://github.com/AllenInstitute/AllenSDK`) as described in de Vries et al. (2020).

Briefly, for each frame of the video an ellipse was fitted to the region corresponding to the pupil as follows: a seed point within the pupil was identified via convolution with a black square; 18 rays were drawn starting at this seed point, spaced 20 degrees apart; the candidate boundary point between the pupil and iris along that ray was identified by a change in pixel intensity above a session-specific threshold; a RANSAC algorithm was used to fit the an ellipse to the candidate boundary points using linear regression with a conic section constraint; and fitted parameters of the regression were converted to ellipse parameters (coordinates of the center, lengths of the semi-major and semi-minor axes, and angle of rotation with respect to the x-axis). Pupil diameter was taken to be twice the semi-major axis of the fitted ellipse.

The resulting timeseries contained some artifacts, which we removed by the same combination of automated and manual methods used for the locomotion timeseries (§ 4.2, Locomotion). Each pupil diameter timeseries was then normalized by dividing by the maximum diameter that occurred within the 10 blocks of stimulus presentations during that session.

Event detection

Discrete calcium events were detected from the $\Delta F/F_0$ traces using the L_0-penalized method of S. Jewell and Witten (2018) and S. W. Jewell et al. (2020). This procedure, which replaces the continuous relative changes in fluorescence with discrete, real valued events, is described in detail in de Vries et al. (2020); code is available at `https://github.com/AllenInstitute/visual_coding_2p_analysis/blob/`

`master/visual_coding_2p_analysis/l0_analysis.py`.

Stimuli

We created twelve 30 s greyscale naturalistic and artificial movie stimuli.

The eight naturalistic stimuli (Figure C.1, top) consisted of three montages of six 5 s clips, spliced together with jump cuts, and four continuous stimuli. The "mouse montage 1" stimulus contained clips of conspecifics, a snake, movement at ground level through the underbrush of a wooded environment, and a cat approaching the camera. The "mouse montage 2" stimulus contained different footage of movement through the wooded environment; different footage of a cat approaching the camera; conspecifics in a home cage filmed from within the cage; crickets in a home cage filmed from within the cage; footage of the interior of the home cage with environmental enrichment (a shelter, running wheel, and nesting material); and a snake filmed at close range orienting towards the camera. The "human montage" contained clips of a man talking animatedly to an off-screen interviewer; a café table where food is being served; automobile traffic on a road viewed from above; a woman in the foreground taking a photo of a city skyline; footage of a road filmed from the passenger seat of a vehicle; and a close shot of a bowl of fruit being tossed. The four continuous stimuli were: footage of a snake at close range orienting towards the camera; crickets in a home cage filmed from within the cage; a man writing at a table; movement through a wooded environment at ground level; and conspecifics in a home cage. No two stimuli contained identical clips.

The four artificial stimuli (Figure C.1, bottom) consisted of two phase-scrambled versions of the "mouse montage 1" stimulus, a phase-scrambled version of the "mousecam" stimulus (§ 4.2, Phase scrambling), and a high-pass-filtered $1/f$ noise stimulus.

Stimuli were presented in a randomized block design with 10 repetitions, with 4 s of static mean-luminance grey presented between stimuli (Figure 4.1F). 60 s of mean-luminance grey (to record spontaneous activity) and a 60 s high-contrast sparse noise stimulus were also presented in the beginning of each session (this stimulus was not analyzed in this work).

Phase scrambling

Two methods of phase scrambling were used: temporal and spatial, described in detail below. Briefly, for the temporal scrambling we independently randomized the phase of each pixel's intensity timeseries in contiguous, nonoverlapping windows of 1 s. For the spatial scrambling, we randomized the phase of the spatial dimensions of the three-dimensional spectrum of each window. The "mouse montage 1" stimulus was phase-scrambled using both procedures to obtain the "mouse montage 1, temporal phase scramble" and "mouse montage 1, spatial phase scramble" stimuli. The "mousecam" stimulus was scrambled using the spatial procedure to obtain the "mousecam, spatial phase scramble" stimulus.

Temporal phase scramble

First, the stimuli were windowed into contiguous, nonoverlapping 1 s segments (30 frames each). For each 1 s window, we applied the following procedure:

We estimated the one-dimensional spectrum of each pixel's intensity timeseries with the discrete Fourier transform (DFT) using the NumPy function `numpy.fft.fft`. The phase and magnitude of each spectrum were computed with `numpy.angle` and `numpy.abs` respectively. For each pixel, we generated a 14-element random vector drawn uniformly from the interval $[0, 2\pi]$. A randomized phase was then obtained for that pixel by concatenating the first element of the original phase, the random vector, the 15th element of the original phase, and the negative reversed random vector. This yielded a 30-element phase vector with the required conjugate symmetry of the spectrum of a 1 s real-valued signal sampled at 30 frames per second. The randomized phase was then combined with the spectral magnitude and transformed back into the time domain with the inverse DFT using `numpy.fft.ifft`, yielding a temporally phase-scrambled version of that pixel's intensity timeseries. Each pixel's timeseries was independently phase-scrambled in this fashion.

This resulted in 30 independently phase-scrambled 1 s windows. These windows were then concatenated to obtain the full 30 s temporally phase-scrambled stimulus.

Spatial phase scramble

First, the stimuli were windowed into contiguous, nonoverlapping 1 s segments (30 frames each). For each window, we applied the following procedure. The three-dimensional Fourier spectrum (frame, width, and height) was estimated with the DFT using `numpy.fft.fftn`. The phase and magnitude of the spectrum were computed with `numpy.angle` and `numpy.abs` respectively. To randomize the phase in the spatial dimensions, we generated a random signal in the time domain with the same dimensions as a stimulus frame (192 pixels wide by 120 pixels high) and computed its phase in the frequency domain as described above. This two-dimensional random spatial phase was added to the spatial dimensions of the three-dimensional stimulus phase. After being randomized in this way, the stimulus phase was recombined with the spectral magnitude and transformed back into a time-domain signal with the inverse DFT using `numpy.fft.ifftn`. The 30 resulting phase-scrambled 1 s windows were then concatenated to obtain the full 30 s spatially phase-scrambled stimulus.

Effect of phase-scrambling

The greyscale movie stimuli were represented in the stimulus presentation software as arrays of unsigned 8-bit integers. The limitations of this representation resulted in phase-scrambled stimuli with power spectra that were close but not identical to the power spectrum of their unscrambled counterparts.

Specifically, although the phase scrambling procedures described above leave the power spectrum unchanged, they do not necessarily preserve the range of the resulting real-valued signal. In our case, applying these procedures to our stimuli resulted in phase-scrambled stimuli in which the pixel intensities occasionally lay outside the range $[0, 255]$. Thus, in order to represent the phase-scrambled stimuli with 8-bit integers, we truncated the result so that negative intensities were set to 0 and intensities greater than 255 were set to 255. This operation does affect the power spectra, and as a result the spectra of the unscrambled and scrambled stimuli are closely matched but not equal.

Differentiation analysis

Spectral differentiation

Our analysis of the responses to the stimuli follows the techniques developed in previous work in humans (Boly et al., 2015; Mensen et al., 2017, 2018). The spectral differentiation measure of ND used by Mensen et al. (2018) was designed for analysis of timeseries responses to continuous movie stimuli, and was found to be positively correlated with subjective reports of stimulus “meaningfulness”. We employed this measure with our calcium imaging data on single-trial responses: (A) for each cell, the

$\Delta F/F_0$ trace of each cell during stimulus presentation was divided into 1 s windows; (B) the power spectrum of each window was estimated using a Fourier transform; (C) the "neurophysiological state" during each 1 s window was defined as a vector in the high-dimensional space of cells and frequencies (*i.e.*, the concatenation of the power spectra in that window for each cell); (D) the ND in response to a given stimulus was calculated as the median of the pairwise Euclidean distances between every state that occurred during the stimulus presentation. A schematic illustration is shown in Figure 4.2, and an illustration of how the measure behaves for different types of signals is shown in Figure C.3.

The relationship between action potentials and the resulting calcium imaging signal is complex (Chen et al., 2013; Deneux et al., 2016; Huang et al., 2021; Ledochowitsch et al., 2019; Pachitariu et al., 2018; Siegle, Ledochowitsch, et al., 2021; Wei et al., 2020). The $\Delta F/F_0$ signal approximately represents a convolution of the underlying spike train with the calcium-dependent fluorescence response kernel, which depends nonlinearly on the spike rate. A consequence of the nonlinearity is that calcium imaging is much more sensitive to burst-like activity than to isolated spikes. Since this convolution affects the spectral properties of the signal, some discussion of its impact on the spectral differentiation measure is warranted. The energy of the GCaMP6f response is concentrated in low frequencies, so for our purposes, the effect of the convolution is that differences among the spectral states are amplified at lower frequencies and attenuated at higher frequencies. Thus, when applied to the $\Delta F/F_0$ signal, the measure will be less sensitive to short-timescale differences between activity patterns than it would if it were applied directly to the ground-truth spike train. This is not necessarily a disadvantage, as our aim in using a spectral measure in the first place was to achieve temporal smoothing to detect differences in temporal structure on the scale of the state window size.

We normalized spectral differentiation values by the square root of the number of cells in the recorded population, reasoning as follows. Consider a hypothetical population of cells that each exhibit the same temporal pattern of activity. The spectral differentiation of such a population will be proportional to the square root of its size because the Euclidean distance is used to compare neurophysiological states. If we have two such populations differing only in the number of cells, their activity should be considered equally differentiated for our purposes, since their temporal patterns are identical; any differences in spectral differentiation would be due to the (arbitrary) number of cells captured in the imaging session. Thus, we divided by the square root of the population size to remove this dependency.

To investigate the properties of the signal that drive differences in spectral differentiation, we applied the measure to discrete L_0 calcium events detected from the $\Delta F/F_0$ traces (§ 4.2, Event detection) and obtained similar results as in the main analysis (Figure C.11). This indicates that the observed differences in spectral differentiation are driven by differences in the large-timescale patterns of responses rather than small-timescale spectral differences within the windows, consistent with the sparsity of calcium responses in this dataset. We also measured ND of $\Delta F/F_0$ traces with transients removed. Transients were defined as the 200 ms (6 imaging samples) following a L_0 calcium event. This analysis yielded similar results as well (Figure 3-9), indicating that ND differences are not driven solely by initial transients in the calcium response.

For the analysis of stimulus differentiation (Figure 4.8), stimuli were first blurred with a circular Gaussian filter whose half width at half maximum was set to the median radius of a L2/3 V1 receptive field as measured by de Vries et al. (2020) (8.92 degrees) to account for the coarseness of mouse vision. Stimulus differentiation was then calculated by treating each pixel of the stimulus as a "cell" and applying the spectral differentiation measure to the traces of pixel intensities over time.

Multivariate differentiation

We also measured ND using a multivariate approach that considers spatiotemporal differences in activity patterns. For each experimental session, we selected $\Delta F/F_0$ traces recorded during presentations of unscrambled stimuli and their scrambled counterparts and concatenated them to obtain an $m \times n$ matrix of responses, where m is the number of two-photon imaging samples and n is the number of traces. We used a nonlinear dimensionality reduction procedure, Uniform Manifold Approximation and Projection for Dimension Reduction (Python package `umap-learn`, McInnes et al., 2020), to reduce this matrix to $m \times 8$ with parameters `UMAP(n_components=8, metric="euclidean", n_neighbors=50, min_dist=0.5)`. Each row of the resulting matrix was an 8-dimensional vector that represented the state of the cell population during the corresponding two-photon sample. We then grouped the rows of the resulting matrix by stimulus presentation. Each row vector can be thought of as a point in $\mathbb{R}^8$, so that each trial was associated with a cloud of points corresponding to the population states that the stimulus evoked during that presentation.

The intuition motivating this approach is that we can operationalize the notion of neurophysiological differentiation by measuring the dispersion of this point cloud. The more distant two points are, the more different are the corresponding responses of the cell population; thus, if a stimulus evokes many different population states, the point cloud will be more spread out in response space. Therefore, we measured ND evoked during each stimulus presentation by finding the centroid of the associated point cloud and taking the mean Euclidean distance of each point to the centroid.

In the multivariate differentiation analysis of calcium events, population response vectors were obtained by summing event magnitudes within 1 s bins to match the definition of the neurophysiological state of the population used in the spectral differentiation analyses. Because the event data were sparse and because including many duplicate instances of the zero vector will reduce the sensitivity of the multivariate differentiation measure to differences among bins in which the population was active, bins with no events were discarded prior to the dimensionality reduction step.

Statistical analyses

All analyses were performed with custom Python and R code, using numpy (C. R. Harris et al., 2020), `scipy` (Virtanen et al., 2020), `pandas` (Reback et al., 2020), `scikit-learn` (Pedregosa et al., 2011), `matplotlib` (Hunter, 2007), `seaborn` (Waskom & the seaborn development team, 2020), `lme4` (Bates et al., 2015), `multcomp` (Hothorn et al., 2008), and `emmeans` (Lenth, 2020).

Linear mixed effects models

For analysis of ND across all experimental sessions (Figures 4.3, 4.5, C.4 and C.11 to C.13), we employed linear mixed effects models using the `lmer()` function from the `lme4` package in R with `REML = FALSE` (Bates et al., 2015). The distributions of ND values for both spectral and multivariate differentiation measures were well-approximated by log-normal distributions, so we applied a logarithmic transformation to ND values prior to statistical modeling.

First, we fit an LME model with cortical layer, stimulus category (unscrambled or scrambled), and their interaction as fixed effects, with experimental session as a random effect (`lme4` formula: `"differentiation ~ 1 + layer * stimulus_category + (1 | session)"`). To test layer specificity, we then fit a reduced model with the interaction removed (`"differentiation ~ 1 + layer + stimulus_category + (1 | session)"`) and used a likelihood ratio test to compare the two models.

Next, we fit an LME model with cortical area, stimulus category, and their interaction as fixed effects, with experimental session as a random effect (`lme4` formula: `"differentiation ~ 1 + area * stimulus_category + (1 | session)"`). To test-area specificity, we fit a reduced model with the interaction removed (`"differentiation ~ 1 + area + stimulus_category + (1 | session)"`) and used a likelihood ratio test to compare the two models.

To test for differences in ND among the unscrambled continuous stimuli ("snake (predator)", "crickets (prey)", "man writing", "mousecam", and "conspecifics"; Figure 4.7), we fit an LME model with stimulus as a fixed effect and experimental session as a random effect (`lme4` formula: `"differentiation ~ 1 + stimulus + (1 | session)"`).

We visualized these results by plotting the difference in mean ND for each experimental session; however, no averaging was performed in the statistical analyses.

Post hoc tests

We performed *post hoc* one-sided z-tests to reject the null hypothesis that mean ND for scrambled $\geq$ mean ND for unscrambled in favor of alternative hypothesis that mean ND for scrambled < mean ND for unscrambled using the `glht()` function from the `multcomp` package in R on each LME model with contrasts between stimulus categories (unscrambled or scrambled) within each layer and area, respectively. p values were adjusted for multiple comparisons using the single-step method in `multcomp` (Hothorn et al., 2008).

Post hoc two-sided z-tests for pairwise differences among the unscrambled continuous stimuli were performed with the emmeans function from the emmeans package in R ("`emmeans(model, pairwise ~ stimulus`"), with p values adjusted for multiple comparisons using Tukey's method (Lenth, 2020).

For all *post hoc* tests, simultaneous 95% confidence intervals were obtained using the `confint` methods of the respective model objects. Effect sizes are reported as the Cohen's d value for each pairwise comparison. Cohen's d was calculated with the pooled standard deviation:

$$\sqrt{\frac{(n_1 - 1)\,\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 2}}$$

Permutation tests

Permutation tests were performed for each experimental session to test whether spectral differentiation evoked by unscrambled stimuli was greater than that evoked by scrambled stimuli (Table 4.1). We obtained a null distribution by randomly permuting the trial labels (unscrambled or scrambled) 20,000 times and computing the difference in mean spectral differentiation on unscrambled and scrambled trials for each permutation. p values were computed as the fraction of permutations for which the permuted difference was greater than the observed difference, and significance is reported at the level of $\alpha = 0.05$.

Mediation analyses

Mediation analyses were conducted using the mediation package in R (Tingley et al., 2019).

We analyzed whether the mean event magnitude during a trial mediated the effect of stimulus category on differentiation values by fitting LME models for the mediator and outcome, including arousal variables as covariates, and using the mediate function (mediator model: `"mean_magnitude ~ 1 + stimulus_type + pupil_diameter + locomotion + (1 | session)"`; outcome model: `"differentiation ~ 1 + mean_magnitude + stimulus_type + pupil_diameter + locomotion + (1 | session)"`; treatment: `"stimulus_category''; mediator: ``mean_magnitude"`). This analysis assesses the contribution of the treatment variables on the outcome variable via each of two causal paths: (1) stimulus category and arousal level affect the mean event magnitude, which then in turn affects the measured ND (mediated), and (2) stimulus category and arousal directly affect the measured ND (direct).

For the analysis shown in Table C.2, we fit LME models for each arousal variable (locomotion and pupil diameter) as a mediator, in each case including the other arousal variable as a covariate (*e.g.* `"locomotion ~ 1 + stimulus + pupil_diameter + (1 | session)"`), and an outcome model (`"differentiation ~ 1 + stimulus + locomotion + pupil_diameter + (1 | session)"`). Mediation for a particular pair was evaluated using the mediate function with the "`treat.value`" and "`control.value`" arguments. For each stimulus pair identified as eliciting significantly different ND in our *post hoc* LME analyses, and for each arousal variable, this analysis assesses the contribution of the effect of stimulus on ND via each of two causal paths: (1) the stimulus affects arousal as measured by pupil diameter or locomotion, which then in turn affects ND (mediated), and (2) stimulus directly affects ND, independent of arousal level (direct).

Decoding analyses

For each experimental session, we decoded stimulus category (unscrambled or scrambled) using linear discriminant analysis with the Python package scikit-learn (Pedregosa et al., 2011). First, the responses to each category were concatenated to form a $s \times n \cdot t$ matrix, where s is the number of stimulus presentation trials, n is the number of cells recorded, and t is the number of two-photon imaging samples in a single trial. To obtain a tractable number of features for linear discriminant analysis, we used PCA to reduce the dimensionality of the matrix such that the number of components c was sufficient to retain 99% of the variance along the rows, yielding an $s \times c$ matrix (`sklearn.decomposition.PCA(n_components=0.99)`). This was then used to train a shrinkage-regularized LDA classifier with fivefold cross-validation (`sklearn.discriminant_analysis.LinearDiscriminantAnalysis(solver='lsqr', shrinkage='auto')`). We report the mean balanced accuracy score (`sklearn.metrics.balanced_accuracy_score`) on the heldout test data across cross-validation folds. Chance performance is 0.5.

For Figure C.14, we used the same procedure as described above, but the classifier was trained to decode stimulus identity rather than category; chance performance is 1/12. For Figure C.15, we used the same procedure but trained the classifier using only responses to the 5 continuous naturalistic stimuli, and classifier performance was evaluated for each stimulus separately with the F1 score.

Code accessibility

The analysis code used in this work is freely available online at `https://github.com/wmayner/openscope-differentiation`. The data are published as Mayner et al. (2021).

4.3 Results

Using *in vivo* two-photon calcium imaging (Figure 4.1A–D), we recorded from the left visual cortex of awake mice while they passively viewed stimuli presented to the contralateral eye. We used the transgenic mouse lines Cux2, Rorb, and Rbp4, in which GCaMP6f is expressed in excitatory neurons predominantly in L2/3, L4, and L5, respectively (3 mice each; Cux2, 2 males; Rorb, 3 males; Rbp4, 1 male; § 4.2, Transgenic mice). Visual cortical areas were delineated via intrinsic signal imaging (ISI; Figure 4.1B). Data were collected from L2/3, L4, and L5 in each of 5 areas (V1, LM, AL, PM, and AM; Figure 1E) across 45 experimental sessions (15 sessions per transgenic line; 9 sessions per area; 5 ± 1 sessions per mouse; number of cells shown in Table C.1). Mice were head-fixed and free to move on a rotating disc while pupil diameter and running velocity were recorded. During each 70-minute session, twelve 30 s movie stimuli were presented in a randomized block design with 10 repetitions, with 4 s of mean-luminance grey shown between stimulus presentations (Figure 4.1F,G, Figure C.1). Stimuli were presented in greyscale but were not otherwise modified (in particular, it should be noted that spatial frequencies beyond the mouse acuity limit will appear blurred to the mice). Representative $\Delta F/F_0$ traces and behavioral data are shown in Figure 1H. One imaging session in L5 of AL was excluded from our analyses because of technical problems with the two-photon recording.

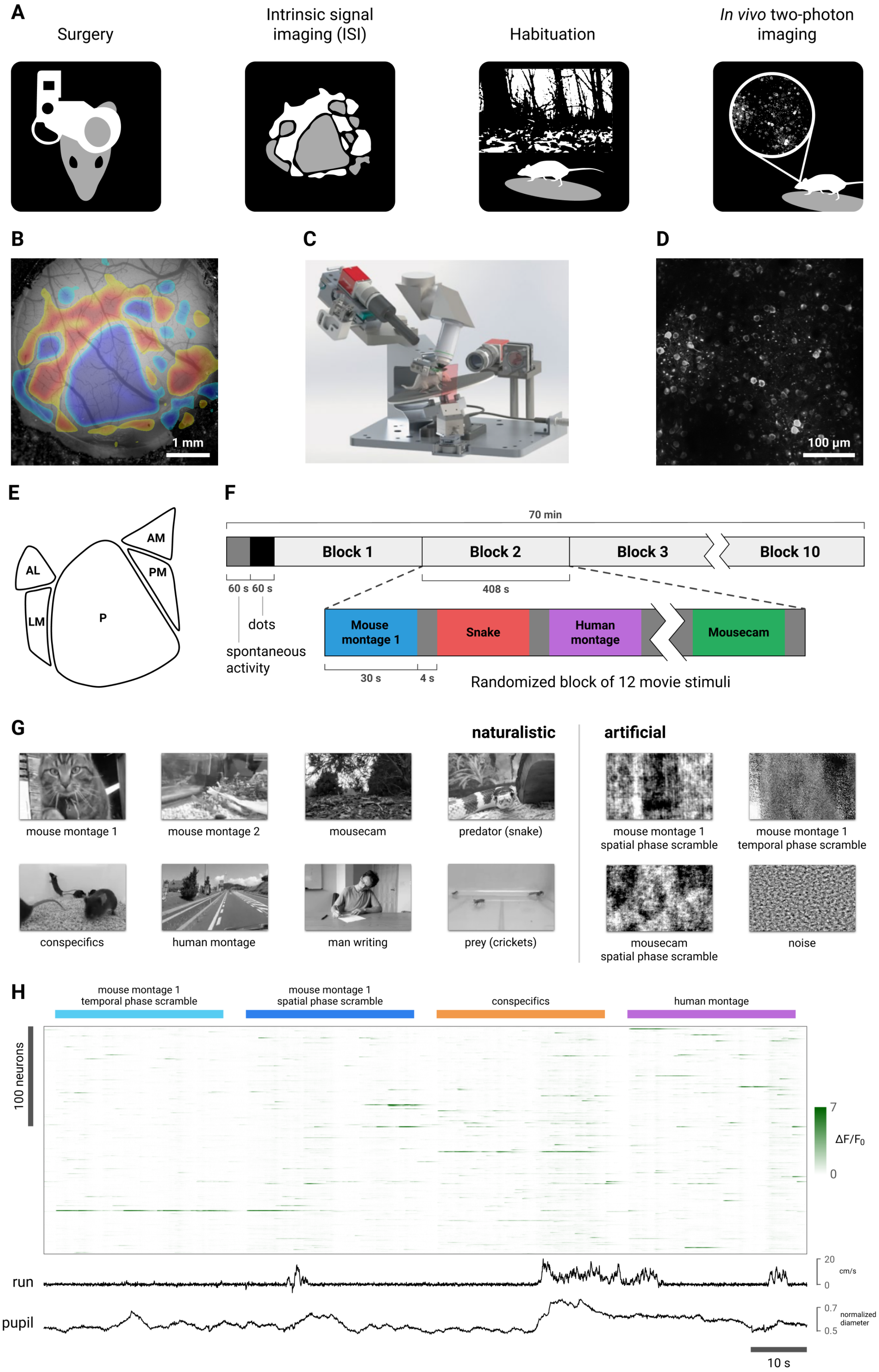

A
Surgery
Intrinsic signal imaging (ISI)
Habituation
In vivo two-photon imaging
B
1 mm
C
D
100 µm
E
AM
AL
PM
LM
P
F
70 min
Block 1
Block 2
Block 3
Block 10
60 s
60 s
408 s
dots
spontaneous activity
Mouse montage 1
Snake
Human montage
Mousecam
30 s
4 s
Randomized block of 12 movie stimuli
G
naturalistic
artificial
mouse montage 1
mouse montage 2
mousecam
predator (snake)
mouse montage 1 spatial phase scramble
mouse montage 1 temporal phase scramble
conspecifics
human montage
man writing
prey (crickets)
mousecam spatial phase scramble
noise
H
mouse montage 1 temporal phase scramble
mouse montage 1 spatial phase scramble
conspecifics
human montage
100 neurons
7
ΔF/F0
0
run
20
cm/s
0
pupil
0.7
normalized diameter
0.5
10 s

Figure 4.1. Experimental design. (A) Data were acquired using a standardized two-photon calcium imaging pipeline based on that described in de Vries et al. (2020) and Groblewski et al. (2020) (§ 4.2, Materials and methods). Briefly, a custom headframe was implanted; intrinsic signal imaging (ISI) was performed to delineate retinotopically mapped visual areas; the mouse was habituated to the passive viewing paradigm over the course of ~2 weeks; and two-photon calcium imaging was performed in the left visual cortex while animals viewed stimuli presented to the contralateral eye in several experimental sessions. **(B)** Example of an ISI map. **(C)** Schematic of the two-photon imaging rig (reproduced with permission from Figure 1D of de Vries et al., 2020). During the imaging sessions, head-fixed mice were free to run on a rotating disc. Locomotion velocity was recorded and pupil diameter was extracted from video of the animal's right eye. **(D)** Example frame from a two-photon movie. Imaging data was processed as described in de Vries et al. (2020) to obtain $\Delta F/F_0$ traces. **(E)** Schematic of the 5 visual areas targeted in this study. **(F)** 10 randomized blocks of twelve 30 s movie stimuli were presented. 4 s of mean-luminance grey was presented between stimuli. The first 60 s was mean-luminance grey (spontaneous activity); the second 60 s period was a high-contrast sparse noise stimulus (not analyzed in this work). **(G)** Still frames from the 8 naturalistic (*left*) and 4 artificial (*right*) movie stimuli (§ 4.2, Stimuli). Two of the naturalistic stimuli, "mouse montage 1" and "mousecam", were phase-scrambled to destroy high-level image features while closely matching low-order statistics (§ 4.2, Phase scrambling). **(H)** Representative calcium imaging and behavioral data. A heatmap of $\Delta F/F_0$ values is shown for 228 neurons simultaneously imaged in L2/3 of AL during presentation of four stimuli, with locomotion velocity and normalized pupil diameter plotted below. Numbers of cells recorded from each layer and area are listed in Table C.1. Calcium indicator kinetics did not differ across cell populations (Figure C.2).

To measure ND, we employed a method from Mensen et al. (2018) for analyzing a set of timeseries recorded during the presentation of a continuous stimulus (Figure 4.2). Briefly, the power spectrum of each cell's $\Delta F/F_0$ trace was estimated in 1 s windows. The cells' power spectra during simultaneous windows were concatenated to form a vector representing the neurophysiological state of the population during that window. We calculated ND for each trial as the median Euclidean distance between the 30 population states elicited over the course of the 30 s stimulus. We computed distances in the frequency domain rather than the time domain to focus on differences in overall population state rather than differences in precise timing of $\Delta F/F_0$ transients. To account for variability in the size of the imaged populations we divided ND values by the square root of the number of cells (§ 4.2, Spectral differentiation). Spectral differentiation is zero when the set of $\Delta F/F_0$ traces is perfectly periodic with a period of 1 s (the window size), and it is high when many traces exhibit temporally varied patterns across the 30 seconds (Figure C.3). The measure scales with the magnitude of the signal and thus has no well-defined maximum.

To compare the differentiation of responses to naturalistic and artificial stimuli, we generated Fourier phase-scrambled versions of two of our movie stimuli. Phase-scrambling destroys the naturalistic structure of the stimulus while closely matching the power spectrum (the spectrum was not conserved exactly because of numerical representational limitations of the stimulus format; § 4.2, Phase scrambling).

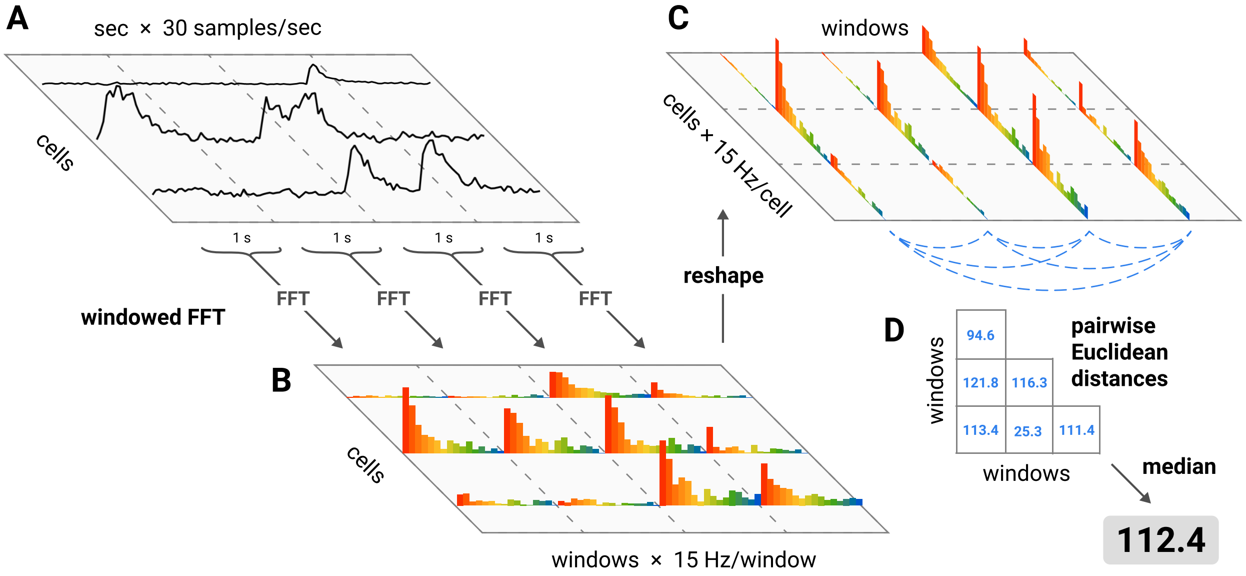


Figure 4.2. Spectral differentiation analysis. ND was computed as follows: **(A)** for each cell, the $\Delta F / F_0$ trace of each cell during stimulus presentation was divided into 1 s windows; **(B)** the power spectrum of each window was estimated; **(C)** the "neurophysiological state" during each 1 s window was defined as a vector in the high-dimensional space of cells and frequencies (*i.e.*, the concatenation of the power spectra in that window for each cell); **(D)** the ND in response to a given stimulus was calculated as the median of the pairwise Euclidean distances between every state that occurred during the stimulus presentation. An illustration of how the measure behaves is shown in Figure C.3.

Note that operations that leave the power spectrum of a signal unchanged will not affect its spectral differentiation.

For the "mouse montage 1" stimulus (a montage of six 5 s naturalistic movie clips), we performed the phase-scrambling in two ways: (1) along the temporal dimension, on each pixel independently, and (2) along the two spatial dimensions, on all pixels. For the "mousecam" stimulus (a continuous 30 s clip of movement at ground level through the underbrush of a forest) we performed only the spatial phase-scrambling. This yielded 2 unscrambled stimuli and 3 scrambled stimuli (Figure C.1). The full set of twelve stimuli was designed to span different levels of putative ethological relevance; here, we focus on the comparison of the unscrambled stimuli to their scrambled versions because low-order stimulus statistics are controlled and thus the contrast can be more easily interpreted.

Unscrambled stimuli elicit more differentiated responses compared to scrambled stimuli

We hypothesized that the unscrambled stimuli would elicit higher ND than their phase-scrambled counterparts. We tested this by fitting linear mixed effects (LME) models with experimental session as a random effect (§ 4.2, Linear mixed effects models); mean differences in ND of responses to unscrambled *vs.* scrambled stimuli are shown in Figure 4.3. We obtained similar results contrasting naturalistic *vs.* artificial stimuli across the entire stimulus set (Figure C.4). ND values were approximately log-normally distributed, so we applied a logarithmic transform to ND in all statistical analyses (§ 4.2, Statistical analyses).

Increased differentiation for unscrambled stimuli is specific to excitatory cells in L2/3

We found that unscrambled stimuli elicited more differentiated responses specifically in L2/3 (Figure 4.3A). We fitted an LME model with stimulus category (unscrambled or scrambled), layer, and their interaction as fixed effects and found a significant interaction (likelihood ratio test, $\chi^2(2) = 13.379$, $p = 0.00124$). *Post hoc* tests showed that the unscrambled *vs.* scrambled difference was specific to L2/3 (one-sided z-test; L2/3, $z = 3.866$, $p = 1.66 \times 10^{-4}$, Cohen's $d = 0.164$, 95% confidence interval (CI) $[0.051, \infty)$; L4, $z = 0.191$, $p = 0.810$, Cohen's $d = 0.011$, 95% CI $[-0.057, \infty)$; L5, $z = -1.168$, $p = 0.998$, Cohen's $d = -0.067$, 95% CI $[-0.100, \infty)$; p values and CIs adjusted for multiple comparisons).

Increased differentiation for unscrambled stimuli is specific to areas AL and AM

The increased ND in response to unscrambled stimuli was area-specific (Figure 4.3B). We fitted an LME model with stimulus category, area, and their interaction as fixed effects and found a significant interaction (likelihood ratio test, $\chi^2(4) = 15.203$, $p = 0.00430$). *Post hoc* tests showed that the unscrambled *vs.* scrambled difference was specific to AL and AM (one-sided z-test; V1, $z = 0.704$, $p = 0.748$, Cohen's $d = 0.054$, 95% CI $[-0.061, \infty)$; LM, $z = -0.234$, $p = 0.989$, Cohen's $d = -0.016$, 95% CI $[-0.097, \infty)$; AL, $z = 2.873$, $p = 0.0101$, Cohen's $d = 0.200$, 95% CI $[0.022, \infty)$; PM, $z = -1.843$, $p > 0.999$, Cohen's $d = -0.122$, 95% CI $[-0.157, \infty)$; AM, $z = 2.446$, $p = 0.0356$, Cohen's $d = 0.268$, 95% CI $[0.128, \infty)$; adjusted for multiple comparisons).

It is conceivable that these results are artifacts of our implementation of the spectral differentiation measure. To check the robustness of our findings, we performed a sensitivity analysis in which we systematically varied (1) the distance metric used to assess differences between population states, (2) the window length that defines the state of the neural population, (3) the frequency bin spacing in the spectra, and (4) the window function and amount of overlap used in the spectral estimation step (Figures C.5 to C.7). The results were qualitatively the same for nearly all combinations of these

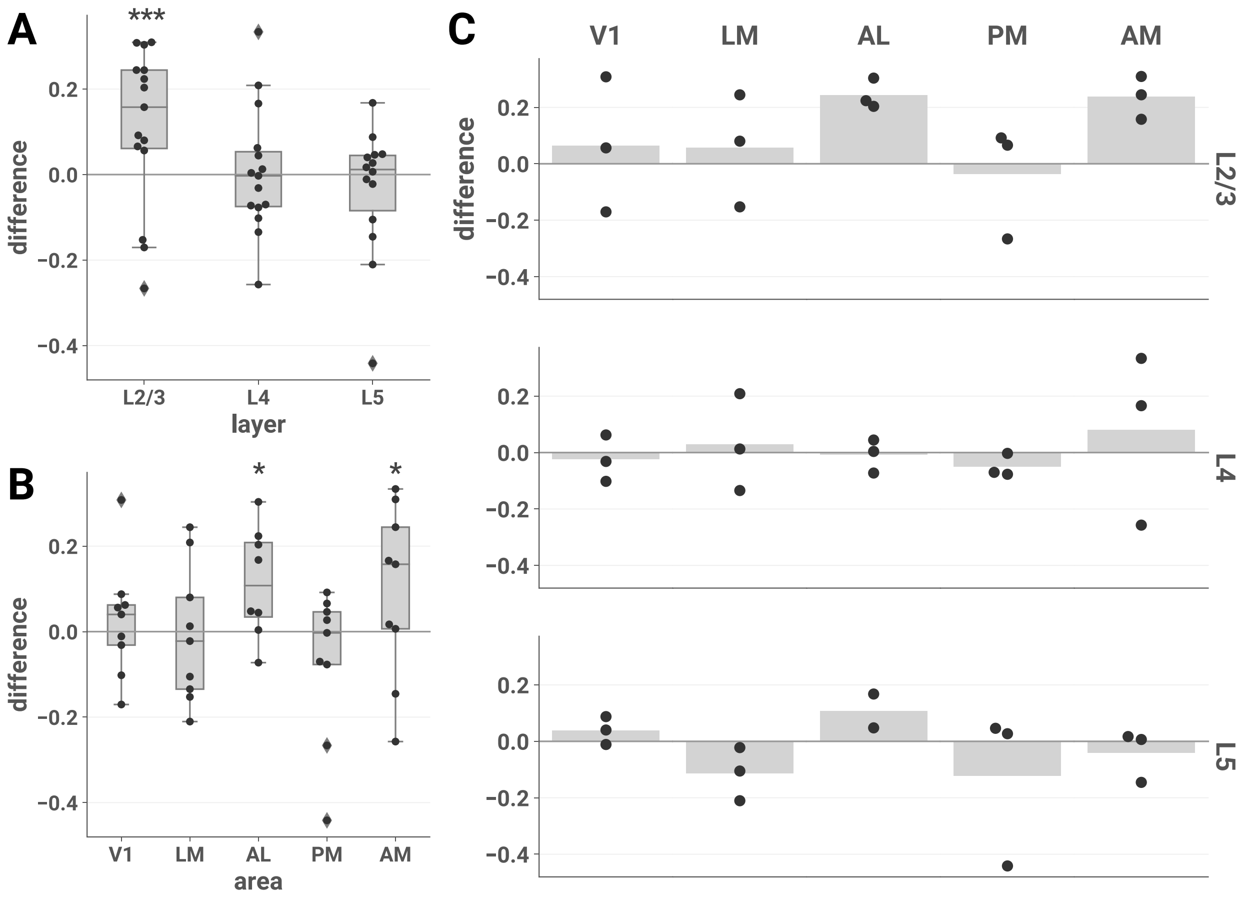


Figure 4.3. ND elicited by unscrambled *vs.* scrambled stimuli is higher in L2/3 of areas AL and AM. The difference in ND of responses to unscrambled *vs.* scrambled stimuli is plotted for each session by layer **(A)**, area **(B)**, and layer-area pair **(C)**. Each point represents the difference between the mean ND of responses to the 2 unscrambled and the 3 scrambled stimuli during a single experimental session. Similar results were found contrasting naturalistic *vs.* artificial stimuli across the entire stimulus set (Figure C.4). To demonstrate the robustness of this effect, we conducted several further analyses. Sensitivity analyses showed similar findings for various choices of analysis parameters (Figures C.5 to C.7) and when pupil diameter and locomotion were included as covariates in the LME models (Figures C.8 to C.10). We found similar results when we performed the same analysis on discrete calcium events detected from the $\Delta F/F_0$ traces with an L_0-regularized algorithm (§ 4.2, Event detection), indicating that the effect is driven by differences in the large-timescale patterns of responses rather than small-timescale spectral differences within windows (Figure C.11). Finally, we also found similar results when we removed event-triggered transients from the $\Delta F/F_0$ traces, indicating that the effect is not driven solely by initial transients in the calcium response (Figure C.12). **(A)** and **(B)**: asterisks indicate significant *post hoc* one-sided z-tests in the layer (A) and area (B) interaction LME models (*, $p < 0.05$; ***, $p < 0.001$). Boxes indicate quartiles; whiskers indicate the minimum and maximum of data lying within 1.5 times the inter-quartile range of the 25% or 75% quartiles; diamonds indicate observations outside this range. **(C)** Mean values are indicated by bars.

parameters we tested.

Since arousal state modulates neuronal activity in visual cortex (Dadarlat & Stryker, 2017; McGinley, Vinck, et al., 2015; Niell & Stryker, 2010; Polack et al., 2013; Reimer et al., 2014; Salkoff et al., 2020; Vinck et al., 2015), the increase in firing rates seen during periods of high arousal raises the possibility that the differences in ND we observed could be due to changes in arousal alone rather than stimulus category. To rule this out, we repeated the sensitivity analysis of our main results while including locomotion and pupil diameter as covariates in the LME models. Consistent with the simpler models, for nearly all parameter combinations, L2/3 of AL & AM emerged as the cell populations in which ND is greater for unscrambled *vs.* scrambled stimuli (Figures C.8 to C.10), indicating that the measured arousal variables are insufficient to fully explain the differences in ND we observed. We also analyzed whether the mean magnitude of calcium events (a proxy for firing rate) mediated the effect of stimulus category and found evidence for both mediated and direct effects (mediated effect: 0.1351, 95% CI $[0.0784, 0.20]$, $p < 2 \times 10^{-16}$; direct effect: 0.0951, 95% CI $[0.0400, 0.15]$, $p = 0.002$; proportion of total effect mediated: 0.5898, 95% CI $[0.3963, 0.80]$, $p < 2 \times 10^{-16}$). That is, unscrambled stimuli led to increased ND relative to scrambled stimuli both directly, independent of mean event magnitude, and indirectly, via increases in mean event magnitude that in turn increased ND. Thus, while a portion of the effect of stimulus category was mediated by changes in population firing rate, this mediated effect is likewise insufficient to fully explain our results.

Permutation tests for individual experimental sessions

The above analysis shows that the mean ND elicited by unscrambled stimuli is greater than for their phase-scrambled counterparts, and that this effect is driven by L2/3 cells in areas AL and AM. We also analyzed ND at the level of individual sessions with non-parametric permutation tests. For each session, we obtained a null distribution by randomly permuting the trial labels (unscrambled or scrambled) 20,000 times and computing the difference in mean ND on unscrambled *vs.* scrambled trials for each permutation. p values were computed as the fraction of permutations for which the permuted difference was greater than the observed difference.

The results of the individual session analyses were consistent with the LME analyses (Table 4.1). In all sessions recorded from L2/3 of AL & AM, responses to unscrambled stimuli were significantly more differentiated than to scrambled stimuli ($p < 0.05$).

Table 4.1. Permutation tests show increased ND for unscrambled *vs.* scrambled stimuli in L2/3 of AL & AM at the level of individual experimental sessions. Entries contain the fraction of sessions in which the mean ND of responses to unscrambled stimuli was significantly greater than responses to their scrambled counterparts *vs.* total number of sessions at a threshold of $\alpha = 0.05$. For each session, a null distribution was obtained by randomly permuting trial labels (unscrambled or scrambled) 20,000 times and computing the difference in mean ND on unscrambled and scrambled trials for each permutation. p values were computed as the fraction of permutations for which the permuted difference was greater than the observed difference.

	V1	**LM**	**AL**	**PM**	**AM**	**All areas**
L2/3	1/3	1/3	3/3	0/3	3/3	8/15
L4	0/3	1/3	0/3	0/3	0/3	1/15
L5	0/3	0/3	0/2	0/3	0/3	0/14
All layers	1/9	2/9	3/8	0/9	3/9	9/44

Arousal is correlated with effect size

Locomotion and pupil diameter can be considered behavioral indications of engagement with the environment (Bennett et al., 2013; Ganea et al., 2020; Jacobs et al., 2020). We found that in L2/3 of AL & AM, effect sizes were positively correlated with locomotion activity (Figure 4.4, top left; Pearson's $r = 0.896$; two-sided t-test; $t(4) = 4.030$, $p = 0.0157$, 95% CI $[0.308, 1.00]$) and pupil diameter (Figure 4, top right; $r = 0.716$; $t(4) = 2.054$, $p = 0.109$, 95% CI $[-0.227, 1.00]$), suggesting that the difference in ND is more clear when the animal is engaged. However, we note that the relatively restricted range of observed mean locomotion fraction and pupil diameter values in the sessions of interest limits the generalizability of these conclusions.

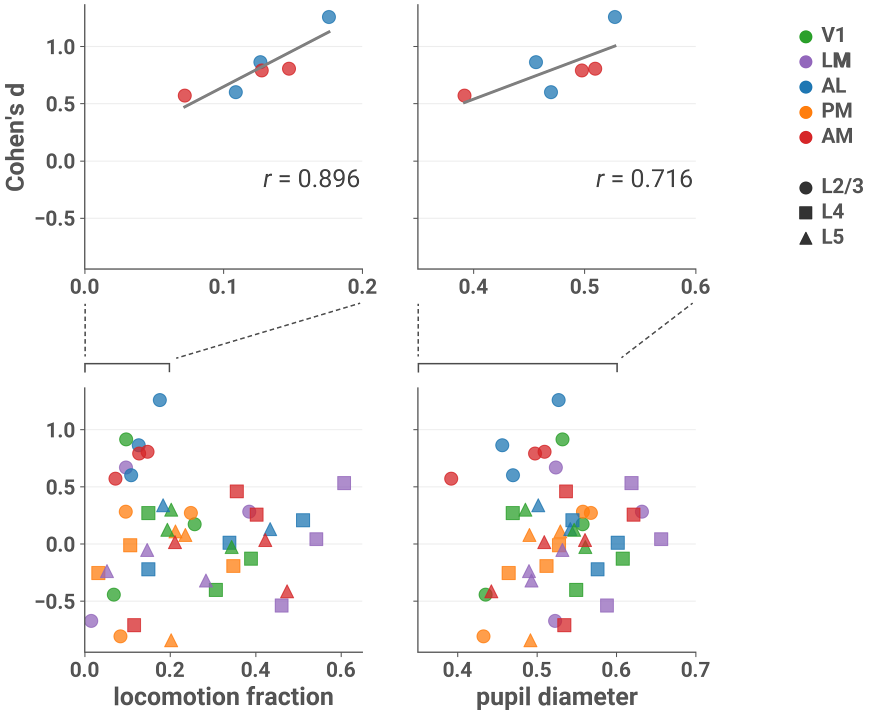


Figure 4.4. Effect sizes in L2/3 of AL & AM are larger in sessions with more locomotion and larger pupil diameter. Cohen's d is plotted against the fraction of locomotion activity (*left column*) and mean normalized pupil diameter (*right column*) during the session, with linear fit in grey. *Top row*: Only sessions recorded from L2/3 and areas AL or AM. *Bottom row*: All sessions (note different scales). *Top left*: Pearson's $r = 0.896$ (two-sided t-test; $t(4) = 4.030$, $p = 0.0157$, 95% CI $[0.308, 1.00)$. *Top right*: $r = 0.716$ ($t(4) = 2.054$, $p = 0.109$, 95% CI $[-0.227, 1.00]$). Running velocity greater than 2.5 cm/s was considered locomotion activity (§ 4.2, Locomotion). Normalized pupil diameter was obtained by dividing by the maximum diameter that occurred during the session (§ 4.2, Pupillometry).

Multivariate analysis also shows increased differentiation for unscrambled stimuli

Spectral differentiation is a univariate measure in the sense that the coordinates of the population state vectors are orthogonal, so that each squared difference term in the Euclidean distance reflects differences only within a given cell's responses across time. To ensure that our results were not due to this method of measuring ND, we also employed a multivariate approach that considers spatiotemporal differences in activity patterns across the cell population. For each session, the dimensionality of the population response vectors was reduced to 8 using UMAP (McInnes et al., 2020). In the resulting 8-dimensional space, ND was measured as the mean Euclidean distance to the centroid of the set of responses corresponding to that stimulus (§ 4.2, Multivariate differentiation).

The results of the multivariate analysis were consistent with those found using the spectral differentiation measure. The mean centroid distance was higher in response to unscrambled compared to scrambled stimuli (Figure 4.5), and this effect was specific to L2/3 (layer × stimulus category interaction: likelihood ratio test, $\chi^2(2) = 18.135$, $p = 1.154 \times 10^{-4}$; *post hoc* one-sided z-tests: L2/3, $z = 5.149$, $p = 3.92 \times 10^{-7}$, Cohen's $d = 0.194$, 95% CI $[0.0181, \infty)$; L4, $z = 1.749$, $p = 0.116$, Cohen's $d = 0.0994$, 95% CI $[-0.00221, \infty)$; L5, $z = -0.938$, $p = 0.995$, Cohen's $d = -0.0651$, 95% CI $[-0.0189, \infty)$) and areas AL and AM (area × stimulus category interaction: likelihood ratio test, $\chi^2(4) = 16.232$, $p = 0.00272$; post hoc tests: V1, $z = 0.420$, $p = 0.872$, Cohen's $d = 0.0281$, 95% CI $[-0.0146, \infty)$; LM, $z = -0.047$, $p = 0.974$, Cohen's $d = -0.00269$, 95% CI $[-0.0182, \infty)$; AL, $z = 2.941$, $p = 0.00816$, Cohen's $d = 0.184$, 95% CI $[0.00508, \infty)$; PM, $z = 0.152$, $p = 0.945$, Cohen's $d = 0.0119$, 95% CI $[-0.0167, \infty)$; AM, $z = 4.436$, $p = 2.29 \times 10^{-5}$, Cohen's $d = 0.277$, 95% CI $[0.0163, \infty)$.

The multivariate differentiation measure is also suitable for use on discrete data. To support our main findings, we analyzed discrete calcium events detected from the $\Delta F / F_0$ traces with an L_0-regularized algorithm (§ 4.2, Event detection). The results of multivariate differentiation analysis of these data were consistent with our results using $\Delta F / F_0$ traces (Figure C.13), with the additional finding of significantly greater differentiation for unscrambled *vs.* scrambled stimuli in L2/3 of V1 as well as AL & AM (layer × stimulus category interaction: likelihood ratio test, $\chi^2 = 15.029$, $p = 0.000545$; *post hoc* one-sided z-tests: L2/3, $z = 5.337$, $p = 1.42 \times 10^{-7}$, Cohen's $d = 0.178$, 95% CI $[0.0158, \infty)$; L4, $z = 1.698$, p = 0.128, Cohen's $d = 0.0611$, 95% CI $[-0.00208, \infty)$; L5, $z = -0.124$, $p = 0.909$, Cohen's $d = -0.00377$, 95% CI $[-0.0114, \infty)$; area × stimulus category interaction: likelihood ratio test, $\chi^2(4) = 20.854$, $p = 0.000339$; *post hoc* tests: V1, $z = 3.132$, $p = 0.00434$, Cohen's $d = 0.167$, 95% CI $[0.00516, \infty)$; LM, $z = -0.798$, $p > 0.999$, Cohen's $d = -0.0451$, 95% CI $[-0.0198, \infty)$; AL, $z = 2.757$, $p = 0.0145$, Cohen's d = 0.100, 95% CI $[0.00295, \infty)$; PM, $z = -0.366$, $p = 0.994$, Cohen's $d = -0.0160$, 95% CI $[-0.0170, \infty)$; AM, $z = 4.372$, $p = 3.07 \times 10^{-5}$, Cohen's $d = 0.158$, 95% CI $[0.0130, \infty)$.

Decoding analysis does not reveal layer or area specificity

We next asked whether the layer and area specificity of our ND results would be reflected in our ability to decode the stimulus category (unscrambled or scrambled) from population responses. We performed fivefold cross-validated linear discriminant analysis to decode stimulus category for each session and scored the classifier using balanced accuracy (§ 4.2, Decoding analyses). Decoding performance was high for most areas and layers (Figure 4.6), in contrast to the unscrambled-scrambled difference in ND. Performance was also high across layers and areas when we decoded stimulus identity, rather than

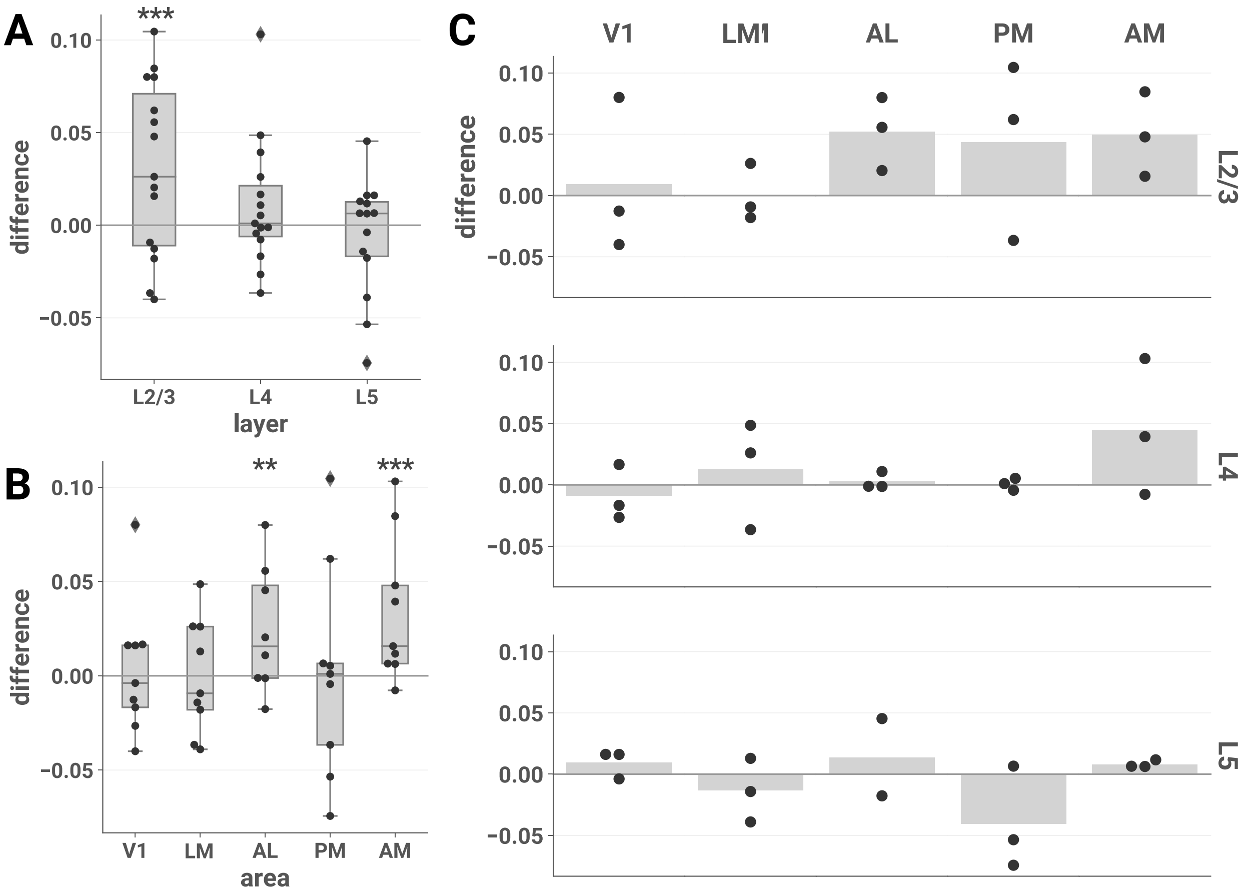


Figure 4.5. Multivariate differentiation analysis. The mean difference in the mean centroid distance of responses to unscrambled *vs.* scrambled stimuli is plotted for each session by layer (A), area (B), and layer-area pair (C). ND elicited by unscrambled *vs.* scrambled stimuli is higher in L2/3 and areas AL and AM, consistent with the spectral differentiation analysis. We found similar results when we analyzed discrete L_0 calcium events detected from the $\Delta F / F_0$ traces (§ 4.2, Event detection, Figure C.13). **(A)** and **(B)**: asterisks indicate significant *post hoc* one-sided z-tests in the layer (A) and area (B) interaction LME models (**, $p < 0.01$; ***, $p < 0.001$). Boxes indicate quartiles; whiskers indicate the minimum and maximum of data lying within 1.5 times the inter-quartile range of the 25% or 75% quartiles; diamonds indicate observations outside this range. **(C)** Mean values are indicated by bars.

category, using responses to all 12 stimuli (Figure C.14).

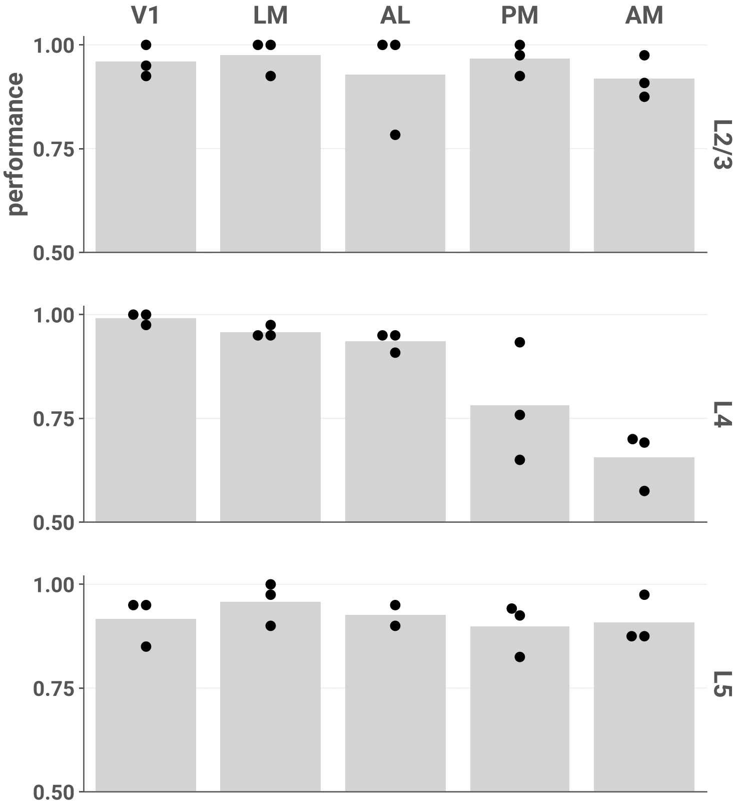


Figure 4.6. Stimulus category (unscrambled or scrambled) can be accurately decoded from most layers and areas. Each point represents the mean fivefold cross-validated balanced accuracy score of linear discriminant analysis performed on a single session (§ 4.2, Decoding analyses). Chance performance is 0.5. We found similar results when decoding stimulus identity across all 12 stimuli (Figure C.14).

Differences in ND among individual stimuli

We also investigated whether ND differed among stimuli within the same category. This analysis was restricted to the set of unscrambled stimuli without jump cuts, *i.e.,* the 5 naturalistic continuous 30 s clips, to avoid potential confounds in comparing stimuli with and without abrupt transitions between different scenes. Here we used data from all layers and areas, since although responses from L2/3 of AL & AM drove the unscrambled/scrambled differences, within-category differences might not be restricted to that subset. We fitted an LME model with stimulus as a fixed effect and found it was significant (likelihood ratio test, $\chi^2(4) = 32.115$, $p = 1.812 \times 10^{-6}$). Post-hoc pairwise two-sided z-tests (adjusted for multiple comparisons), shown in Figure 4.7, revealed that the predator stimulus (a snake) evoked significantly higher differentiation than clips of conspecifics ($z = 3.229$, $p = 0.0110$, Cohen's $d = 0.156$, 95% CI $[0.015, 0.180]$; prey (crickets) ($z = 3.928$, $p = 8.149 \times 10^{-4}$, Cohen's $d = 0.181$, 95% CI $[0.036, 0.201]$), and a man writing ($z = 5.249$, $p = 1.522 \times 10^{-6}$, Cohen's $d = 0.232$, 95%

CI $[0.076, 0.241]$). The “mousecam” clip of movement through a wooded environment also evoked a significantly higher differentiation than the clip of a man writing ($z = 3.396$, $p = 0.00615$, Cohen’s $d = 0.154$, 95% CI $[0.020, 0.185]$). Mediation analysis showed a mixture of direct and arousal-mediated effects, indicating that changes in arousal cannot fully account for these differences (Table C.2). Here we present the main effect of stimulus; for an exploration of interactions with layer and area, and a comparison to decoding, see Figure C.15.

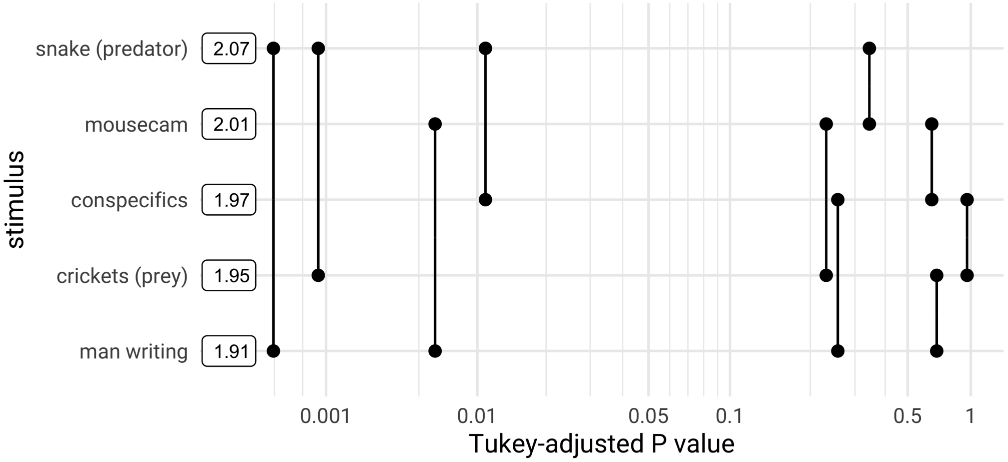


Figure 4.7. Pairwise differences in ND among unscrambled, continuous stimuli. *Post hoc* pairwise comparisons using data from all neuronal populations are plotted against their p values (adjusted for multiple comparisons). Boxes show mean ND for each stimulus. ND of the snake stimulus is significantly greater than that of crickets and man writing at a threshold of $\alpha = 0.01$, and greater than conspecifics at $\alpha = 0.05$. ND of the mousecam stimulus is greater than that of man writing at $\alpha = 0.01$. Mediation analysis showed a mixture of direct and arousal-mediated effects, indicating that changes in arousal cannot fully account for these differences (Table C.2). Pairwise differences in ND and decoding performance stratified by layer and area are shown in Figure C.15.

Stimulus differentiation does not explain ND

It is possible that ND does not reflect functionally relevant visual processing but is instead merely inherited from the differentiation of the stimulus itself. To rule out this possibility, we computed the stimulus differentiation (SD) by treating each pixel of the stimulus as a “cell” and applying the spectral differentiation measure to the traces of pixel intensities over time after blurring the stimulus to account for the coarseness of mouse vision (§ 4.2, Spectral differentiation). Within L2/3 of AL and AM, the mean ND elicited by each stimulus was positively correlated with SD (Figure 4.8; Pearson’s $r = 0.746$, one-sided t-test; $t(10) = 3.542$, $p = 0.00267$, 95% CI $[0.393, 1.00]$). However, the noise stimulus is a highly influential observation (Cook’s $D = 2.318$, an order of magnitude larger than the next most influential observation). If we exclude this stimulus, we find a weaker correlation ($r = 0.258$; one-sided t-test; $t(9) = 0.801$, $p = 0.222$, 95% CI $[-0.307, 1.00]$). Furthermore, there was no evidence of a relationship with ND when considering only the scrambled stimuli and their unscrambled counterparts ($r = -0.378$; two-sided t-test; $t(3) = -0.708$, $p = 0.530$, 95% CI $[-0.945, 0.756]$). Thus, we conclude that ND is not inherited from SD. We also did not find a relationship with stimulus luminance, contrast, or spectral energy (Figure C.16).

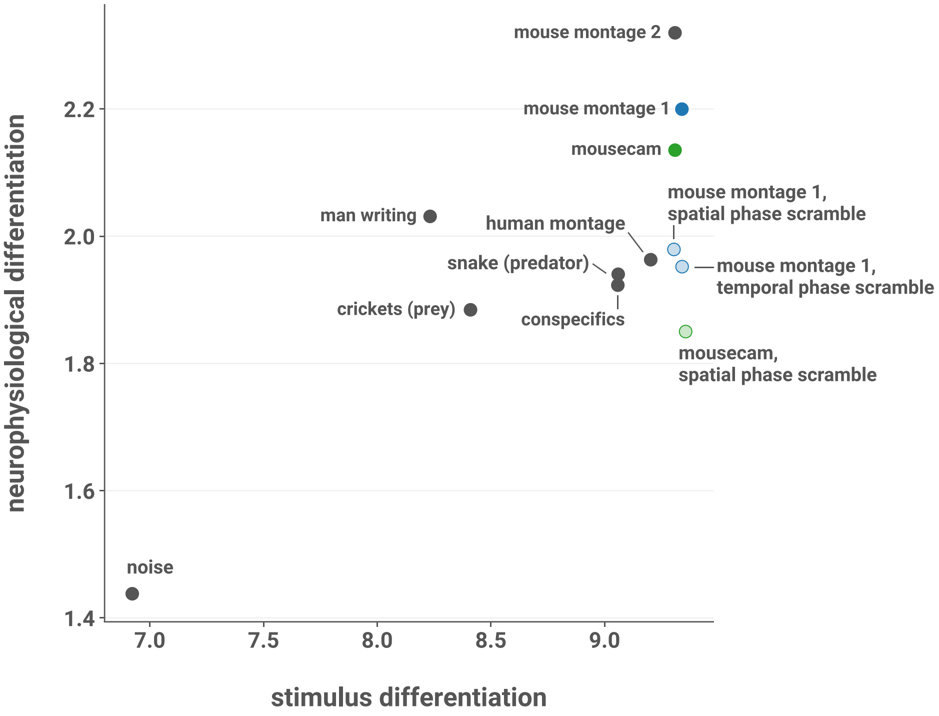


Figure 4.8. SD does not explain ND. Mean ND elicited by each stimulus in L2/3 of AL and AM, plotted against SD. SD was computed by treating each pixel of the movie as a "cell" and applying the spectral differentiation measure to traces of pixel intensities over time after blurring the movie with a Gaussian filter to account for the coarseness of mouse vision. Across all stimuli, mean ND is positively correlated with SD (Pearson's $r = 0.746$; one-sided t-test; $t(10) = 3.542$, $p = 0.00267$, 95% CI $[0.393, 1.00]$). However, here the noise stimulus is a highly influential observation (Cook's $D = 2.318$, an order of magnitude larger than the next most influential observation). With the noise stimulus excluded, the correlation is weaker ($r = 0.258$; one-sided t-test; $t(9) = 0.801$, $p = 0.222$, 95% CI $[-0.307, 1.00]$). Moreover, there was no evidence of a relationship with ND when considering only the scrambled stimuli and their unscrambled counterparts ($r = -0.378$; two-sided t-test; $t(3) = -0.708$, $p = 0.523$, 95% CI $[-0.945, 0.756]$). ND was also not explained by variation in stimulus luminance, contrast, or spectral energy (Figure C.16).

4.4 Discussion

Our results show that excitatory L2/3 neurons in visual areas AL and AM have more differentiated responses to stimuli with naturalistic structure than to phase-scrambled stimuli with closely matched low-order statistics, indicating that these populations are uniquely sensitive to high-level natural features in this stimulus set. We found this difference in neurophysiological differentiation (ND) at the level of single experimental sessions, and it was robust to complementary methods of measuring ND. We found that effect sizes were larger with increasing pupil diameter and locomotion, suggesting sensitivity to the animal' s arousal level. Decoding analysis showed a marked lack of area and layer specificity: stimulus category could be accurately decoded from the activity of most cell populations we surveyed. Besides the differences between unscrambled and scrambled stimuli, we found differences in ND among unscrambled stimuli. Finally, we argued that ND is not merely inherited from the stimulus differentiation.

The precise functional specialization of visual areas in the mouse remains unclear (Glickfeld & Olsen, 2017). Recent large-scale anatomical (J. A. Harris et al., 2019) and functional (Siegle, Jia, et al., 2021) studies have uncovered a "shallow hierarchy" in which V1 lies at the base, followed by LM, RL, AL, and PM, with AM at the top. In this light, our findings that ND in L2/3 of AL and AM is sensitive to high-level naturalistic structure could be interpreted as a reflection of hierarchical processing, which may be constructing a richer dynamical repertoire for perception of naturalistic stimuli at higher hierarchical levels. Interestingly, we did not find this effect in PM, despite its intermediate position between AL and AM in the hierarchy, suggesting that such hypothetical processing towards richer repertoires is not fully determined by the one-dimensional hierarchy, but may involve specific pathways through subsets of visual areas. These observations indicate that differentiation analysis may help refine our understanding of functional specialization of brain areas and uncover differences between them that can be used to direct further investigations.

A recent study found that feedback projections from higher visual areas to L2/3 excitatory neurons in V1 create a second receptive field (RF) surrounding the feedforward RF and that these RFs are mutually antagonistic, pointing to a role for these neurons in predictive processing (Keller et al., 2020). If this pattern is present at higher levels of the visual hierarchy, then the layer specificity we find could be explained by a scenario in which feedback to AL and AM from areas higher in the putative dorsal stream (Marshel et al., 2011; Wang et al., 2012) are integrated with feedforward inputs in L2/3 to compute prediction errors about high-level visual features. In this scenario, the naturalistic stimuli, which contain high-level features that are presumably less predictable, would elicit more prediction errors and thus more differentiated activity.

Stimulus-evoked activity in cortex is modulated by arousal level and behavioral state (McGinley, Vinck, et al., 2015; Salkoff et al., 2020). Locomotion is associated with heightened arousal, increased membrane depolarization, firing rates, and signal-to-noise ratio, and enhanced stimulus encoding (Bennett et al., 2013; Dadarlat & Stryker, 2017; Niell & Stryker, 2010; Polack et al., 2013; Vinck et al., 2015). Pupil diameter can serve as an index of arousal (Larsen & Waters, 2018; McGinley, David, & McCormick, 2015; McGinley, Vinck, et al., 2015). Larger pupils are associated with increases in the gain, amplitude, signal-to-noise ratio, and reliability of responses in V1 (Reimer et al., 2014). Thus, our finding that increased pupil diameter and locomotion are associated with larger effect sizes could be explained by an increase in response gain or amplitude in V1 that is inherited by downstream AL and AM: since the ND in these areas is selective for naturalistic structure, increased bottom-up drive could accentuate unscrambled-scrambled differences in ND.

Alternatively, response gain or amplitude in higher visual areas could be modulated directly by subcortical arousal systems. The noradrenergic and cholinergic systems are likely candidates, although it is not clear why noradrenergic modulation would cause an effect specific to L2/3; as for cholinergic modulation, Pafundo et al. (2016) showed that V1 and LM are differentially modulated by basal forebrain stimulation such that the response gain and reliability of excitatory L2/3 neurons was enhanced in V1 but not in LM, despite an even distribution of basal forebrain axons across all layers in both areas. However, neuromodulatory regulation of activity in other visual areas, in particular AL and AM, has not yet been characterized in great detail. Another possibility is a top-down effect, where increases in arousal and locomotion reflect increased attentional engagement that favors processing of high-level stimulus features, selectively increasing ND for the unscrambled stimuli. In the passive viewing paradigm employed here, in which the animal is not motivated to attend to the stimuli, the top-down modulation of sensory processing may vary considerably across the experimental session as arousal and attention fluctuate.

Though differentiation analysis revealed area- and layer-specific differences in responses to unscrambled and phase-scrambled stimuli, our ability to decode stimulus category from neural responses was remarkably similar across areas and layers. These findings are consistent with a growing literature that reveals a dissociation between encoding and function (Erlich et al., 2015; Jin & Glickfeld, 2020; Katz et al., 2016; Liu & Pack, 2017; Tsunada et al., 2016; Zatka-Haas et al., 2021). The contrast between our ND and decoding results highlights an important distinction: decoding reveals information content, but this information is necessarily measured from the extrinsic perspective (Buzsáki, 2019; Oizumi et al., 2014; Tononi, 2004; Tononi et al., 2016). The presence of information about a stimulus in a neural circuit does not imply that the information is functionally relevant (Brette, 2019). As an extreme example, stimulus category would presumably be perfectly decodable from photons impinging on the retina, but this would reveal nothing of interest about perception. By contrast, ND is an intrinsic measure defined without reference to a stimulus (Boly et al., 2015; Mensen et al., 2017, 2018). In the brain, a complex evolved system in which activity is energetically costly, ND may be a signature of functionally relevant dynamics. The dissociation we find between ND and decoding indicates that differentiation analysis can point to populations of interest that are not revealed by detecting stimulus information.

Finally, we also found that the predator stimulus and the "mousecam" stimulus elicited significantly higher ND than other unscrambled continuous stimuli. The predator stimulus finding is intriguing because that stimulus has lower luminance, contrast, and spectral energy than the clip of conspecifics in a home cage (Figure C.16); given the importance of detecting natural predators, the high ND evoked by this stimulus may reflect its salience to the visual system, driven by high-level features such as the presence of the predator rather than low-order stimulus statistics. This also demonstrates that differentiation analysis can probe differences in visual responses at the level of individual stimuli.

It is important to note the limitations of these data. First, calcium imaging provides an imperfect proxy for neuronal activity. The fluorescence signal from calcium indicators is more sensitive to bursts of spikes than sparse, low-frequency spiking (Chen et al., 2013; Huang et al., 2021; Ledochowitsch et al., 2019; Pachitariu et al., 2018; Siegle, Ledochowitsch, et al., 2021; Wei et al., 2020). Such sparse activity may contribute to ND but would not be present in this dataset. However, given the typically sparse spiking activity of L2/3 excitatory neurons compared to deeper layers (Barth & Poulet, 2012), it is possible that this limitation only obscures even stronger L2/3 specificity. Second, for this exploratory study we used a range of naturalistic stimuli and a limited number of phase-scrambled control stimuli in order to include diverse high-level features. Future studies could test our findings using a larger set of artificial stimuli controlling for other low-level characteristics, *e.g.,* optical flow, in addition to the power spectrum. Third, the restricted range of the average arousal measures we observed in our L2/3 AL & AM experiments limits the generalizability of the association we observed between effect size and arousal state. Fourth, while we observed medium to very large effect sizes within individual experimental sessions in L2/3 of AL & AM (Cohen's d = 0.57–1.25), the overall effect was relatively subtle (Cohen’s $d = 0.34$) due to variability in ND values across sessions. There was also considerable variability in arousal state and locomotor activity across trials. To the extent that these factors modulate effect size, future work might uncover larger effects by employing an active paradigm where the animal is motivated to attend to the stimuli.

In summary, we measured stimulus-evoked differentiation of neural activity with cellular resolution and found increased ND in response to unscrambled versus scrambled stimuli. This effect was specific to L2/3 excitatory cells in AL and AM and was enhanced at higher arousal levels. To our knowledge, this study is the first to systematically measure stimulus-evoked differentiation with cellular resolution across multiple cortical areas and layers. These results advance our understanding of the functional differences among visual areas, and future work should integrate our findings into the emerging picture of a shallow hierarchy in the mouse visual system, for example by investigating potential differences in neuromodulation among areas or the contrast between AL/AM and PM. Differentiation analysis is motivated by IIT, and provides an intrinsic, “inside-out” analytical approach that complements extrinsic, “outside-in” measures such as decoding performance, which in this dataset did not distinguish specific cell populations. This method can be used to compare individual stimuli and may provide a readout of the degree to which a given stimulus induces a rich and varied perceptual experience. Future studies should investigate stimulus-evoked differentiation with cellular resolution in humans and non-human primates, where subjective reports are available, and thereby determine the contributions of distinct cell populations to ND while correlating ND with phenomenology.

CHAPTER 5

Meaning, perception, and matching

Quantifying how the structure of experience matches the environment

William G. P. Mayner[1,2,☯], Bjørn E. Juel[2,3,☯] Giulio Tononi[2,]

1 Neuroscience Training Program, University of Wisconsin–Madison, Madison, WI, USA

2 Wisconsin Institute for Sleep and Consciousness, University of Wisconsin–Madison, Madison, WI, USA

3 Brain Signalling Group, University of Oslo, Oslo, Norway

☯ These authors contributed equally to this work.

Acknowledgments

We thank Larissa Albantakis and William Marshall for invaluable feedback during the early stages of this project; Leonardo S. Barbosa, Melanie Boly, Armand Mensen, and Shuntaro Sasai for helpful discussions; and Alireza Zaeemzadeh for helpful comments on the manuscript.

5.1 Introduction

Cast a glance at the scene outside the window. In a blink of the eye, you take in the forest, with its intricate canopy of trees. It is usually assumed that, in doing so, a stimulus from the environment impinges on the retina, and the stimulus conveys information that is relayed to the brain; the information is processed through a hierarchy of sensory areas, aided by top-down signals that disambiguate or fill in noisy bottom-up data; and finally, the meaning of the information is decoded, with the ultimate goal of guiding behavior. The very idea of processing suggests that the information is in the stimulus, waiting to be decoded, and that its meaning is the result of that processing. Somewhere along this information processing chain, we happen to perceive the stimulus and become conscious of it—a case of conscious processing.

Integrated information theory (IIT; Albantakis et al., 2022) offers a different approach to what it means to perceive a stimulus, because its starting point is consciousness itself rather than its behavioral, functional, or neural correlates. IIT begins by considering the essential properties of consciousness—those that are immediately given and true of every conceivable experience: every experience is from the intrinsic perspective of the subject (intrinsicality), specific (information), unitary (integration), definite (exclusion), and structured by phenomenal distinctions bound by phenomenal relations (composition). Next, IIT formulates these properties in physical terms, defined as cause-effect power—the ability to take and make a difference. A substrate of consciousness, called a *complex*, is thus a set of units whose cause-effect power is upon itself (intrinsicality), selecting a specific cause-effect state (information), in a way that is irreducible (integration), maximally so among all overlapping substrates (definite), and structured by the cause-effect power of subsets of its units (composition). Finally, IIT argues that the quality and quantity of the experience are accounted for in full by the *cause-effect structure*, or Φ-*structure*, specified by the complex in its state. The Φ-structure is composed by the distinctions (cause-effects) and relations (overlaps among cause-effects) specified by subsets of the complex. It defines the *intrinsic meaning* or information content of the experience regardless of any reference to anything outside the complex (Tononi, forthcoming; Tononi, Albantakis, et al., 2022).

How, then, is the role of environmental stimuli to be understood from the intrinsic perspective taken by IIT? What is required to perceive a stimulus? And how does the intrinsic meaning of experiences refer to regularities in the environment? In this paper, we extend IIT's framework to address these questions.

We present the mathematical formalism in the Theory section. In the Results section, we demonstrate the formalism in a simple model system. First, we briefly summarize how a complex, disconnected from the environment, supports a Φ-structure that corresponds to a “dreaming” experience, which fully specifies its intrinsic meaning (Figure 5.1A). We then connect the complex to the environment, let it quickly settle into a state triggered by a stimulus sampled from that environment, and unfold the Φ-structure specified by the triggered state (Figure 5.1B). We employ the formalism of actual causation (Albantakis et al., 2019), itself based on the principles of IIT, to calculate *connectedness* and the associated, normalized *triggering coefficient*—the extent to which the current state of each subset of the complex is triggered by the stimulus. We can then calculate *perception* as the product of how much each distinction and relation exists intrinsically (its φ value) and how much it is caused by the stimulus (the triggering coefficient), yielding a *perceptual structure*. The intrinsic meaning of the experience is the same whether it occurs spontaneously or is triggered by an external stimulus, implying that every perception is an interpretation. We then calculate *stimulus-specific matching* as the expected difference between the intrinsic meaning triggered by a stimulus (the perceptual structure) sampled from an environment and stimuli sampled from independent noise sources (Figure 5.1C). If matching is positive, intrinsic meaning is attuned to regularities in the environment that are responsible for that stimulus to occur above chance. Finally, we calculate *perceptual differentiation* as the sum of the perception values associated with the set of unique distinctions and relations triggered by a set of stimuli from the environment vs. independent noise sources (*stimulus sequence matching*). If stimulus sequence matching is positive, the connectivity of the complex has internalized different aspects of causal processes in its environment. Matching thus reveals a link between intrinsic meaning and extrinsic reference.

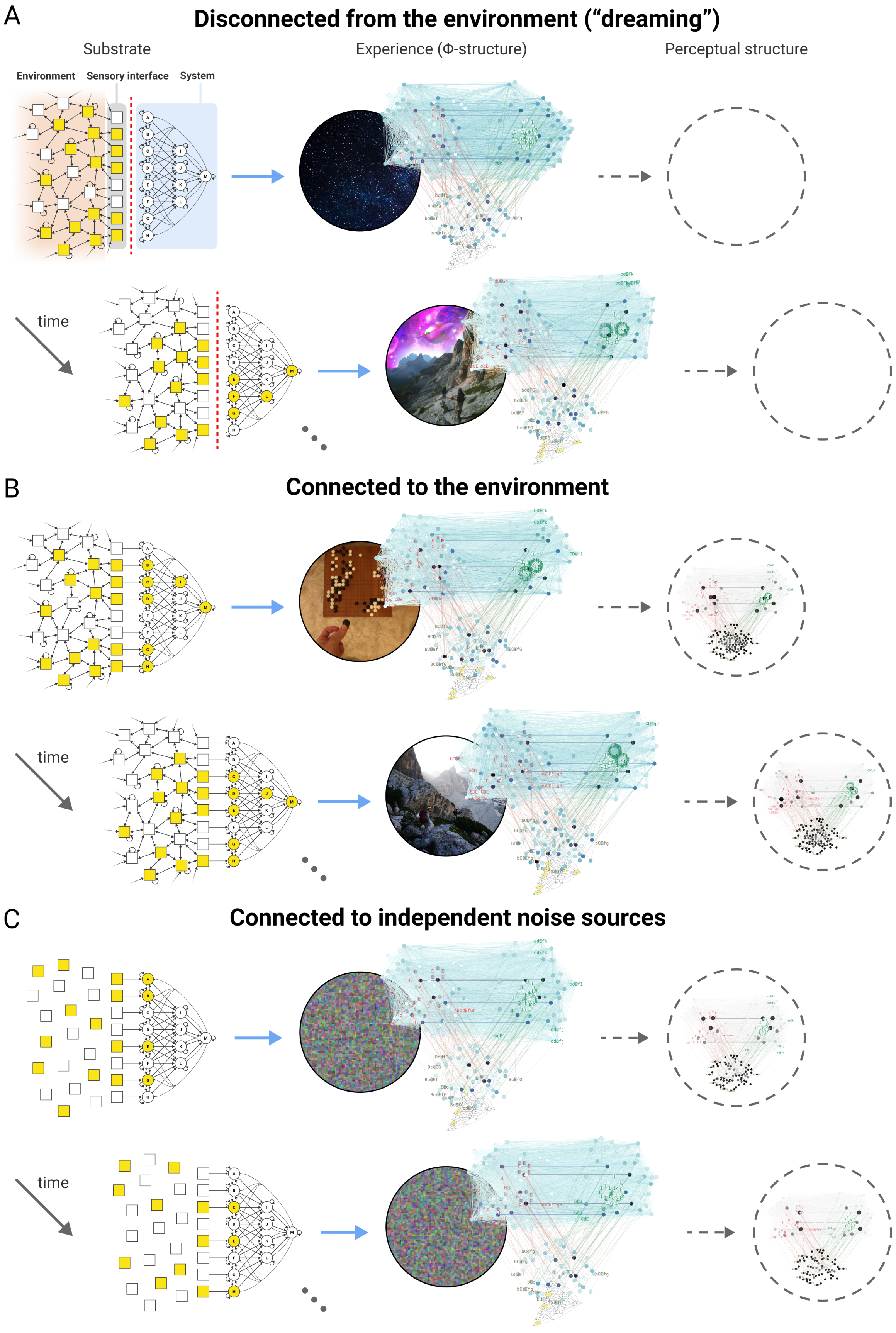

A
Disconnected from the environment (“dreaming”)
Substrate
Experience (Φ-structure)
Perceptual structure
Environment
Sensory interface
System
time
B
Connected to the environment
time
C
Connected to independent noise sources
time

Figure 5.1. Conceptual summary. (A) A simple model system, showing a complex disconnected from the environment (left) that supports Φ-structures corresponding to "dreaming" experiences (middle; Φ-structures are overlaid on suggestive illustrations of visual experiences). None of the distinctions and relations composing the dreamt Φ-structures are percepts, since they were not triggered by the environment—so there is no perceptual structure (right). **(B)** When the complex is connected to the environment, stimuli impinging on the sensory interface trigger system states that likewise support Φ-structures corresponding to experiences. Here, the components of the Φ-structure are specified by subsets of the system whose state was caused by the stimulus (positive triggering coefficient). Those components have a positive perception value, yielding a perceptual structure that captures the extent to which the stimulus is perceived. **(C)** To quantify the matching between the system's subjective experience and objective regularities in the environment, we connect the system to an environment with no regularities—*i.e.*, independent noise sources—and measure the difference between the perceptual differentiation elicited by a stimulus sequence sampled from the environment and that elicited by a random stimulus sequence sampled from noise.

5.2 Theory

The system and its environment

In IIT, physical existence is synonymous with having cause-effect power, the ability to take and make a difference. A physical substrate U with state space Ω_U is operationally defined by its potential causal interactions, assessed in terms of conditional probabilities. Accordingly, as in our prior work (Albantakis et al., 2022), the starting point of our analysis is a stochastic system $U = \{U_1, U_2, \dots, U_n\}$ of n discrete interacting binary units with state space $\Omega_U = \prod_i \Omega_{U_i}$ and current state $u \in \Omega_U$. We denote the complete transition probability function of U over a system update $u \rightarrow \bar{u}$ as

$$\mathcal{T}_U \equiv \Pr(\bar{u} \mid u), \quad \forall\, u, \bar{u} \in \Omega_U. \tag{5.1}$$

We assume that the system state updates in discrete steps, that the state space Ω_U is finite, and that the individual random variables $U_i \in U$ are conditionally independent from each other given the preceding state of U:

$$\Pr(\bar{u} \mid u) = \prod_{i=1}^{n} \Pr(\bar{u}_i \mid u). \tag{5.2}$$

We also assume a complete description of the system, meaning that we can determine the conditional probabilities in (5.2) for every state, with $\Pr(\bar{u} \mid u) = \Pr(\bar{u} \mid \mathrm{do}(u))$, where the "do-operator" indicates that u is imposed by intervention (Albantakis et al., 2019; Ay & Polani, 2008; Janzing et al., 2013; Pearl, 2009). This implies that U is a causal network (Albantakis et al., 2019) and $\mathcal{T}_U$ is a transition probability matrix (TPM) of size $|\Omega_U|$.

In this work, we divide U into two parts: the system in question $S \subseteq U$ and its environment $E = U \setminus S$. We define the *sensory interface* $\partial S \subseteq E$ to be the part of the environment that has an effect on S over one update step, such that the next state of S depends only on its current state and the current state of ∂S.

Experience and intrinsic meaning

In this section we briefly recapitulate IIT's account of consciousness. For a complete description, we refer the reader to (Albantakis et al., 2022).

IIT identifies five essential properties of experience that are immediately given and true of every conceivable experience, termed 'axioms': phenomenal experience is (1) *intrinsic* (it exists *for itself*), (2) *specific* (it is *this one*), (3) *unitary* (it is *a whole*, irreducible to separate experiences), (4) *definite* (it is *this whole*), and (5) *structured* (it is composed of *distinctions* and the *relations* that bind them together, yielding a *phenomenal structure* that feels *the way it feels*). The theory then makes an "inference to a good explanation" for these phenomenal properties by postulating that the substrate of experience must jointly possess certain physical properties that correspond to them. IIT formulates these properties in physical terms as its five 'postulates' (*intrinsicality*, *information*, *integration*, *exclusion*, and *composition*), translates these in mathematical terms, and provides an algorithm for operationally assessing the extent to which a candidate substrate satisfies the postulates Chapter 2. According to IIT, the physical properties characterized by the postulates are necessary and sufficient for a system to be a substrate of consciousness. Furthermore, IIT proposes a fundamental explanatory identity: an experience is identical to the Φ-structure unfolded from a maximal substrate, or complex (defined below). According to this identity, all phenomenal properties of experience must have a good explanation in terms of properties of the corresponding Φ-structure.

Identifying the substrate of consciousness

The first step in the analysis is to identify the substrate of consciousness. This is done by applying the first four postulates:

Intrinsicality: The intrinsicality postulate requires that a system exert cause-effect power *within itself*.

To assess this, we causally marginalize the background units of the system $W = U \setminus S$, conditional on their current state, which renders them causally inert with respect to S.

Information: The information postulate requires that a system's cause-effect power be specific: the system in its current state s must select a specific cause-effect state for its units. This is the state for which *system intrinsic information* ii_s is maximal (Barbosa et al., 2020).

Integration: The integration postulate requires that the system must specify its cause-effect state in a way that is irreducible, *i.e.*, that cannot be reduced to the joint specification of its parts. This is assessed by *partitioning* the system into parts and quantifying how much system intrinsic information is lost due to the partition; *system integrated information* φ_s is then evaluated as the amount lost over the partition that makes the least difference (Marshall et al., 2023).

Exclusion: Finally, the exclusion postulate requires that the substrate of consciousness must be constituted of a definite set of units, neither less nor more. Moreover, the units and updates of the substrate must have a definite grain. Since multiple overlapping subsets of U may have a positive value of φ_s, exclusion is enforced operationally by selecting the set of units that maximizes φ_s over itself. This set of units is called a *maximal substrate* or *complex*.

Again, we refer the reader to Albantakis et al. (2022) for definitions of these quantities and a full description of the formalism.

Unfolding the *Φ*-structure and characterizing intrinsic meaning

Once a complex has been identified, we apply the postulate of composition. This requires that we characterize the complex's *cause-effect structure*, or Φ-*structure*, [1] by considering all its subsets and *unfolding* its cause-effect power. We denote the *Φ*-structure of a complex S in state y as $C(y)$.

To contribute to the *Φ*-structure of a complex, a system subset must both take and make a difference within the system. A subset $M \subseteq S$ in state $m \in \Omega_M$ is called a *mechanism* if it *links* a cause and effect state over subsets of units $Z_\mathrm{c} \subseteq S$ and $Z_\mathrm{e} \subseteq S$, called the cause and effect *purviews*. A mechanism together with its cause and effect is called a *causal distinction*, denoted $d(m)$. Distinctions are evaluated

[1] We use "cause-effect structure" to refer to the unfolded cause-effect power of a given system, whether or not it has been identified as a complex. If the system is a complex, then we use the term "*Φ*-structure."

based on whether they satisfy the postulates of IIT (except for composition, as distinction are themselves components). Briefly, for each candidate purview $Z_{c/e} \subseteq S$, the cause and effect states z_c and z_e are those for which the mechanism specifies maximal intrinsic cause information ii_c and intrinsic effect information ii_e, respectively (Barbosa et al., 2020). Integration is then assessed for a candidate purview by evaluating all partitions of the mechanism and purview, where the associated integrated information $\varphi_{c/e}$ is the amount of $\text{ii}_{c/e}$ lost over the partition that makes the least difference. Next, exclusion is enforced by selecting from among the candidate purviews $Z_{c/e}$ the cause and effect purview that respectively maximize the φ_c and φ_e values, yielding the cause purview state $z_c^*(m)$ and effect purview state $z_e^*(m)$. The integrated information of the distinction is the minimum of the cause and effect integrated information, $\varphi_d(m) = \min(\varphi_c(m), \varphi_e(m))$, which quantifies its irreducibility. Finally, by the information postulate, the distinctions that exist for the complex are only those whose cause-effect state is congruent with that of the complex as a whole.

A set of distinctions $\mathbf{d}$ are *bound* together by a *causal relation* $r(\mathbf{d})$ if the cause-effect state of each distinction $d \in \mathbf{d}$ overlaps congruently over a shared set of units, called the *relation purview*, which may be part of the cause, effect, or both the cause and effect of each distinction. For a given set of distinctions, there are potentially many "relating" sets of causes and/or effects $\mathbf{z}$ with non-empty intersection. These specify unique aspects about the relation $r(\mathbf{d})$ and constitute its *faces* $\mathbf{f}(\mathbf{d})$ (by analogy to the faces of a simplex). For a given face $f \in \mathbf{f}(\mathbf{d})$ the set $\mathbf{z}$ is called the *face purview*; the union of the face purviews forms the relation purview. A relation $r(\mathbf{d})$ that binds together $h = |\mathbf{d}|$ distinctions is called an *h-degree relation*, and a relation face $f(\mathbf{d}) \in \mathbf{f}(\mathbf{d})$ over $k = |\mathbf{z}|$ cause/effect purviews is called a *k-degree face*. Briefly, the irreducibility $\varphi_r(\mathbf{d})$ of a causal relation is measured by "unbinding" distinctions from their joint purviews, taking into account all faces of the relation, as follows. For each distinction $d \in \mathbf{d}$, the average integrated information φ_d per unique distinction purview unit is multiplied by the number of units in the relation purview. The minimum of this value across distinctions is the relation's irreducibility $\varphi_r(\mathbf{d})$, corresponding to the integrated information lost by partitioning that distinction from the relation.

The Φ-structure of the complex in its current state is composed of these distinctions and relations. By the fundamental explanatory identity of IIT, these account in full, with no additional ingredients, for the quality or feeling of an experience, which is the same as its meaning ("the meaning is the feeling"). The sum of integrated information values φ of all the distinctions and relations that compose the Φ-structure, called *structure integrated information*, denoted Φ, corresponds to the quantity of consciousness.

Distinction Φ-folds

Here we introduce a convenient decomposition of the Φ-structure $C(y)$ into sub-structures, called Φ-folds, associated with each distinction.

For a given mechanism $M \subseteq S$ in state m that specifies a distinction $d(m)$, we call the sub-structure consisting of that distinction and all relations involving it the *distinction* Φ*-fold* of $d(m)$, denoted $C(d(m))$, or simply $C(d)$ when d ranges over distinctions. If the Φ-structure is thought of as a hypergraph, with distinctions as vertices and relations as edges, then a distinction Φ-fold corresponds to a single vertex and its incident edges.

We can define a quantity $\Phi_d(C(d(m)))$ that captures the contribution of the mechanism m to the total $\Phi(C(y))$ as

$$\Phi_d(C(d(m)))) \;=\; \sum_{c \,\in\, C(d(m))} \frac{\varphi_c}{|c|}, \tag{5.3}$$

where for a distinction $c = d(m)$, $|c| = |d(m)| = |\{m\}| = 1$, and for a relation $c = r(\mathbf{d})$, $|c| = |r(\mathbf{d})| = |\mathbf{d}| > 1$.

Note that this expression counts the entire φ_d value of the distinction, but a fraction of the φ_r of relations involving $d(m)$. This assumes that a relation's integrated information φ_r is distributed uniformly across the distinctions it binds together. Since the relation is an irreducible component of the Φ-structure and removing any one of the distinctions would "unbind" them, we consider the contribution of each $d \in \mathbf{d}$ to the relation's φ_r value to be $\varphi_r / \, |\mathbf{d}|$.

We then have that the total $\Phi(C(y))$ is partitioned by the Φ_d values:

$$\Phi(C(y)) \;=\; \sum_{d \,\in\, C(y)} \Phi_d(C(d)). \tag{5.4}$$

In words, the Φ value of the entire Φ-structure can be expressed as the sum of the Φ_d values of the distinction Φ-folds specified by each of the system's irreducible mechanisms.

Perception

Connectedness & the triggering coefficient

Let p be the conditional probability of $M = m$ at time t given $\partial S = x$ at time $t - \tau$,

$$\Pr(M_t = m \mid \partial S_{t-\tau} = x),$$

and let q be the marginal probability of $M = m$ at time t,

$$\Pr(M_t = m).$$

We define the connectedness $c(x, m)$ of a set of units $M \subseteq S$ in state m at time t to the sensory interface ∂S in state x at $t - \tau$ as

$$c(x,m) \;=\; \begin{cases} \log_2(p/q) & \text{if } p > 0,\ q > 0, \text{ and } p \geq q, \\ 0 & \text{otherwise.} \end{cases} \tag{5.5}$$

This definition is based on the measure of actual effect information developed in Albantakis et al. (2019), although here we do not consider the question of which subset of the sensory interface caused the current state of the substet ("what causes what").[2] The expression in the first case is also known as the pointwise mutual information (PMI), and is positive when the stimulus raises the probability of the mechanism's state occurring compared to when the influence of the stimulus is ignored.[3]

Connectedness $c(x, m)$ is maximized when $p = 1$, *i.e.*, when the stimulus x causes the mechanism state m deterministically. It is therefore bounded by the self-information of the mechanism state:

$$c(x,m) \;\leq\; \log_2(1/q). \tag{5.6}$$

[2] To exemplify, consider a subset M of eight binary units such that a single stimulus $\partial S = x$ fully determines the state $M = m \in \{0,1\}^8$ (so that $p = 1$), and m never occurs in response to any other stimulus (so that $q = 1/2^{|\partial S|}$). If the sensory interface also consists of eight binary units, then $c(x, m) = 8$ bits. If $|\partial S| = 8$ but $|M| = 1$, we likewise have $c(x, m) = 8$ bits. But if $|\partial S| = 1$ and $|M| = 8$, then $c(x, m) = 1$ bit. In general, if m always and only occurs in response to a single stimulus, connectedness is limited by the information specified by $\partial S = x$. Conversely, if $p = 1$ but m occurs in response to other stimuli, then $q > 1/2^{|\partial S|}$, and the connectedness value is instead limited by the information specified by $M = m$.

[3] If $p < q$, the PMI is negative. In this case, we take the stance that the environment did not bring about the state $M = m$ and define connectedness to be zero, in line with the actual causation framework (Albantakis et al., 2019), rather than

We use this bound to define a normalized form of connectedness, called the *triggering coefficient*:

$$t(x,m) \;=\; \frac{c(x,m)}{\log_2(1/q)}, \tag{5.7}$$

so that $0 \leq t(x,m) \leq 1$. When the mechanism m specifies a distinction d, we will also refer to its triggering coefficient as $t(x,d)$ or $t(x,d(m))$.

The triggering coefficient expresses the extent to which the stimulus $\partial S_{t-\tau} = x$ caused $M_t = m$ relative to how strong that cause could have been, taking into account the size of the mechanism as well as the system's intrinsic connectivity. While the connectedness value is dependent on the size of the subsets (larger subsets can specify more information), the triggering coefficient is not. A value of $t(x,m) = 1$ indicates that $M_t = m$ can be fully attributed to $\partial S_{t-\tau} = x$, irrespective of the amount of information specified by $M_t = m$, while $t(x,m) = 0$ indicates the stimulus had no role in bringing about $M_t = m$.[4]

The triggering coefficient also permits the unbiased comparison of different subsets by accounting not only for subset size,[5] but also for the role of the system's intrinsic connectivity. In general, subsets that are more directly connected to the sensory interface will be more strongly connected to the environment, while internal subsets will tend to be affected more by intrinsic dynamics. Normalizing by the self-information of the mechanism state, $\log_2(1/\Pr(m))$, discounts such differences. For example, a neuronal assembly deep in the brain whose state is determined largely by intrinsic dynamics, but which is nonetheless reliably triggered by the stimulus, will be evaluated on the same terms as an assembly in earlier sensory areas.

considering this a "preventative effect" on the subset state (Korb et al., 2011). In an information-theoretic context, this measure has been termed the positive PMI (Dagan et al., 1993).

[4] There are other possible normalizations for the connectedness value. The positive PMI is also bounded by the self-information of the stimulus, $-\log_2(\Pr(x))$; the minimum of both self-information values, $\min(-\log(\Pr(x)), -\log(\Pr(m)))$; and the joint self-information, $-\log_2(\Pr(x,m))$. Normalizing by each of these bounds respectively yields a measure of (1) non-degeneracy, *i.e.* the selectivity of the mechanism in responding to the stimulus (maximized when $\Pr(x \mid m) = 1$); (2) either determinism or non-degeneracy (maximized when either $\Pr(x \mid m) = 1$ or $\Pr(m \mid x) = 1$); and (3) determinism and non-degeneracy, *i.e.* perfect co-occurrence (maximized when both $\Pr(x \mid m) = Pr(m \mid x) = 1$). Bouma (2009) noted these options and investigated the latter in the context of linguistics, calling it the normalized PMI (Bouma, 2009). In our case, by contrast, the measure should be sensitive to determinism but not to degeneracy (when $M_t = m$ also occurs in response to other stimuli, raising the marginal probability q). Given a stimulus and a response that actually occurred, the triggering coefficient quantifies the causal role of the stimulus in producing the response regardless of whether other stimuli might have also caused the response counterfactually.

[5] This is crucial because, for IIT, what matters is only whether a subset has irreducible cause-effect power within the system, not how many units it contains.

Note that because we take the intrinsic perspective of the system under analysis, we assume the uniform distribution over stimuli, $\Pr(x) = 1/\,|\Omega_{\partial S}|$, rather than the observed distribution (Albantakis et al., 2022). Note also that τ, the delay at which the stimulus' effect is evaluated, is a free parameter and should be chosen to maximize the efficacy of the stimulus. In practice, the appropriate choice of τ will depend on the experimental setup.

Perception value & perceptual richness

We define the *perception value* of a distinction $d(m)$ as

$$p(x, d(m)) \;=\; t(x, m)\; \varphi_d(m). \tag{5.8}$$

In words, a distinction is perceived to the extent that it was triggered by the stimulus. There is no perception when either the subset of the system m is reducible and does not specify a distinction ($\varphi_d(m) = 0$), or the when state of the subset was not caused by the stimulus ($t(x, m) = 0$). Note that because $t(x, m) \leq 1$, the maximum perception value for a distinction is its φ_d value.

This approach can be straightforwardly extended to the case of relations. Since each relation $r(\mathbf{d})$ binds together several distinctions $\mathbf{d}$, we can define the triggering coefficient for a relation $t(x, r(\mathbf{d}))$ as a weighted average of the triggering coefficients of each $d \in \mathbf{d}$. To determine the weights, we use the same rationale as for the definition of the Φ_d of a distinction Φ-fold: since the relation is irreducible, we assume that its integrated information $\varphi_r(\mathbf{d})$ is distributed uniformly across the distinctions it binds together (as removing any one of them would "unbind" all of them). Thus, each mechanism's triggering coefficient is given equal weight in the average:

$$t(x, r(\mathbf{d})) \;=\; \frac{1}{|\mathbf{d}|} \sum_{d \in \mathbf{d}} t(x, d). \tag{5.9}$$

The perception value of a relation $r(\mathbf{d})$ with respect to a stimulus x is then defined as

$$p(x, r(\mathbf{d})) \;=\; t(x, r(\mathbf{d}))\; \varphi_r(\mathbf{d}). \tag{5.10}$$

Note that for distinction $\varPhi$-folds, this implies

$$\sum_{c \in C(d(m))} p(x,c) \;=\; t(x,m)\; \varPhi_d(C(d(m))). \tag{5.11}$$

Eqs. (5.8) and (5.10) can be combined into a general expression for an arbitrary component c of the $\varPhi$-structure (distinction or relation):

$$p(x,c) \;=\; t(x,c)\frac{\varphi_c}{|c|}, \tag{5.12}$$

where $|c|$ denotes the number of distinctions involved in the component, as in (5.3).

Next, we consider the relationship of the stimulus to the $\varPhi$-structure as a whole. Analogously to $\varPhi$, we define the total *perceptual richness* to be the sum of the perception values of the components in the specified $\varPhi$-structure:

$$\mathcal{P}(x,y) \;=\; \sum_{c \in C(y)} p(x,c). \tag{5.13}$$

For a distinction $\varPhi$-fold, we denote the sum of perception values as $\mathcal{P}_d(x,m)$. Note that by Eqs. (5.3) and (5.12), perceptual richness can be partitioned into the $\varPhi_d$ values of each distinction $\varPhi$-fold in $C(y)$, weighted by their triggering coefficients:

$$\mathcal{P}(x,y) \;=\; \sum_{m \subseteq y} \mathcal{P}_d(x,m) \;=\; \sum_{d \in C(y)} t(x,d)\, \varPhi_d(C(d)) \tag{5.14}$$

Perceptual richness quantifies the extent to which a $\varPhi$-structure is a percept. Said otherwise, it quantifies the extent to which the experience of a complex in a state—which defines the intrinsic meaning for the complex—was caused by an external stimulus, and can thus be said to refer to it (Tononi, forthcoming; Tononi, Albantakis, et al., 2022).

Matching

Stimulus-specific matching

To define *stimulus-specific matching*, we introduce an environment N comprising $|\partial S|$ independent noise sources. Each noise source provides input to a distinct unit in ∂S, so that stimuli are drawn uniformly at random from all possible stimuli that could impinge on the system; *i.e.*, every stimulus or sequence of stimuli is as likely as any other. We refer to this environment with the random variable N, for "noise".

Given a stimulus $\partial S = x$ produced by the actual environment E that impinges on the system at time $t - \tau$, we define stimulus-specific matching as

$$m(x, S) \;=\; \mathbb{E}\left[\, \mathcal{P}(x, Y) - \mathcal{P}(N, Y') \,\right], \tag{5.15}$$

where $\mathbb{E}$ denotes the expected value, Y denotes the response of the system to x, and Y' denotes the response of the system to a stimulus sampled from N.[6] In words, stimulus-specific matching is the expected difference between the perceptual richness elicited by a given stimulus and that elicited by a stimulus drawn uniformly at random.

Differentiation (stimulus sequence matching)

We now extend our analysis to a sequence of k stimuli $x(1, k) = (x_1, x_2, \ldots, x_i, \ldots, x_k)$ that triggers the sequence of system states $Y(1, k) = (Y_1, Y_2, \ldots, Y_i, \ldots, Y_k)$.

First we define the *differentiation* elicited by $x(1, k)$ given a particular system response sequence $Y(1, k) = y(1, k)$. We count every unique component c that appears in the associated set of $\varPhi$-structures; that is, we consider the union of the $\varPhi$-structures associated with $y(1, k)$, called the *differentiation structure*:

$$C_{\mathcal{D}}\left(y(1, k)\right) \;=\; \bigcup_{i=1}^{k} \left\{ c \mid c \,\in\, C(y_i) \right\}. \tag{5.16}$$

The amount of *(structural) differentiation* is then defined analogously to $\varPhi$ as

$$\mathcal{D}\left(y(1, k)\right) \;=\; \sum_{c \,\in\, C_{\mathcal{D}}(y(1,k))} \varphi_c. \tag{5.17}$$

We can then define the *evoked differentiation capacity* of a system as the total differentiation that elicited by all possible stimuli:

$$\overline{\mathcal{D}}(S) \;=\; \mathcal{D}(\{y \in \Omega_S \mid \exists x \in \Omega_{\partial S} \text{ such that } \Pr(y \mid x) > 0\}) \tag{5.18}$$

[6] In general, Y_t depends on both x_t and the previous system state Y_{t-1}. This definition leaves open the choice of distribution

where Ω_S denotes the set of all possible system states (note that order does not matter here).

We can also define the *intrinsic differentiation capacity* of a system as the differentiation of all its possible states (including those that cannot be evoked by stimuli), which is the maximum differentiation the system can attain:

$$\mathcal{D}^{\max}(S) \;=\; \mathcal{D}(\Omega_S). \tag{5.19}$$

We define the *perceptual differentiation* triggered by $x(1,k)$ similarly to differentiation, using perception values in place of φ values. Since for each occurrence of a given component $c \in C_{\mathcal{D}}\left(y(1,k)\right)$ the triggering coefficient will generally differ depending on the stimulus x_i, the perception value of a component is taken to be its maximum perception value across all i:

$$\mathcal{D}_p\left(x(1,k), y(1,k)\right) \;=\; \sum_{c\,\in\, C_{\mathcal{D}}(y(1,k))} \max_{i=1}^{k}\; p\left(x_i, c_i\right), \tag{5.20}$$

where c_i denotes the component as it occurs in $C(y_i)$. The *evoked perceptual differentiation capacity* is then defined as

$$\overline{\mathcal{D}}(\partial S, S) \;=\; \mathcal{D}_p\Big(\,\{(x,y) \mid x \in \Omega_{\partial S},\; y \in \Omega_S \text{ such that } \Pr(y \mid x) > 0\}\,\Big) \tag{5.21}$$

By considering the the union of the perceptual structures elicited by each stimulus, perceptual differentiation measures the extent to which the sequence not only elicits high perceptual richness, but a varied sequence of percepts.

Finally, we define the *stimulus sequence matching* as

$$\mathcal{M}\left(x(1,k), S\right) \;=\; \max_{a,b\,\in\,(1,\ldots,k),\; a<b} \;\mathbb{E}\left[\,\mathcal{D}_p\left(x(a,b), Y(a,b)\right) - \mathcal{D}_p\left(N(a,b), Y'(a,b)\right)\,\right], \tag{5.22}$$

In words, stimulus sequence matching is the maximum expected difference between the perceptual differentiation elicited by a stimulus sequence sampled from the environment and that elicited by a random sequence of stimuli sampled from independent noise sources, where the maximum is taken

for Y_{t-1}. In this work, we use the uniform distribution, but the observed distribution can also be used; in practice this should be determined based on experimental design considerations.

over contiguous subsequences.[7] This isolates the portion of perceptual differentiation that is due to the causal processes in E that produced the stimulus sequence by subtracting out the portion due to the intrinsic structure of the system (in the sense that it occurs regardless of environmental structure, *i.e.*, in response to N). The use of perceptual differentiation (rather than, *e.g.*, the sum of the perceptual richness) allows the measure to distinguish between systems that have incorporated more or less of the regularities in E, all else being equal.

The reasoning for taking the maximum over contiguous subsequences is as follows. Since every possible stimulus has nonzero probability in N, as the sequence length k grows, the probability that all possible stimuli occur in $N(1,k)$ approaches 1, so that

$$\lim_{k \to \infty} \mathcal{D}_p \left(N(1,k), Y'(1,k) \right) = \overline{\mathcal{D}_p}(\partial S, S). \tag{5.23}$$

That is, the differentiation evoked by random stimuli approaches the system's evoked perceptual differentiation capacity. This implies that

$$\lim_{k \to \infty} \mathbb{E} \left[\mathcal{D}_p \left(x(1,k), Y(1,k) \right) - \mathcal{D}_p \left(N(1,k), Y'(1,k) \right) \right] \leq 0. \tag{5.24}$$

(If every possible stimulus also has nonzero probability in E, then (5.24) becomes an equality.) Therefore, for completeness, formally the maximum is taken over contiguous subsequences so that the measure behaves as intended even for large k. However, since the probability of a given stimulus being sampled from N decreases exponentially as $|\partial S|$ increases, this limit is approached very slowly; for sequence lengths of any practical relevance, the maximum will likely be attained with the full sequence.

In general, matching will be positive when (1) the system is a complex, *i.e.* a maximal substrate that specifies a Φ-structure; (2) the causal processes in the environment impose nonuniform statistical structure on the stimuli impinging on ∂S, and (3) the system's intrinsic causal structure is such that likely environmental stimuli elicit greater perceptual richness than unlikely stimuli.

[7] Note that we do not treat the stimulus sequence as a random variable. Instead we define $\mathcal{M}$ in terms of a particular sample $x(1,k)$ from the environment and leave open the characterization of its statistical properties, because we take the stance that in general the environment is not stationary and cannot be described by a single distribution.

5.3 Results

As a proof of principle, we demonstrate the formalism introduced above in a simple *in silico* model of a system in an environment. After describing the model system, we apply the IIT analysis (Albantakis et al., 2022) to unfold its intrinsic cause-effect structure—its Φ-structure—which fully characterizes the intrinsic meaning specified by the system in its current state. Next, we demonstrate how this formalism characterizes perception: the extent to which parts of the system's Φ-structure are triggered by an external stimulus. Finally, we use the perception formalism to characterize the matching between the system's intrinsic causal structure and causal processes in the environment.

The substrate model

The substrate model U consists of a system of three hierarchical levels of stochastic binary units with distinct, state-dependent activation functions (loosely inspired by the mammalian visual system) and an environment E, which impinges on the system at the lowest level via the sensory interface $\partial S \subseteq E$.[8] We chose to use a small model that can be fully characterized by a TPM of tractable size in order to permit exact computation of the quantities in the formalism. Here, we briefly describe its main properties (for details, see Appendix D).

In brief, the model system is a neural network designed to detect the presence of a single 'segment' stimulus—three adjacent 'ON' units surrounded by 'OFF' units (01110)—anywhere on the sensory interface. Level 1 of the model system comprises 'lattice' units that receive a bottom-up input, each from a distinct unit in the sensory interface, lateral input from adjacent lattice units, and top-down input from the top level. These units are strongly driven by the bottom-up input whenever the input differs from their current state (*e.g.* when the input is 'ON' and the lattice unit is 'OFF'). Level 2 is composed of 'configuration detectors,' each of which receives bottom-up input from the five lattice units below it as well as input from the top level. The configuration detectors are activated with high probability when the level 1 inputs in their receptive field (RF) form the configuration 01110, and are otherwise inactive with high probability. Level 3 consists of a single 'segment' unit that receives bottom-up input

[8] In this work, we model the environment implicitly as a probability distribution over the states of the sensory interface. However, the results do not depend on modeling the environment in this way, and in general, we take the stance that the environment is not stationary and cannot be adequately described by a single distribution. In fact, the formalism should be applied to sequences of stimuli that are actually sampled from the environment, making it unnecessary to model the distribution over states of the sensory interface in practice.

from the four configuration detectors and is activated strongly whenever exactly one of them is active, thus detecting the presence of a segment on the sensory interface. Fig. 5.2A shows a schematic of the system; the activation functions of each class of units are shown schematically in Fig. 5.2B. Every unit also features a self-connection which tends to stabilize the unit's state.

In the simulations, *in silico*, a stimulus is presented to the system by clamping the sensory interface to a fixed state for $\tau = 5$ time steps and allowing its effects to percolate through the system.[9] While the system state stabilizes, driven by bottom-up inputs, short-term plasticity changes the strength of the interactions along lateral, top-down, and self connections, with the net effect that the intrinsic connectivity is strengthened. However, in line with some empirical data (Chance et al., 1999; Douglas et al., 1995; L.-y. Li et al., 2013; Y.-t. Li et al., 2013; Lien & Scanziani, 2013; Peron et al., 2020; Zerlaut et al., 2019), the strengthened intrinsic connectivity 'endorses' the activity pattern triggered by the stimulus, rather than allowing lateral or top-down signals to 'override' it. To achieve this "*intrinsic connectivity endorsement*" (*ICE*), the sign of the interaction—whether positive (excitatory) or negative (inhibitory)—is set such that the input reinforces the unit's current state (Appendix D). For example, if a unit's input is 'ON' and the unit's state is also 'ON', the interaction becomes strongly positive. On the other hand, if a unit's input is 'ON' and the unit state is 'OFF', then the interaction becomes negative, as in lateral inhibition. The strength of the interaction differs for the four combinations of input state and unit state, reaching the highest strength when both the input and the unit are 'ON' (reminiscent of the opening of NMDA receptors in cortical synapses; Larkum et al., 2009). This internalizes the co-activation of neighboring units, which occurs more frequently than chance owing to the smoothness of visual inputs (Field, 1987; Hubel & Wiesel, 1962), in line with the principle that "what fires together wires together" (Hebb, 2002).

An example of the model's dynamics when presented with stimuli is shown in Figure 5.2C. The level 1 units relay the stimuli from the sensory interface to the level 2 configuration detectors, which preferentially activate when they receive the input configuration 01110. The level 3 segment unit then activates when one of the configuration detectors is active, indicating the presence of a segment on the sensory interface.

[9] Note that the sensory surface state need not be fixed, but to simplify the computations, we used fixed state for the duration of the delay τ so as to minimize the number of possible stimuli. For a given stimulus x, we denote this by $\partial S_{t-\tau} = x$.

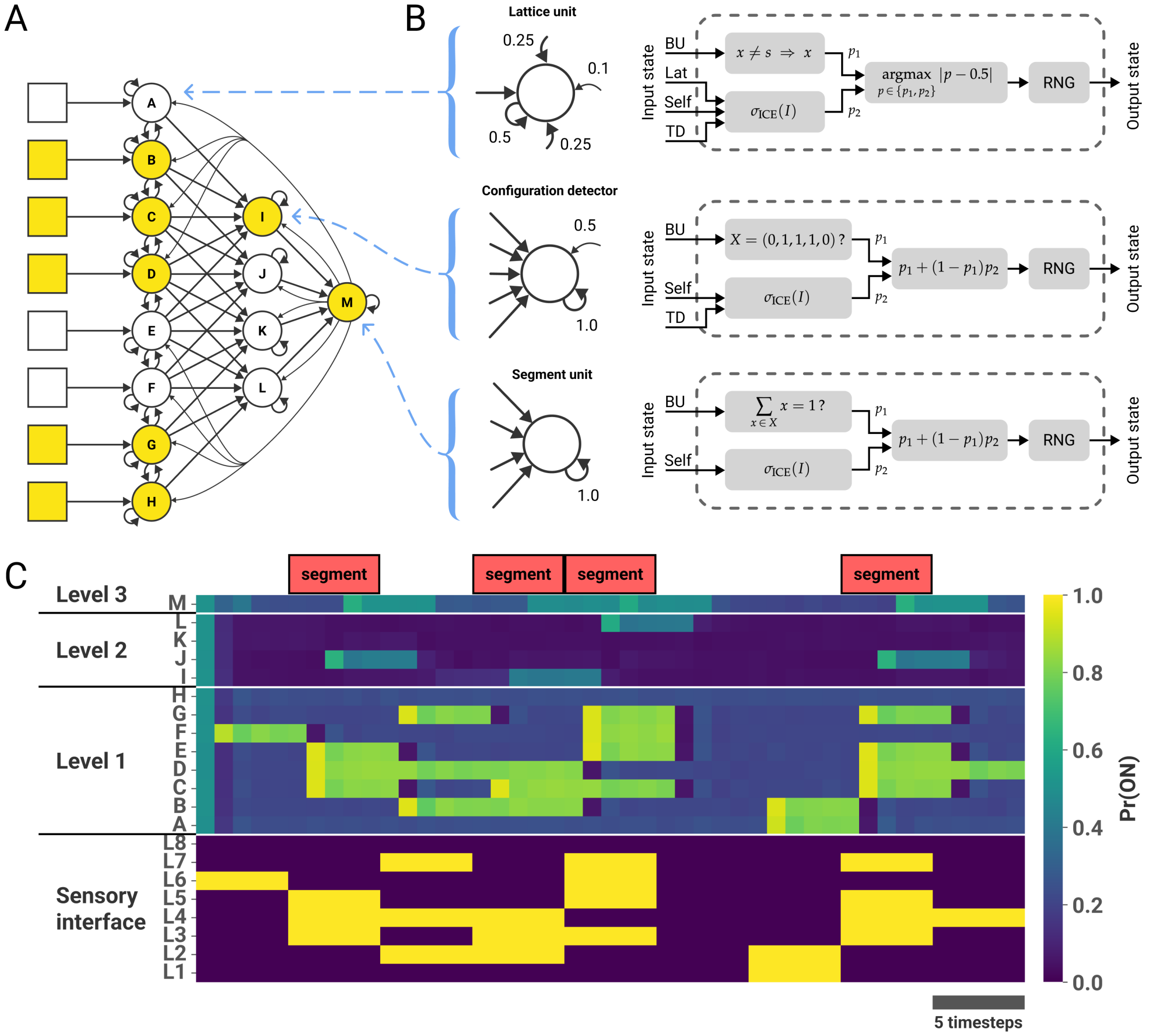


Figure 5.2. The substrate model and its dynamics. **(A)** The model system consists of stochastic binary units (labeled circles). Yellow indicates the 'ON' state ('1'), white indicates 'OFF' ('0'). Units are arranged in hierarchical levels in a rough analogy to the visual system. The sensory interface (squares) impinges on level 1 via bottom-up connections, which are shown disconnected here ('dreaming' condition). Level 1 (*A*–*H*) consists of 'lattice' units that are strongly driven by the sensory input. Level 2 consists of 'configuration detector' units (*I*–*L*), which are selective for the 01110 configuration of their receptive field (their five inputs from level 1). Level 3 consists of a single 'segment' unit M that activates when one of the level 2 units detects the 01110 configuration. All units are nondeterministic and feature a self-connection that reinforces the current state. The activation functions of all units are state dependent. The state-dependency is such that the input to each unit is strengthened in a way that reinforces or 'endorses' the unit's current state, and can be thought of as implementing short-term synaptic plasticity. **(B)** Schematic depiction of the units at each level and their activation functions. Bottom-up connections (BU) are shown as straight arrows pointing right, lateral connections (Lat) are shown as curved arrows pointing up and down, the self-connection is shown as a looping arrow, and top-down connections (TD) from the segment unit are shown as curved arrows pointing left. Numbers on lateral and top-down connections indicate the connection strength used in the sigmoidal module σ_{ICE} of each activation function. To the right of the unit diagrams, activation functions are represented schematically as compositions of simpler functions. For details, see Appendix D. **(C)** Example of the model's dynamics, showing the probability of observing an 'ON' state for each unit in a sample of 5000 simulations in which the same sequence of stimuli was presented to the model. Each stimulus was presented for 5 timesteps and 4 of the stimuli contained 'segments' (highlighted in the top row). As activity percolates upwards, the level 1 lattice units are strongly driven by the bottom-up inputs and reproduce the input pattern with high probability; level 2 configuration detectors activate in the presence of a segment in their receptive field; and the level 3 segment unit M preferentially activates in the presence of a segment anywhere on the sensory interface.

Intrinsic meaning: unfolding the *Φ*-structure of the system

We first unfold the *Φ*-structure of our system in a 'dreaming' condition, *i.e.*, while it is disconnected from the environment, to emphasize that intrinsic meaning is fully accounted for by the distinctions and relations composing the *Φ*-structure and does not require an external referent (Tononi, forthcoming; Tononi, Albantakis, et al., 2022). Fig. 5.3A shows the results of this analysis for an example state (all units 'OFF', inset). The *Φ*-structure is composed of 80 distinctions with 21127 relations among them, with structure integrated information $\Phi = 106.74$.

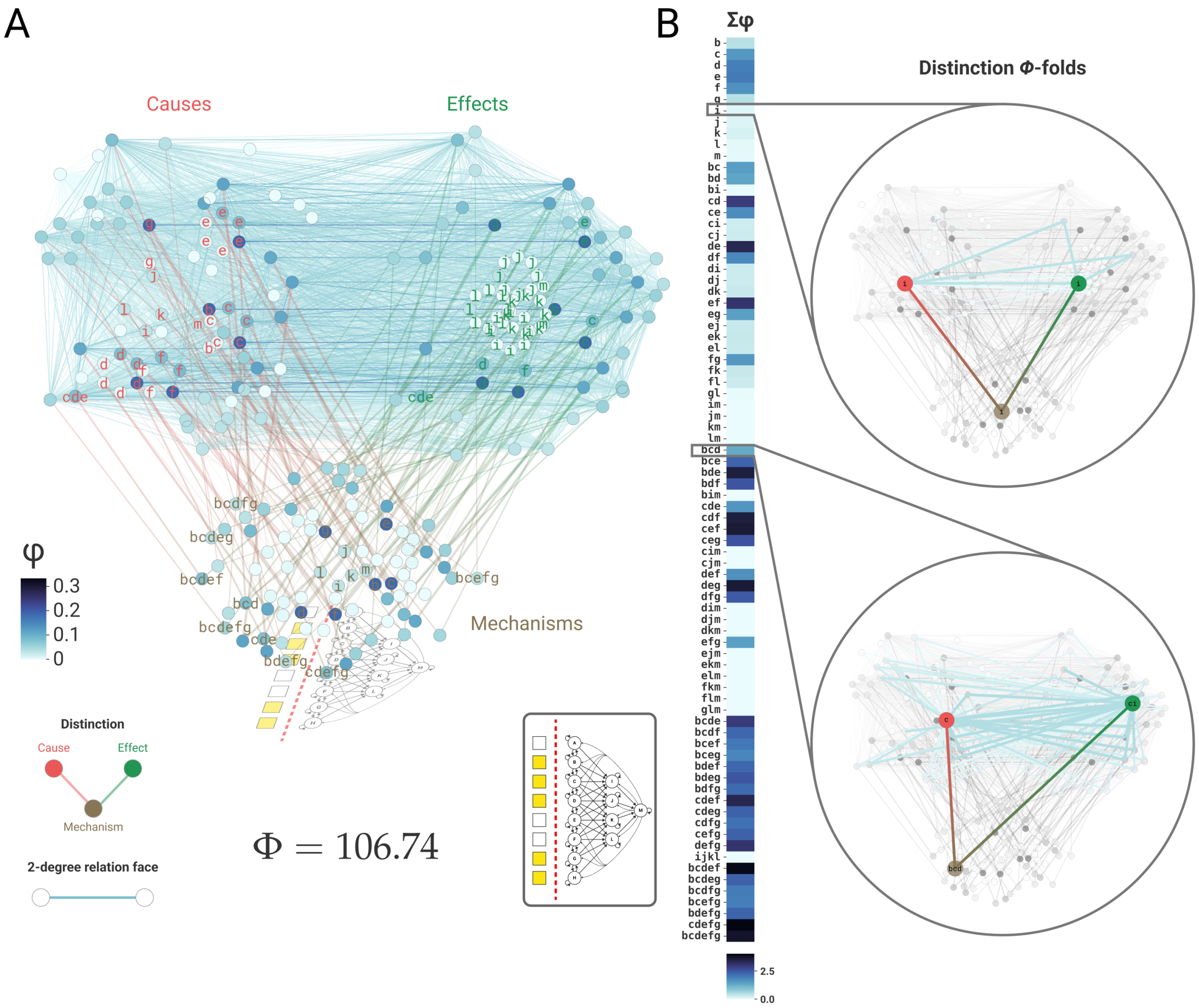


Figure 5.3. A *Φ*-structure and its intrinsic meaning in a disconnected state. (A) Illustration of the *Φ*-structure specified by the system in its current state (inset), when disconnected from the environment. Subsets of the system (*mechanisms*) that specify an irreducible cause-effect pair (a *distinction*) are plotted as the lowest cluster of points, superimposed on the substrate. The causes of each distinction are shown as points in the top left cluster with purview labels in red; effects are plotted on the top right with purview labels in green (only some labels are shown, for legibility). Red and green lines link the mechanisms to the cause and effect pair they specify. Distinctions are bound together by irreducible *cause-effect relations*, which comprise sets of causes and/or effects that congruently specify common purview units (*relation faces*). Relation faces consisting of two cause and/or effect purviews (2^{nd}*-degree faces*) are plotted as lines between the points corresponding to those purviews. Higher degree faces are not shown; this visualization shows only a low-dimensional aspect of the full structure. Mechanisms, purviews, and 2^{nd}-degree relation faces are colored according to their *φ* value. **(B)** Heatmap showing the sum of *φ* values for the distinction *Φ*-fold stemming from each distinction in the *Φ*-structure. *Insets:* Example distinction *Φ*-folds superimposed on the background of the full structure in gray. The mechanism (brown) is shown connected to its cause purview (red) and effect purview (green). Relations in the *Φ*-fold (those involving either the distinction's cause or its effect) are colored as in **A**.

Fully unfolding the *Φ*-structure is computationally infeasible for even this small model system (Mayner et al., 2018), so for these analyses we only computed a representative subset of the distinctions and relations specified by the system (see §5.2). Furthermore, the visualizations of Φ-structures only plot the mechanisms & distinctions (as points) and the 2^{nd}-degree relation faces that bind two distinctions together (as lines), since these parts of the structure are amenable to two-dimensional representation. However, the Φ-structure is high-dimensional and also comprises many higher-degree relation faces. Finally, for the purposes of this demonstration we assume that the system is maximally irreducible; in general, however, the first four postulates of IIT should be applied to all subsets of the substrate and across grains to identify the system that maximizes φ_s.

The Φ_d values of each distinction Φ-folds are plotted in Fig. 5.3B, showing how the structure integrated information Φ is distributed across the system's irreducible mechanisms. The top inset shows the Φ-fold stemming from mechanism i; the bottom inset shows that of mechanism bcd (uppercase and lowercase unit labels denote 'ON' and 'OFF', respectively). The lattice units in level 1 are designed in such a way that when unfolded, the distinctions and relations they specify form an 'extension.' As as described in (Haun & Tononi, 2019), this is a Φ-fold that captures, in physical terms, the phenomenal structure of experienced space, which is the same as its intrinsic meaning. As argued there, phenomenal space is composed of distinctions ('spots') that are related to themselves (reflexivity) and overlap partially with other distinctions such that their intersection and union are themselves spots (connection and fusion). Similarly, the configuration detectors and segment detector together specify distinctions and relations that capture the notion of an object, binding together an invariant concept and a specific configuration of features (to be discussed in (Grasso, in preparation)).

Perception

When the system is connected to the environment ('awake' condition), the stimulus percolates through the system and triggers a response. Accordingly, the subsets of the system have positive connectedness values and triggering coefficients, shown in Figure 5.4.[10] Because the stimulus $x = 01110011$ contains the pattern 01110, it causes the configuration detector I and the segment unit M to activate (Figure 5.4A).

[10] Because we have the full TPM $\mathcal{T}_U$ of the substrate in our simulation, the connectedness values do not need to be estimated from repeated trials. Instead, the relevant probabilities can be calculated directly by matrix operations on $\mathcal{T}_U$. First, we compute $\mathcal{T}_S(\partial S_{t-\tau}) = \Pr(S_t \mid \partial S_{t-\tau})$ as follows. For each stimulus x, $\mathcal{T}_U$ is conditioned on $\partial S = x$ and the resulting conditional TPM of S is exponentiated by τ to yield $\Pr(S_t \mid \partial S_{t-\tau} = x, S_{t-\tau})$ for each initial system state $S_{t-\tau} = y_{t-\tau}$. We then marginalize over $S_{t-\tau}$ to obtain $\Pr(S_t \mid \partial S_{t-\tau})$ as desired. Marginalizing over stimuli $\partial S_{t-\tau}$ then yields $\Pr(S_t)$. Finally, the probabilities $\Pr(M_t = m \mid \partial S_{t-\tau} = x)$ and $\Pr(M_t = m)$ for each subset $M \subseteq S$ can then be obtained by marginalizing over the appropriate columns and used to compute the connectedness value and triggering coefficient.

M has connectedness $c(x, M) = 0.708$ and triggering coefficient $t(x, M) = 0.404$, reflecting the fact that its state was caused by the stimulus (Figure 5.4B). Connectedness values and triggering coefficients for subsets of the level 2 detector units are higher for those subsets that include the activated detector I and lower for those that include only non-active detectors, reflecting the fact that the configuration detectors are highly selective for segment states (Figure 5.4C). For the level 1 subsets, two patterns can be seen. First, connectedness values of level 1 subsets are generally proportional to their size (Figure 5.4D). This is because the lattice units in level 1 function largely as feed-forward relays, so the marginal probability of a subset Z being in a particular state is on the order of $1/|\Omega_Z| = 1/2^{|Z|}$ (since we impose the uniform distribution over stimuli). Second, there is smaller-scale variability in connectedness superimposed on the large-scale pattern. This reflects the differential strength of the lateral connections due to short-term plasticity, which is maximal when both units are active. When the connectedness values are normalized to yield the triggering coefficient, the influence of subset size is discounted and the variability due to the lateral connections dominates. For example, $t(x, C)$ is higher than that of its neighboring units $t(x, B)$ and $t(x, D)$ because C forms the middle of the segment and its co-active neighbors both provide relatively strong lateral input, while B and D are at the border and their respective lateral inputs from a and e are less potentiated. Similar effects can be seen with larger subsets, *e.g.* comparing $t(x, BCD)$ to $t(x, aBC)$ and $t(x, CDe)$.

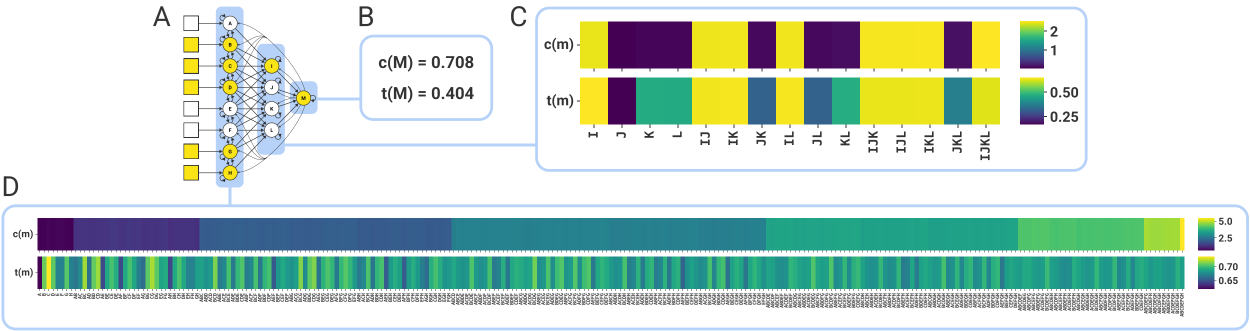


Figure 5.4. Connectedness and triggering coefficients. When the system is connected to the sensory interface, the stimulus $\partial S = x = 01110011$ percolates through the system, triggering the state $ABCDEFGH = 01110011$, $IJKL = 1000$, $M = 1$ in this example (panel **A**). Values of $c(x, m)$ and $t(x, m)$ for this stimulus-response pair are shown for the subsets of levels 3, 2, and 1 in panels **B**, **C**, and **D**, respectively. Due to space constraints, we do not show the values for the full power set, omitting the 7921 subsets that span levels.

As in the dreaming condition, the system specifies a Φ-structure intrinsically (Figure 5.5A).[11] The

[11] By definition, the sensory interface constitutes the background conditions in IIT's analysis. Therefore when unfolding the Φ-structure we condition the system's TPM on the state of the sensory interface.

leftmost plot in Figure 5.5B shows how the structure integrated information Φ is partitioned into the Φ_d values of the distinction Φ-folds specified by the system's irreducible mechanisms.

Note that Φ tends to be concentrated in the distinction Φ-folds of mechanisms that include the lattice units BCD, the corresponding detector unit I, or the segment unit M. For example, the Φ-fold of the 1st-order mechanism I contributes a relatively large amount to the total Φ. This is partly due to the convergence of many irreducible effects of other mechanisms onto I (note the density of I labels in the effect cluster on the top right of the Φ-structure plot). Because I specifies its own state as its maximally-irreducible effect, this convergence in turn leads to a high density of irreducible relations among the mechanisms whose effect purviews include $\{I\}$ (these relations contribute $\sum \varphi_r = 1.92$ to the $\Phi_d(I) = 5.25$, whereas for the distinction Φ-fold stemming from J, for example, the corresponding relations contribute $\sum \varphi_r = 0.37$).

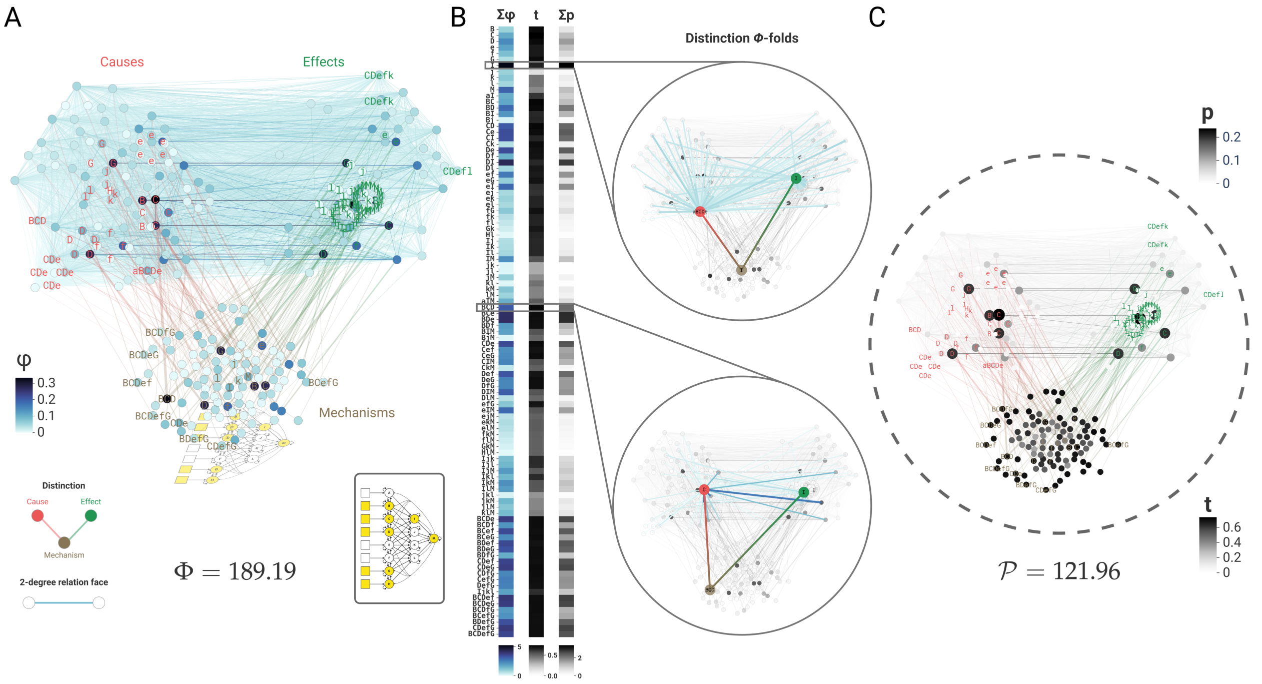


Figure 5.5. Perception. To the extent that intrinsic meanings are triggered by extrinsic stimuli, they can be considered percepts. **(A)** $\varPhi$-structure visualized as in Fig 5.3A, with the system now connected to the environment via the sensory interface ('awake' condition). *Inset*: Stimulus and response as in Figure 5.4; level 1 units relay the stimulus to the levels above, which detect the presence of a segment ($aBCDe = 01110$). Note that in this state, the system specifies a $\varPhi$-structure with higher structure integrated information $\varPhi$ than in the all 'OFF' state shown in Figure 5.3. **(B)** Left heatmap and insets as in Fig 5.3B. The middle and right heatmaps respectively show the triggering coefficient and perception values associated with each distinction $\varPhi$-fold. **(C)** The perceptual structure triggered by the stimulus. Points corresponding to irreducible mechanisms m within the system (bottom cluster) are colored according to their triggering coefficient $t(x, m)$. The causes (top left cluster), effects (top right cluster), and relations among them (lines within and between clusters) specified by the irreducible mechanisms are colored according to their perception value $p(x, d(m)) = t(x, d(m))\ \varphi_d(m)$. Grayscale is used to emphasize that the perceptual structure does not exist in its own right as a $\varPhi$-structure, but rather is the fraction of the $\varPhi$-structure that is triggered by a stimulus.

The triggering coefficients $t(x, m)$ and perceptual richness $\mathcal{P}_d(x, m) = t(x, m)\ \varPhi_d(C(d(m))$ of the distinction $\varPhi$-folds are shown in the middle and rightmost plot in Figure 5.5B, respectively. As seen in Figure 5.4, the triggering coefficients are relatively large for mechanisms that include the active units involved in detecting the segment: B, C, D, I, and M. The distribution of $\mathcal{P}_d(m)$ values indicates that the intrinsic meanings defined by certain $\varPhi$-folds, such as those of I, DI, DIM, BCe, BDe, and of higher-order mechanisms including combinations of B, C, and D, can be largely attributed to the stimulus's triggering action and are thus highly perceived. Figure 5.5C illustrates the resulting perceptual structure. This depiction is necessarily partial—higher-degree relations are omitted, as they are difficult to visualize due to their density, even though, as mentioned earlier, they contribute significantly to the $\mathcal{P}_d$. Nonetheless, one can see for example that mechanisms B, C, D and BCD specify distinctions with large values of p (color of circles indicating purviews), and each distinction's cause and effect are bound by a relation with a perception value as well (color of lines connecting purviews).

Note that in the model all triggering coefficients are positive, so the perceptual structure includes all components of the Φ-structure. In general, however, a component of the Φ-structure may have a triggering coefficient of zero—*i.e.*, it was not caused by the stimulus and would not appear in the perceptual structure at all. Conversely, if a subset does not specify an irreducible cause and effect within the system ($\varphi = 0$) or does so by specifying a state that is incongruent with that specified by complex as a whole, it would not appear in the perceptual structure despite having a positive triggering coefficient. This implies that even units within the neural substrate of consciousness may be activated by a stimulus and yet not contribute to experience and perception, as may be the case during bistable perception or during certain stages of sleep.

Stimulus-specific matching

For a single stimulus x, *stimulus-specific matching* is computed as the expected difference between the perceptual richness elicited by that stimulus and that elicited by a stimulus drawn uniformly at random from all possible states of the sensory interface (see §5.2). Figure 5.6 illustrates this calculation for a single sample from N (note that properly estimating stimulus-specific matching requires an appropriate sample size for the noise stimuli). The stimulus-specific matching value is estimated to be $\hat{m}(x, y) = 75.74$.

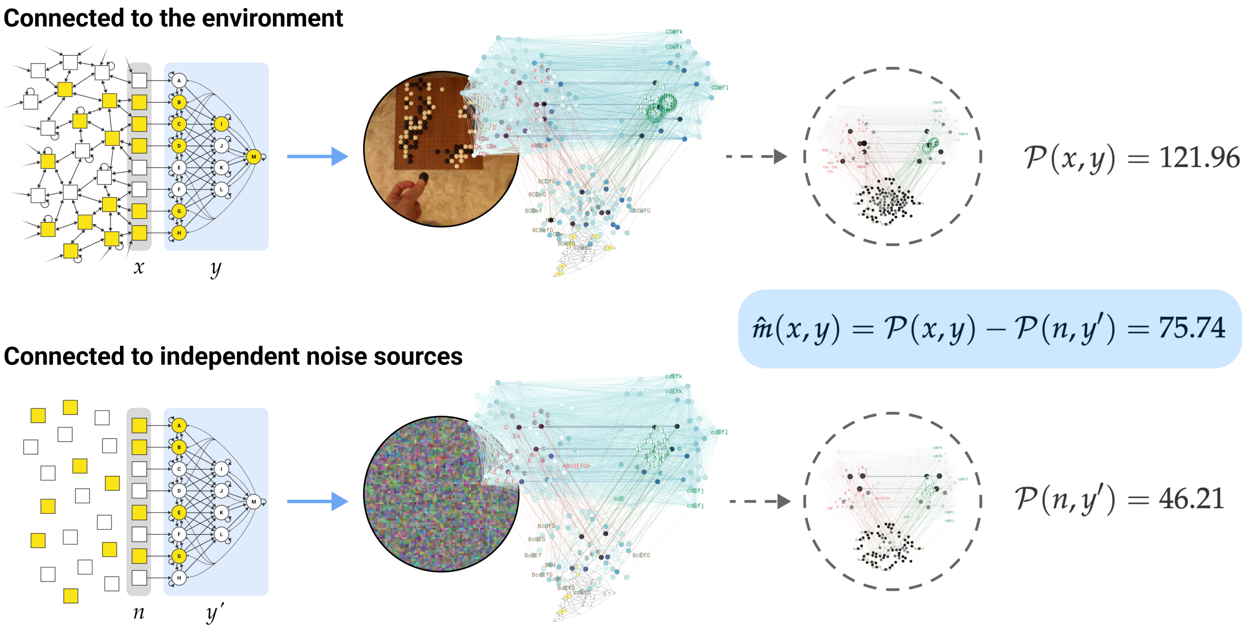

Connected to the environment
x
y
$\mathcal{P}(x,y) = 121.96$
$\hat{m}(x,y) = \mathcal{P}(x,y) - \mathcal{P}(n,y') = 75.74$
Connected to independent noise sources
n
y'
$\mathcal{P}(n,y') = 46.21$

Figure 5.6. Stimulus-specific matching. *Top*: A stimulus sampled from the environment E (the segment stimulus analyzed in Figure 5.5). *Bottom*: A stimulus sampled from independent noise sources N, depicted here as disconnected units. The stimulus from E causes a system state that specifies a perceptual structure with higher perceptual richness ($\mathcal{P}(x,y) = 121.966$) than that triggered by the stimulus sampled from N ($\mathcal{P}(N = n, Y' = y') = 42.214$), so for this stimulus-response pair the estimate of stimulus-specific matching is $\hat{m}(x,y) = 75.742$.

In general, if a system has internalized regularities in E, stimuli typical of E will tend to trigger Φ-folds of high Φ_d and/or percolate deeply into the system, resulting in large triggering coefficients for many mechanisms. This will bring about internal states that specify rich intrinsic meaning (Φ-structures with large Φ values), hence the perceptual richness associated with the environmental stimuli will be large. On the other hand, this will not be the case for stimuli sampled from N, yielding a positive value of stimulus-specific matching.

Differentiation (stimulus sequence matching)

stimulus sequence matching is calculated as the expected difference between the perceptual differentiation of a sequence sampled from E and that of a random sequence of the same length sampled from N. This is illustrated in Figure 5.7: the system is connected to an environment characterized by streams of segments moving smoothly across the sensory interface. This deviation from randomness means that sequences of stimuli sampled from the environment (top) are more likely to contain different 'segment' stimuli compared to sequences sampled from independent noise sources (bottom). As we saw in Figure 5.5, this results in greater perceptual richness on average. Furthermore, when the union is taken over the perceptual structures (right), we see that because the different segments activate different detectors (and consequently specify different distinctions and relations with high perception values), the perceptual richness triggered by each stimulus is largely additive. By contrast, sequences sampled from N tend to trigger the same state of the detector units and segment unit (all 'OFF'), so that the higher-level units contribute less to the total perceptual differentiation. Thus, the estimate of stimulus sequence matching is positive, reflecting the fact that the model was designed to perceive segments. In Figure D.1, we show estimates of matching values for different sequence lengths when the system is exposed to an environment in which segments are 5 times more likely to impinge on the sensory interface than non-segment stimuli.

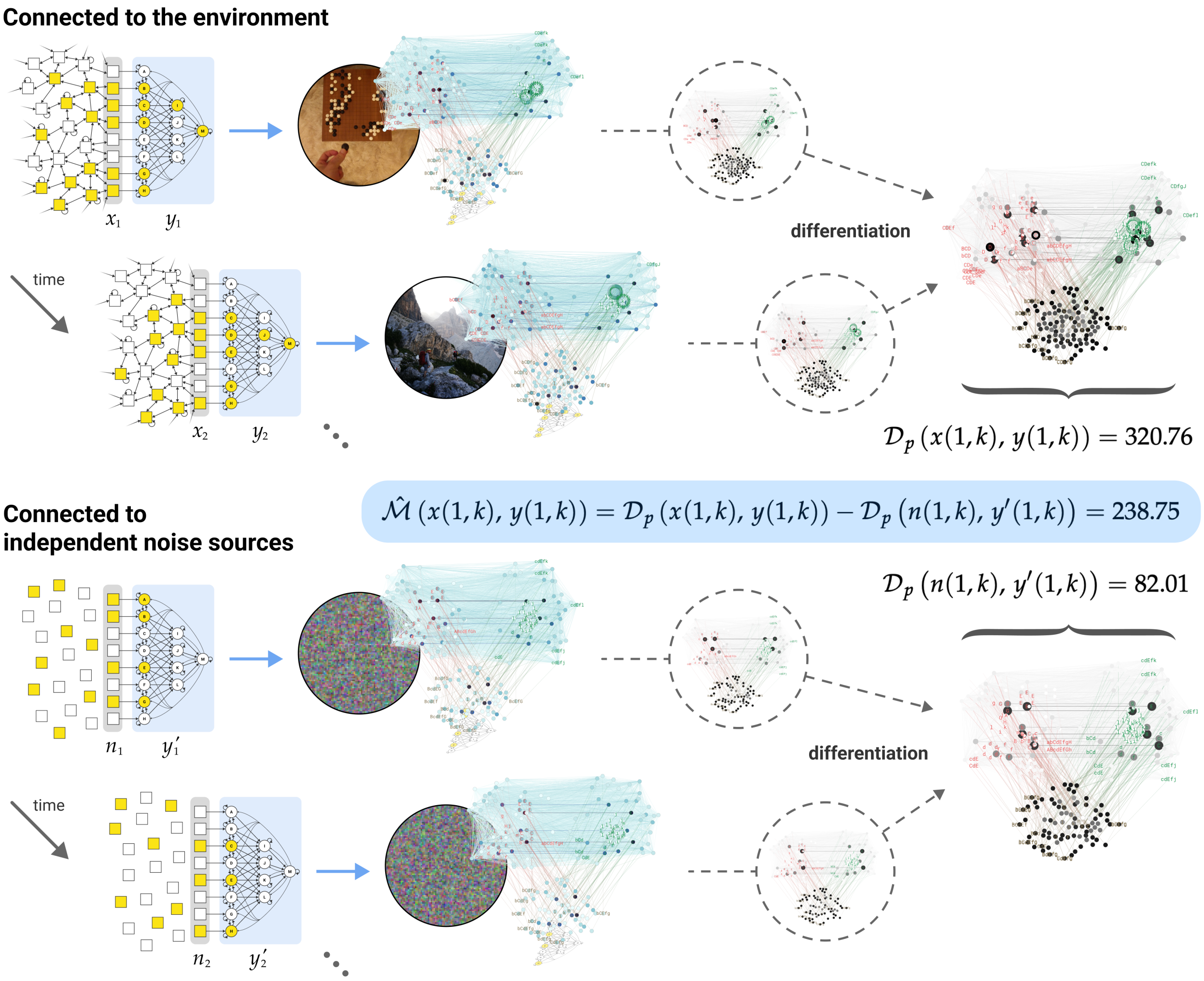

Connected to the environment
x_1
y_1
time
x_2
y_2
differentiation
$\mathcal{D}_p\left(x(1,k),\, y(1,k)\right) = 320.76$
$\hat{\mathcal{M}}\left(x(1,k),\, y(1,k)\right) = \mathcal{D}_p\left(x(1,k),\, y(1,k)\right) - \mathcal{D}_p\left(n(1,k),\, y'(1,k)\right) = 238.75$
Connected to independent noise sources
$\mathcal{D}_p\left(n(1,k),\, y'(1,k)\right) = 82.01$
n_1
y_1'
differentiation
time
n_2
y_2'

Figure 5.7. Differentiation (stimulus sequence matching). The stimulus sequence sampled from the environment E (top) is likely to contain 'segment' stimuli. Since the model system is designed to perceive segments, E is more likely to elicit perceptual structures with greater perceptual richness than those elicited by uniform noise N (bottom). Moreover, when the union is taken over the perceptual structures to yield the perceptual differentiation structure (right), the different percepts elicited by the environment all contribute to the total $\mathcal{D}_p$ value, whereas for the noise sequence there are fewer unique percepts (note that much of the difference again lies in the higher-order relations, which are not visualized here). stimulus sequence matching is estimated as the difference between the perceptual differentiation of the E sequence and that of the N sequence (highlighted in blue). The estimate is positive ($\hat{\mathcal{M}} = 238.75$), reflecting the fact that the system's intrinsic connectivity matches regularities in the environment.

Because triggered system states in general depend not only on the stimulus but also on the previous system state, this formalism indirectly captures ordering effects mediated by the system's memory. For example, when viewing a movie, the perceptual richness associated with a particular scene in a movie may be larger if the scene is the culmination of the plot than if it were presented first, because the state of the viewer's brain was "prepared" by viewing the preceding scenes.

In practice, the choice of stimulus sequence is obviously important. In order to use this measure to draw general conclusions about the matching between the system's internal causal structure and the causal processes in a given environment, care must be taken that the environmental sequences form a comprehensive sample and are representative of its regularities. Also, as with stimulus-specific matching, a sufficient sample of noise sequences must be used to obtain a good estimate.

5.4 Discussion

IIT argues that every experience is accounted for operationally by a Φ-structure specified by a complex in a state. The Φ-structure captures in full the way the experience feels, which is the same as its intrinsic meaning (Albantakis et al., 2022; Tononi, Albantakis, et al., 2022). This paper extends the IIT framework by presenting a principled approach to quantify connectedness, perception, and matching to an environment, and applies the associated measures to a minimal model of a sensory system. Below we will briefly review the framework presented in the Theory and Results section and examine some of its implications.

Intrinsic meaning, connectedness, perception, and matching

The theoretical framework presented here, and its illustration through simple computer models, starts deliberately from the intrinsic meaning of an experience, and only then proceeds to assess the effects of stimuli and their perception. Thereafter, measures of matching quantify the extent to which intrinsic meanings capture causal processes in the environment.

Intrinsic meaning. By IIT, the intrinsic meaning of an activity pattern over the complex is defined intrinsically by the distinctions and relations specified by its subsets. The quantity or richness of meaning—structure integrated information Φ—is the sum of the integrated information values φ of all the distinctions and relations that compose the Φ-structure. The intrinsic meaning of an experience is the same whether the experience occurs spontaneously, as in a dream, or it is triggered by stimuli from the environment. In fact, phenomenally it may at times be difficult to tell whether a particular scene is dreamt or perceived (Nir & Tononi, 2010; Wamsley et al., 2014).

Connectedness. As defined here, connectedness assesses the extent to which the state of subsets of units within a system is *caused* by a stimulus (the state of units of the sensory interface). The measure is based on the actual causation formalism, which employs the principles of IIT to quantify “what caused what” (Albantakis et al., 2019). Subset connectedness is high if the probability of occurrence of the current subset state is increased by the stimulus compared to any possible stimulus. The value of subset connectedness normalized by the self-information of the subset state is the *triggering coefficient*. It varies between 0, if a stimulus makes no difference to the probability of occurrence of that subset state (say, owing to a disconnection of sensory pathways), and 1, when that subset state always occurs in response to that stimulus.

Perception. Perception measures to what extent an experience is triggered by a stimulus. A *perception value* can be calculated for every subset of the complex as the product of the $\varPhi_d$ value of its distinction $\varPhi$-fold and its triggering coefficient. A subset’s distinction $\varPhi$-fold is composed of the distinction it specifies and the associated relations. Its $\varPhi_d$ value is the sum of the distinction’s φ_d value and the φ_r values of each associated relation, divided by the number of distinctions which that relation binds together. *Perceptual richness* is the sum of the perception values for all distinction $\varPhi$-folds. A *perceptual structure* represents the fraction of the $\varPhi$-structure that is triggered by a stimulus. Perceptual richness is high if the effects of stimuli percolate broadly, deeply, and strongly over the subsets of a complex (high connectedness) *and* those subsets specify distinction $\varPhi$-folds of high $\varPhi_d$. On the other hand, if the effects of stimuli do not percolate well (low connectedness) or do not trigger distinction $\varPhi$-folds with high $\varPhi_d$, then the stimuli are either not perceived or perceived suboptimally. This could happen, for example, if a stimulus affects units that do not specify distinctions contributing to the main complex, or if it affects the complex at the wrong grain (finer or coarser; Hoel et al., 2016; Marshall et al., 2018).

Regardless of how well a stimulus percolates, perceptual richness is zero if units are unable to support a $\varPhi$-structure. In other words, without consciousness there is no perception. This is clearly the case in deep slow wave sleep, when we lose consciousness due to neuronal bistability (Tononi & Massimini, 2008). Even though responses to sensory stimuli can be detected in many brain areas (Yang & Wu, 2007), we perceive nothing because we experience (next to) nothing. On the other hand, dreaming sleep exemplifies how we can be conscious without perceiving stimuli, due to mechanisms that enforce sensory disconnection (Aru et al., 2020). But even if perceptual richness is maximal, in the sense that an experience is reliably triggered by an external stimulus, the meaning of the activity pattern triggered by the stimulus is specified intrinsically, by the distinctions and relations composing the Φ-structure. In this sense, a perception is always an interpretation.

Matching (stimulus-specific). Matching is assessed by taking the expected difference between the perceptual richness triggered by a stimulus sampled from an environment (E) and that triggered by stimuli sampled from independent noise sources (N). In general, causal processes within E ensure that some stimuli sampled from E (or sequences of stimuli) are more likely than others. Independent noise sources exclude any causal process, such that every stimulus and sequence of stimuli is equally likely. For *stimulus-specific matching* to be high, many distinctions and relations triggered by the E stimulus must have higher φ value than those triggered by typical N stimuli. This can happen if the complex's intrinsic connectivity leads to more selective responses to E stimuli whose probability differs from chance. In the model, for example, the level 2 detector units and the level 3 'segment' unit were wired to selectively activate for 'segment' stimuli which, due to causal processes in E, occur more frequently than chance. Crucially, when these units are active, they specify more selective cause-effects, yielding higher φ values for the distinctions and relations to which they contribute. Furthermore, the dynamic increase in connection strength (short-term plasticity), which is maximal if both connected units are active, yields even higher φ values. 'Lattice' units at level 1 are also more likely to contribute higher φ values when a stimulus is sampled from E. Owing to the 'spatial' structure of E (resulting from causal processes within it), stimuli sampled from E cause nearby lattice units to be in the same state more frequently than chance. The dynamic increase in connection strength is higher for units in the same state, yielding φ values that are higher, on average, for E stimuli than for N stimuli. Altogether, then, high stimulus-specific matching indicates that a stimulus sampled from E triggers greater intrinsic meaning than typical ones from N. Moreover, this increase in meaning can be said to *refer* to some regularity in E. Said otherwise, the intrinsic connectivity of the complex must have adjusted to preferentially "recognize," by triggering distinctions and relations of higher φ, stimuli whose frequency is higher than chance due to causal processes in E. Thus, matching provides a link between intrinsic meaning and extrinsic reference to environmental regularities.

Phenomenally, we can easily recognize this situation by considering how we experience a frame of a movie versus typical "TV noise" frames. The former, which percolates deeply within the brain (Boly et al., 2015; Mayner et al., 2022; Mensen et al., 2017, 2018), triggers an experience rich in structure, with many meaningful, high-level interpretations bound to low-level contours and features. The latter, which percolate much less, affecting mostly primary sensory areas, will trigger an experience whose structure will be almost exclusively spatial, with hardly any high-level meaning, except perhaps for the concept of "noise." (Consistent with IIT, we cannot help interpreting perceived sensory input as spatially organized even if its source is independent noise; Haun and Tononi, 2019).

Differentiation (stimulus sequence matching). Finally, *stimulus sequence matching* is assessed by taking the expected difference between the *perceptual differentiation* of a sequence of stimuli from E to that of an equally long sequence of stimuli sampled from N. Perceptual differentiation is high if the environment triggers a large set of unique distinctions and relations. This is the case if stimuli not only trigger high perceptual richness, but different stimuli trigger different distinctions and relations. In the model, this occurs because different segment stimuli activate different detector units at level 2 (with the associated distinctions and relations), whereas random stimuli typically do not. A model with additional high-level units for invariants such as dots and lines, not to mention a large system such as the brain, would have many high-level, specialized units activated by stimuli reflecting different regularities in E, further increasing differentiation. By contrast, stimuli sampled from N would hardly activate any high-level units, with the possible exception of a few "noise" units that would be activated almost every time. Thus, stimulus sequence matching reflects the preferential triggering of different intrinsic meanings by different regularities in the environment, to which they refer.

Again, we can easily recognize this situation phenomenally by considering a movie and a sequence of TV noise frames. Watching a movie triggers all kinds of experiential contents—different people and animals and plants, many different objects, countless distinct events—yielding high perceptual differentiation. Watching a TV out of tune does not: different noise frames will not trigger different high-level contents, and will only differ in the details of local spatial features, yielding low perceptual differentiation.

In essence, stimulus sequence matching reflects how well the distinctions and relations available to the main complex, across any of its states, capture different regularities in the structure of the environment. Accordingly, stimulus sequence matching is high if an organism has internalized many regularities about different aspects of its environment, and low otherwise. Conversely, matching is zero if a system lacks a main complex (it has no cause-effect structure) or if the environment lacks causal processes (the stimulus sequence will be indistinguishable from N). Matching is also zero if a complex is disconnected, or if its intrinsic connectivity has not internalized any regularity from E.

Some implications and outlook

IIT's intrinsic perspective on meaning, perception, and matching has several implications that will be addressed in forthcoming publications (Comolatti & Grasso, in preparation; Grasso, in preparation) and can only be touched upon briefly here.

Extrinsic triggering of activity patterns and their endorsement by the intrinsic connectivity. We chose a minimal model to illustrate our formalism in order to permit the exact calculation of the relevant quantities. Even so, the model incorporates some features that may be neurobiologically relevant. It is well established that, at last in sensory areas of posterior cortex, the connectivity intrinsic to each area far outweighs the extrinsic connectivity originating from subcortical inputs or other cortical areas, except for directly adjacent areas (Binzegger et al., 2009; Binzegger et al., 2004; Vezoli et al., 2021). For example, 85–90% of connections to primary visual cortex originate within it, but less than 1% originate in the lateral geniculate nucleus of the thalamus, its primary sensory afferent (Binzegger et al., 2004; Markov et al., 2011). The intrinsic connectivity is extraordinarily dense locally (within 1–2 millimeters) and especially so within supragranular layers (Binzegger et al., 2009; Binzegger et al., 2004). The latter have greater neuronal density than infragranular layers in much of posterior cortex, while the opposite is true for much of prefrontal cortex (Castrillon et al., 2023). Also, in much of posterior cortex, this dense supragranular mesh of intrinsic connections is organized as a lattice of specialized units, which fits the requirement of maximal integration for the substrate of consciousness (Albantakis et al., 2022; Tononi et al., 2016), while prefrontal regions tend to be organized in a modular manner (Watakabe et al., 2023). However, the dominance of this intrinsic lattice-like connectivity poses the problem of how sensory inputs (and possibly long-range inputs from distant brain areas) can be effective in driving cortical activity patterns and ensuring connectedness.

A common idea is that the intrinsic connectivity "amplifies" and sustains the initial effects of sensory inputs through various specializations (Chance et al., 1999; Cottaris & De Valois, 1998; Douglas et al., 1995; Lamme & Roelfsema, 2000; L.-y. Li et al., 2013; Y.-t. Li et al., 2013; Lien & Scanziani, 2013; Peron et al., 2020; Ringach et al., 2003; Zerlaut et al., 2019). In the model, we systematically employ short-term synaptic plasticity (Citri & Malenka, 2008; Kandel & Tauc, 1965) to dynamically adjust the strength of connections at the time scale of perception so as to endorse, rather than override, the activity pattern triggered by a stimulus. Mechanistically, the stimulus conveyed by bottom-up connections rapidly activates, bottom-up, an appropriate set of units within the complex. Short-term synaptic plasticity then kicks in, increasing the efficacy of intrinsic connections for tens to hundreds of milliseconds. This "*intrinsic connectivity endorsement*" (*ICE*) of the activity pattern triggered by the stimulus prevents the occurrence of "hallucinatory" patterns and associated experiences when the complex is connected to the environment (minor refinements of noisy activity patterns remain compatible with this scenario). The increased efficacy of intrinsic connections also ensures that, at the time scale of experience, the system's cause-effect power over itself remains maximal. Furthermore, short-term synaptic plasticity affects connections differentially, yielding maximal efficacy when both pre- and post-synaptic units are active. This mimics the action of NMDA receptors, especially concentrated in supragranular layers (Zilles & Palomero-Gallagher, 2017), which multiply the strength of activated synapses (Larkum et al., 1999; Larkum et al., 2004). In this way, active units contribute more than inactive ones to the information content of an experience (by specifying distinction Φ-folds of higher Φ_d). This fits with the notion that in the brain, due to energy constraints, strong activation should be reserved for the selective signaling of relatively infrequent but highly meaningful stimuli (a face, an object, and so on), that need to percolate deeply within the system (Balduzzi & Tononi, 2013).

Assessing differentiation and matching with neurophysiological data. Unfolding Φ-structures systematically to evaluate matching is only possible for extremely simple substrates such as the model employed here. However, it is possible to obtain practical approximations using *neurophysiological differentiation* as a proxy (Chapter 4; Boly et al., 2015; Mensen et al., 2017, 2018). The reason is that activity patterns over a complex fully determine the associated Φ-structures, as long as we can reasonably make assumptions about the border and grain of the substrate of consciousness and its relative stability. Thus, even a coarse estimate of neurophysiological differentiation may be adequate to rank relative levels of perceptual differentiation for brain regions likely to support consciousness. For example, we previously showed using fMRI and estimates of Lempel-Ziv complexity that neurophysiological differentiation within cortex (especially its posterior regions) was higher for a movie than for an equivalent sequence of TV noise frames, and that it was higher if the movie was in the proper sequence rather than scrambled (Boly et al., 2015). Furthermore, we showed using high-density EEG that an estimate of neurophysiological differentiation was higher for sets of stimuli or movie clips that the subjects found more meaningful (Mensen et al., 2018). Similar results were obtained at cellular resolution with calcium imaging (Chapter 4; Mayner et al., 2022) and Neuropixels recordings (Gandhi et al., 2023) in mice. On the other hand, measuring neurophysiological differentiation over the retina (or the cerebellum) would say nothing about perceptual differentiation, because these systems cannot support consciousness due to their internal organization (Albantakis et al., 2022). Similarly, measuring differentiation at a spatial or temporal grain that is either too fine or too coarse could produce misleading results (Hoel et al., 2016; Marshall et al., 2018).

In future work, measures of neurophysiological differentiation could be employed to monitor brain development and the effectiveness of learning in neurotypical individuals. They could also help in assessing which stimulus sequences are most meaningful for subjects who are not neurotypical, such as individuals with autism or schizophrenia. In all cases, measures of differentiation should be obtained with careful consideration of the likely location of the substrate of consciousness, while ensuring that subjects remain behaviorally engaged and therefore connected. Stimulus sequences should be chosen such that they are as comprehensive and representative as possible, and the control sequences should be adequately scrambled (by scrambling the original sequence in space and time or by generating random stimuli with the same first-order statistics; § 4.2, Phase scrambling; Boly et al., 2015). Stimuli should also be highly relevant for the subjects of study. In principle, estimates of neurophysiological differentiation could also be employed to evaluate the relative meaningfulness of different stimulus sets for species whose ecological habitat and brain organization is little known. In such cases, a search strategy aimed at maximizing differentiation might be used to infer what might be meaningful to that species. This point is important conceptually, because it illustrates how optimizing matching can establish what aspects of an environment are meaningful to an organism without presupposing or predefining what those aspects might be.

Maximizing neurophysiological differentiation using a large set of diverse stimuli likely to be meaningful to a subject would also help in estimating a complex's *intrinsic differentiation capacity*. Theoretically, this corresponds to the sum of the φ values of all the unique distinctions and relations specified by the units of a complex when initialized in all possible states (at the optimal grain). Under certain assumptions about the border and grain of a complex, its intrinsic differentiation capacity should be proportional to the Φ value of the Φ-structure it specifies when in a typical state (see Marshall et al., 2016). In other words, maximal neurophysiological differentiation could also serve as an estimate for the quantity of consciousness Φ.

Intrinsic and extrinsic approaches to meaning and information. For IIT, once an activity pattern is established over the main complex (at the right grain), whether it was triggered primarily by exogenous signals, endogenous ones, or their interplay, its meaning is defined intrinsically by the distinctions and relations that compose the associated Φ-structure. As a Φ-structure, meaning is defined intrinsically, "here and now," rather than by reference to extrinsic factors (Zaeemzadeh, in preparation). This contrasts with common approaches that take an extrinsic perspective on information and meaning. For example, the brain is often portrayed as an "information processing" device, where information is conceptualized along the lines of Shannon's information theory (Zaeemzadeh, in preparation). Accordingly, the brain is supposed to "decode" information provided by a stimulus, understood as a message or symbol, and "compute" appropriate outputs based on that information, memories, and goals. Top-down "priors" are thought to act as error correction codes, carrying "contextual" information to properly decode of fill in noisy or incomplete stimuli. It is also suggested that messages that are broadcast globally through long-range connections correspond to "conscious processing" (Dehaene et al., 2011). Alternatively, the brain has been portrayed as a "predictive processing" device that implements a reverse message-passing scheme, where top-down inferences are updated based on error signals provided by stimulus information (Cao, 2020).

In both cases, however, it is not clear what would determine the meaning of the activity pattern resulting from such processing. Extrinsic approaches to meaning often point, on the input side, to sources in the environment from which meaning is somehow inherited. On the output side, they point to subsequent actions or intermediary computations within the brain. More generally, functionalist approaches equate meaning to a function being performed with respect to inputs, outputs, and internal states, such as memories and goals (Putnam, 1973). Meaning would then be revealed by what happens next (a process), rather than being identical to what exists here and now (a structure). It would be about doing rather than being (Albantakis & Tononi, 2021). IIT's intrinsic view of the meaning of an experience as the cause-effect structure specified, at the time scale of perception, by a complex's activity pattern through its intrinsic connectivity, also differs from extrinsic views that assign meaning based on the similarities among activity patterns (Kriegeskorte et al., 2008) or their location within "conceptual spaces" (Gärdenfors, 2000).

Of course, the composition of a Φ-structure for a specific activity pattern at the time scale of perception is determined by the intrinsic connectivity of the main complex. To a large extent, the latter is organized the way it is because it has internalized regularities sampled from the environment over the course of evolution, development, and learning. In this sense, the environment is partially responsible for the meaning of an experience, but only indirectly so. When the experience occurs, its meaning is defined intrinsically, here (over the main complex) and now (at this particular moment), and it goes much "beyond the information given" (Bruner, 1973).

Explanatory and generative power. As we have seen, for matching to be high, different aspects of causal processes within the environment must have been internalized by different portions of the connectivity of the complex. When an organism samples different stimuli, each will trigger a different activity pattern and an associated Φ-structure. This means that the complex can provide a specific interpretation or "explanation" for many different aspects of the environment. High matching then implies high *explanatory power*.

A complex with high matching to a given environment is also an excellent substrate for high *generative power*. This can be understood as the ability of a complex to produce, through its intrinsic connectivity, sequences of intrinsic states that are similar to those observed when it is connected to the environment's causal processes. For example, consider the organization of posterior cortex as a hierarchical lattice poised to be triggered by stimuli via bottom-up input. Being maximally integrated, the same lattice can also enable, under the right neuromodulatory conditions, the propagation of top-down activity. This is what seems to be happening during dreaming, imagination, or mind-wandering, when the corticothalamic system generates, usually in a top-down manner, activity patterns that may be highly similar to those triggered by environmental stimuli. Regardless of whether these activity patterns are triggered exogenously and bottom-up (from the particular to the general), or endogenously and top-down (from the general to the particular), the underlying intrinsic connectivity is the same. Therefore the associated and corresponding feeling—which is to say, the corresponding intrinsic meaning—will be just as similar.

Dreams provide the most obvious demonstration that the substrate of consciousness has both explanatory and generative power. They reflect an immense amount of internalized information about causal processes in our environment: we dream of humans, animals, plants, and objects, faces and places, colors and sounds, not of concepts or entities that we are unequipped to experience, in principle, during wakefulness. In short, dreams reflect the explanatory power of the substrate of consciousness. But dreams also demonstrate the substrate's generative side, because we can evidently bring into existence world-like experiences endogenously, without the need to trigger them exogenously (Nir & Tononi, 2010; Pearson, 2019). Imagination during wakefulness also testifies to our ability to trigger different experiences endogenously, although the vividness and detail of imagination vary greatly among individuals. This generative power thus frees the organism from the "tyranny of the here and now" and allows it to plan ahead and try out imaginary scenarios.[12]

Intrinsic and extrinsic matching. Finally, a key implication of IIT concerns computers implementing artificial intelligence (as well as computers potentially simulating our brain). From a human vantage point, such computers might behave *as if* they were perceiving and understanding the world just as we do. In fact, extrinsic measures of differentiation and matching could also be defined by analogy with intrinsic measures. For example, high values of mutual information between sensory interfaces and different subsets of specialized "hidden" units, much higher than for noise stimuli, would be consistent with high extrinsic matching (see also Tononi et al., 1996). However, because their internal organization is incompatible with the existence of a large complex, computers cannot support Φ-structures of high Φ (Findlay et al., 2019). Thus, even though they may become functionally equivalent to us—navigating the environment, answering questions, and pursuing seemingly goal-directed behaviors—they would experience nothing, perceive nothing, and have neither intrinsic goals nor free will (Tononi, Albantakis, et al., 2022).

Future extensions. In future work, IIT's approach to meaning will be extended in several directions, some of which we will mention briefly. A natural extension is to evaluate connectedness and triggering coefficients on the output side. Just as a stimulus over a sensory interface is connected to the main complex if it is the actual cause of the state of the complex's subsets, an action is connected to the

[12] More generally, while off-line during sleep, the brain endogenously triggers many different activity patterns, and down-selects connections that underlie less congruent patterns (Tononi & Cirelli, 2014). The repeated cycling between comprehensive synaptic down-selection in sleep and selective synaptic potentiation, guided by environmental regularities, during wakefulness, can promote matching while maintaining flexibility in the face of environmental changes (Hashmi et al., 2013; Nere et al., 2013).

main complex if the latter has actual effects on a motor interface. Accordingly, just as a stimulus is perceived and interpreted to the extent that it triggers a distinction Φ-fold within the Φ-structure of the main complex, an action is intended to the extent that a Φ-fold specifies effects that trigger that output. And just as a perceived Φ-fold can include high-level concepts that capture objects and events, an intention Φ-fold may include high-level concepts that capture goals and action plans. Of course, because of its ability to control its outputs (say, eye or hand movements), a complex can change the way its samples inputs, and thus influence its own matching. Active exploration of the environment is important not only for perception here and now, but during learning, when it can greatly enhance our ability to sample suspicious coincidences and sequences and thereby internalize regularities in the environment. Furthermore, actions can modify causal processes within the environment, and even create new things or processes, to match novel concepts and meanings that may have been "invented" intrinsically.

Another extension will be to characterize causal processes dynamically evolving in the environment—their border, beginning, end, and grain—based on the actual causation formalism (Albantakis et al., 2019) and to distinguish them from stable causal entities (Tononi, Albantakis, et al., 2022). This will help to ground the notion of matching as subjective reference to regularities in the environment and to distinguish it from objective, intersubjective knowledge (Tononi, forthcoming). Finally, another extension of the present framework will be the exploration of the bounds of intrinsic meaning. In practice, complexes with an intrinsic connectivity capable of specifying large numbers of unique distinctions and relations come about because of adaptive interactions with a rich environment, as suggested by computer simulations using genetic algorithms (Albantakis et al., 2014). In principle, one could imagine constructing complexes capable of achieving maximal differentiation capacity in a way that is completely disconnected from matching and reference and only limited by bounds on integrated information (Zaeemzadeh & Tononi, 2023). But even a system optimized in this way would not remotely be able to specify intrinsically all the distinctions (or "concepts") that could be specified about its current state (see Feldman, 2003), together with the associated relations. Therefore, in one relevant sense, it would not be able to fully interpret itself.

CHAPTER 6

Discussion

The work presented in this dissertation is part of a coherent effort to develop a principled, parsimonious, and comprehensive theory of consciousness. Integrated information theory offers a novel approach to bridging the explanatory gap between consciousness and its physical substrate by starting from consciousness itself. It proposes a fundamental explanatory identity between an experience and a Φ-structure specified by a maximally-irreducible substrate, and thereby offers an account of the presence and quality of consciousness.

In Chapter 2, I described my software package PyPhi, which provides a reference implementation of IIT's mathematical formalism. Chapter 3 presented the latest version of the theory, IIT 4.0, which incorporates several developments of the past ten years. Chapter 4 reported the results of an empirical application of IIT, in which we measured stimulus-evoked neurophysiological differentiation elicited by unscrambled compared to scrambled stimuli. In Chapter 5, I introduced an extension of IIT's formalism that characterizes the relationship between intrinsic meaning and environmental stimuli, providing a theoretical foundation for the experiments of Chapter 4. Viewed in light of this theoretical account of perception and matching, those experiments, as well as previous work in our laboratory (Boly et al., 2015; Mensen et al., 2017, 2018), can be recognized as attempts to estimate matching empirically. In this chapter, I discuss potential directions for future research, and close with some remarks on the intrinsic perspective taken by IIT.

PyPhi

PyPhi was initially conceived merely as the backend of an interactive online tool for the computation and visualization of Φ-structures. In the course of building that tool, I realized that a full-fledged software package that implemented IIT's formalism would be valuable in its own right, and developing PyPhi became the primary aim of the project. Since its initial release in 2014, PyPhi has become the main tool for researchers in our group and for the IIT research community at large, and has provided a crucial means of implementing and testing ideas during the development of IIT 4.0 over the last several years. It will be important to continue its development and maintenance so that it keeps pace with theoretical advances and continues to serve the growing IIT research community.

There are several natural directions for future development of the software. First, the code that implements the formalism for perception and matching introduced in Chapter 5 should be integrated into PyPhi. Second, a module implementing existing IIT-inspired measures that can be used to analyze empirical data would greatly increase PyPhi's scope and utility as a centralized tool for researchers interested in applying IIT's ideas in their work. (Several heuristic approximations are currently available in PyPhi, but are not applicable to empirical data.) Third, prior work extending PyPhi's analysis to systems of multi-valued units (Gomez et al., 2020) should be properly incorporated into the main branch of the software. Fourth, the automated test suite and comprehensive documentation that accompanied PyPhi's implementation of IIT 3.0 should be updated to reflect its implementation of IIT 4.0. Finally, a project is currently underway to redesign the software's internal representation of the dynamical system under analysis so that only the local TPM of each system unit (Figure 2.2) is instantiated in memory, rather than the full joint distribution of system states. This representation is significantly more efficient in sparse networks, and would enable the manipulation of larger systems of units. This will be critical, for example, in ongoing work applying IIT to account of the experience of objects (Grasso, in preparation) and the flow of time (Comolatti & Grasso, in preparation), in which the Φ-folds specified by local connectivity motifs within relatively large model systems are of interest.

Empirical validation of IIT

As IIT has gained prominence as a potential explanation for consciousness, it has naturally attracted criticism. Targets include IIT's formulation in terms of axioms and their translation into postulates (Bayne, 2018; McQueen, 2019; Mediano, Rosas, Bor, et al., 2022; Morch, 2019; but see Leung and Tsuchiya, 2023; Negro, 2022b); its counterintuitive predictions about certain hypothetical substrates and some panpsychist implications (Aaronson, 2014; Merker et al., 2022; but see Negro, 2022a; Tononi, 2014; Tononi, Boly, et al., 2022; Tononi and Koch, 2015); the applicability of its mathematical formalism to actual physical systems (A. Barrett, 2014; A. B. Barrett & Mediano, 2019); causal structure theories in general (Doerig et al., 2019, 2021; Hanson and Walker, 2021; but see Albantakis, 2020b; Fahrenfort and van Gaal, 2021; Kleiner, 2020; Kleiner and Hoel, 2021; Negro, 2020; Tsuchiya et al., 2020); and, perhaps most importantly, the empirical evidence marshalled in its support (Merker et al., 2022; Michel & Lau, 2020; Seth & Bayne, 2022).

Empirical tests of IIT have so far been constrained by the exponential time complexity of its algorithm stemming from combinatorial explosion (§ 2.3, Limitations), which precludes exact calculation of the Φ-structure. Evidence for IIT thus necessarily rests on the plausibility of the correlation between an experimentally measurable quantity and the actual integrated information of the measured system, or on the plausibility of inferences about integrated-information-theoretic properties of various brain regions, which then enable experimental predictions.

For example, PCI, which was designed based on the principles of IIT (Casali et al., 2013), measures the joint presence of integration and differentiation, and can thus be reasonably assumed to correlate with Φ. But the relationship between PCI and Φ has not been formally characterized and is likely quite complex, limiting its value as an empirical tool for validating IIT's predictions beyond a certain level of precision.

Likewise, IIT's predictions that posterior cortical areas are sufficient for consciousness, and that prefrontal cortical regions are not, rest on inferences about how IIT's causal analysis would characterize those regions based on anatomical and functional considerations (Boly et al., 2017). Indeed, although the first results from the ongoing adversarial collaboration between proponents of IIT and Global Neuronal Workspace Theory (GNWT) validated IIT's main predictions and disconfirmed several predictions of GNWT, Dehaene argued that because Φ was not calculated directly, the experiments did not specifically test IIT *per se* (Cogitate Consortium et al., 2023).[1]

This kind of evidence cannot simply be dismissed; on the contrary, using such inferences to make and test predictions is essential to making progress. The fact that IIT's mathematical framework enables quantitatively precise predictions, in addition to the more general predictions that were tested in the adversarial collaboration, is a strength, not a weakness, when it comes to empirical testing—especially considering that GNWT cannot make such predictions. Nonetheless, the empirical case for IIT would be greatly strengthened if the connection between experimentally measurable quantities and IIT's

[1] It must be noted, however, that the predictions of GNWT in these experiments were equally unspecific, contrary to Dehaene's claim of "considerable asymmetry" (Cogitate Consortium et al., 2023, p. 27). Moreover, in such situations disconfirmatory evidence is relatively stronger than confirmatory evidence. The null results of these large-scale experiments are severely challenging for GNWT, despite their minimization in Dehaene's discussion.

mathematical formalism were placed on firmer theoretical footing.

What is needed, then, are formal approximations of the IIT algorithm: measures that can be tractably computed on experimental data and that are provably within some known factor of the exact result. Developing such approximations is a daunting task because of the highly nonlinear and discontinuous nature of the algorithm. Nonetheless, with the latest refinement of the theory completed (Chapter 3), IIT has reached a relatively mature state, and the time is ripe to advance a research agenda focused on forging a stronger link between theory and experiment.

Despite the difficulties posed by the complexity of IIT's algorithm, there are some exciting potential avenues towards this goal. Among the various investigations into alternative information-theoretic complexity measures inspired by IIT (A. B. Barrett & Seth, 2011; Luppi et al., 2021, 2023; Mediano, Rosas, Luppi, et al., 2022; Mediano et al., 2019, 2021; Tegmark, 2015, 2016), approaches based on information geometry (Amari, 1983, 2016) seem particularly promising. Several researchers have been pursuing this line of inquiry in recent years (Amari et al., 2018; Ito et al., 2020; Kitazono et al., 2018; Oizumi, Amari, et al., 2016; Oizumi, Tsuchiya, & Amari, 2016). For example, Kitazono et al. (2018) have demonstrated some success in approximating the minimum information partition by exploiting Queyranne's submodular optimization algorithm, though it remains to be seen whether their approach can be adapted to IIT 4.0, and whether such an approximation's validity can be formally proven.

Differentiation analysis (Chapter 4) will also be an important tool in this endeavor. One of its key advantages is that measures based on this approach can be easily calculated from empirical data (see § 4.2 and § 4.2). Marshall et al. (2016) were able to prove theorems relating simple measures—the size of a system's dynamical state repertoire, ($\mathcal{D}_1$) and the cumulative variance of the system's units ($\mathcal{D}_2$)—to the system's Φ value. However, that work was also based on IIT 3.0 (in particular, the definitions of cause and effect information) and will need to be adapted to IIT 4.0. Moreover, an important limitation of those theorems is the assumption that the system is integrated. A potentially fruitful approach for future research may therefore be to combine information-geometrical methods with differentiation analysis to develop tractable approximations that simultaneously assess integration and differentiation, while formally characterizing their precise relationship to the exact quantities in IIT.

Differentiation, perception, & matching

The account of perception and matching presented in Chapter 5 provides a theoretical foundation for empirical studies of stimulus-evoked neurophysiological differentiation. We have already begun to apply this measure in experiments, both at the whole-brain in humans (Boly et al., 2015; Mensen et al., 2017, 2018) and at the level of neuronal populations with cellular resolution (Chapter 4; Gandhi et al., 2023). Here I discuss some possibilities for future work in this direction.

In several recent studies, a closed-loop experimental paradigm combining *in vivo* recordings with *in silico* modeling of neuronal responses via artificial neural networks, termed the 'inception loop,' has been used to implement a gradient ascent search strategy for stimuli that maximally drive visual neurons (Bashivan et al., 2019; Ding et al., 2023; Fu et al., 2023; Walker et al., 2019). This approach could be adapted to maximize measures of neurophysiological differentiation, rather than activation strength. A comparison against scrambled stimuli could be employed to maximize stimulus-specific matching, and the search domain could be expanded to sequences of stimuli to maximize stimulus sequence matching. Such experiments would constitute a direct application of the matching formalism to characterize which regularities in the environment the recorded brain regions have internalized in their connectivity through evolution, development, and learning.

Furthermore, maximizing neurophysiological differentiation would help in estimating the differentiation capacity of a substrate. When the integration of the substrate can be assumed, as in intact cortex, this can provide an estimate of the Φ value of the Φ-structure specified by the substrate in a typical state. According to IIT, then, an estimate of the differentiation capacity can serve as an estimate for the quantity of consciousness. This approach could be especially useful in cases where no subjective reports are available, such as in patients with disorders of consciousness, or in animals.

Goh et al. (2023) recently conducted a series of elegant experiments which demonstrated that periods of silence are perceived, not merely cognitively inferred, by employing auditory temporal illusions induced by silence. The authors noted that their results challenge traditional accounts of perception that rule it out in the absence of positive sensory content. These results can be naturally framed within our account, since one of IIT's predictions is that inactive units also contribute to experience by specifying Φ-folds within the Φ-structure, just as active units do (Tononi et al., 2016). If an inactive unit's state was triggered by the sensory interface—whether by a positive stimulus or the absence of one—the associated triggering coefficient and perception value will be high. Future experiments could use a similar design, but with the additional manipulation of selectively and reversibly inactivating early auditory areas only during the silent periods.[2] Though technically difficult, this would enable a contrast between silence-induced temporal illusions in the inactive *vs.* inactivated conditions, with the prediction that the illusion only occurs in the former.

Future theoretical work should extend our formalism of perception and matching in the sensory domain to account for intentions and actions. Connectedness and triggering coefficients could be evaluated on the output side by considering the effect of internal subsets of the system on its motor interface. In the same way that a stimulus is perceived to the extent that it triggers a distinction Φ-fold within the Φ-structure of the main complex, an action is intended to the extent that a Φ-fold specifies effects that triggered it. And just as a perceived Φ-fold can include high-level concepts corresponding to abstractions such as objects and events, an intention Φ-fold may correspond to high-level concepts that correspond to goals and action plans. Closed-loop paradigms using brain-machine interfaces (Sorrell et al., 2021), such as that employed by Clancy and Mrsic-Flogel (2020), could be adapted to explore this extension to the motor domain experimentally. This would enable the empirical study of how an organism can modify its own sampling of environmental inputs through goals, decisions, and actions, and thereby influence its own matching.

The intrinsic perspective

The work presented here can be situated within a larger re-evaluation of classical methodologies that invoke notions such as “neural representations,” “neural codes,” and “information processing” in the brain (Baker et al., 2022; Buzsáki, 2020; Ebitz & Hayden, 2021; Favela, 2021; Hayden & Niv, 2021a, 2021b). As pointed out by Brette (2019), appeals to neural coding often sidestep the critical question of who is interpreting the code. Indeed, in his seminal paper, Shannon (1948) explicitly noted the irrelevance of meaning for information theory. One consequence is that, for example, decoding stimulus information from neural activity does not imply that the activity necessarily refers to the stimulus, nor even that it is functionally relevant for the brain. In Chapter 4, such considerations motivated the contrast between

[2] This would likely require optogenetic techniques in animals; analogous silence illusions that can be read out behaviorally would first need to be demonstrated in the model organism.

differentiation analysis and stimulus decoding: we found that neurophysiological differentiation was significantly higher in response to naturalistic *vs.* phase-scrambled stimuli in specific cell populations, while decoding performance did not vary by stimulus category or identity. There is no homunculus at the end of the line—the meaning of neural activity must be specified intrinsically, by the brain and for the brain, and cannot be inherited from correlations with extrinsic experimental variables.

By taking the intrinsic perspective, IIT can provide a theoretical foundation for this emerging "inside-out" methodology (Buzsáki, 2019), and can ground notions of meaning and reference. Moreover, IIT supplies concrete empirical methods that we have applied in experiments. Ultimately, if IIT's research program succeeds, it may yield a scientific explanation for consciousness. It is my hope that the work presented here is a step towards that goal.

REFERENCES

Aaronson, S. (2014, May 21). *Why I Am Not An Integrated Information Theorist (or, The Unconscious Expander)*. Shtetl-Optimized. Retrieved August 14, 2023, from https://scottaaronson.blog/?p=1799

Albantakis, L. (2020a, July). Integrated information theory. In *Beyond neural correlates of consciousness* (pp. 87–103). https://doi.org/10.4324/9781315205267-6

Albantakis, L. (2020b, September 14). *Unfolding the Substitution Argument*. Conscious(ness) Realist. Retrieved August 14, 2023, from https://consciousnessrealist.com/unfolding-argument-commentary/

Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A. M., Marshall, W., Mayner, W. G. P., Zaeemzadeh, A., Boly, M., Juel, B. E., Sasai, S., Fujii, K., David, I., Hendren, J., Lang, J. P., & Tononi, G. (2022, December 30). *Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms*. https://doi.org/10.48550/arXiv.2212.14787

Albantakis, L., Hintze, A., Koch, C., Adami, C., & Tononi, G. (2014). Evolution of Integrated Causal Structures in Animats Exposed to Environments of Increasing Complexity. *PLoS Computational Biology*, *10*(12). https://doi.org/10.1371/journal.pcbi.1003966

Albantakis, L., Marshall, W., Hoel, E., & Tononi, G. (2019). What caused what? A quantitative account of actual causation using dynamical causal networks. *Entropy. An International and Interdisciplinary Journal of Entropy and Information Studies*, *21*(5), 459. https://doi.org/10.3390/e21050459

Albantakis, L., Prentner, R., & Durham, I. (2023). Measuring the integrated information of a quantum mechanism. *Entropy. An International and Interdisciplinary Journal of Entropy and Information Studies*, *25*(3).

Albantakis, L., & Tononi, G. (2015). The Intrinsic Cause-Effect Power of Discrete Dynamical Systems—From Elementary Cellular Automata to Adapting Animats. *Entropy*, *17*(12), 5472–5502. https://doi.org/10.3390/e17085472

Albantakis, L., & Tononi, G. (2017). Chapter 14: Automata and animats. *From Matter to Life: Information and Causality*, 334. https://doi.org/10.1017/9781316584200.014

Albantakis, L., & Tononi, G. (2019). Causal Composition: Structural Differences among Dynamically Equivalent Systems. *Entropy*, *21*(10), 989. https://doi.org/10.3390/e21100989

Albantakis, L., & Tononi, G. (2021, January). *What we are is more than what we do*.

Amari, S.-i. (1983). A foundation of information geometry. *Electronics and Communications in Japan (Part I: Communications)*, *66*(6), 1–10. https://doi.org/10.1002/ecja.4400660602

Amari, S.-i. (2016). *Information geometry and its applications* (Vol. 194). Springer.

Amari, S.-i., Tsuchiya, N., & Oizumi, M. (2018). Geometry of Information Integration. In *Information Geometry and Its Applications* (pp. 3–17). https://doi.org/10.1007/978-3-319-97798-0_1

Aru, J., Siclari, F., Phillips, W. A., & Storm, J. F. (2020). Apical drive—A cellular mechanism of dreaming? *Neuroscience and biobehavioral reviews*, *119*, 440–455. https://doi.org/10.1016/j.neubiorev.2020.09.018

Ay, N., & Polani, D. (2008). Information flows in causal networks. *Advances in Complex Systems*, *11*(01), 17–41. https://doi.org/10.1142/S0219525908001465

Baker, B., Lansdell, B., & Kording, K. P. (2022). Three aspects of representation in neuroscience. *Trends in Cognitive Sciences*, *26*(11), 942–958. https://doi.org/10.1016/j.tics.2022.08.014

Balduzzi, D., & Tononi, G. (2008). Integrated Information in Discrete Dynamical Systems: Motivation and Theoretical Framework. *PLOS Computational Biology*, *4*(6), e1000091. https://doi.org/10.1371/journal.pcbi.1000091

Balduzzi, D., & Tononi, G. (2009). Qualia: The geometry of integrated information. *PLoS computational biology*, *5*(8), e1000462. https://doi.org/10.1371/journal.pcbi.1000462

Balduzzi, D., & Tononi, G. (2013). What can neurons do for their brain? Communicate selectivity with bursts. *Theory in Biosciences*, *132*(1), 27–39. https://doi.org/10.1007/s12064-012-0165-0

Barbosa, L. S., Marshall, W., Albantakis, L., & Tononi, G. (2021). Mechanism Integrated Information. *Entropy*, *23*(3), 362. https://doi.org/10.3390/e23030362

Barbosa, L. S., Marshall, W., Streipert, S., Albantakis, L., & Tononi, G. (2020). A measure for intrinsic information. *Scientific Reports*, *10*(1), 18803. https://doi.org/10.1038/s41598-020-75943-4

Barrett, A. (2014). An integration of integrated information theory with fundamental physics. *Frontiers in Psychology*, *5*.

Barrett, A. B., & Mediano, P. A. (2019). The phi measure of integrated information is not well-defined for general physical systems. *Journal of Consciousness Studies*, *26*(1-2), 11–20.

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

Barth, A. L., & Poulet, J. F. A. (2012). Experimental evidence for sparse firing in the neocortex. *Trends in Neurosciences*, *35*(6), 345–355. https://doi.org/10.1016/j.tins.2012.03.008

Barttfeld, P., Uhrig, L., Sitt, J. D., Sigman, M., Jarraya, B., & Dehaene, S. (2015). Signature of consciousness in the dynamics of resting-state brain activity. *Proceedings of the National Academy of Sciences*, *112*(3), 887–892. https://doi.org/10.1073/pnas.1418031112

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, *364*(6439), eaav9436. https://doi.org/10.1126/science.aav9436

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bayne, T. (2018). On the axiomatic foundations of the integrated information theory of consciousness. *Neuroscience of Consciousness*, *2018*(1). https://doi.org/10.1093/nc/niy007

Beer, R. D., & Williams, P. L. (2015). Information processing and dynamics in minimally cognitive agents. *Cognitive Science*, *39*(1), 1–38. https://doi.org/10.1111/cogs.12142

Ben-Kiki, O., Evans, C., & Net, I. döt. (2009). *YAML specification*. http://yaml.org/spec/

Bennett, C., Arroyo, S., & Hestrin, S. (2013). Subthreshold Mechanisms Underlying State-Dependent Modulation of Visual Responses. *Neuron*, *80*(2), 350–357. https://doi.org/10.1016/j.neuron.2013.08.007

Binzegger, T., Douglas, R. J., & Martin, K. A. C. (2009). Topology and dynamics of the canonical circuit of cat V1. *Neural networks: the official journal of the International Neural Network Society*, *22*(8), 1071–1078. https://doi.org/10.1016/j.neunet.2009.07.011

Binzegger, T., Douglas, R. J., & Martin, K. A. C. (2004). A quantitative map of the circuit of cat primary visual cortex. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, *24*(39), 8441–8453. https://doi.org/10.1523/JNEUROSCI.1400-04.2004

Boly, M. (in preparation). *Neural correlates of pure presence*.

Boly, M., Massimini, M., Tsuchiya, N., Postle, B. R., Koch, C., & Tononi, G. (2017). Are the Neural Correlates of Consciousness in the Front or in the Back of the Cerebral Cortex? Clinical and Neuroimaging Evidence. *The Journal of Neuroscience*, *37*(40), 9603–9613. https://doi.org/10.1523/JNEUROSCI.3218-16.2017

Boly, M., Sasai, S., Gosseries, O., Oizumi, M., Casali, A., Massimini, M., & Tononi, G. (2015). Stimulus Set Meaningfulness and Neurophysiological Differentiation: A Functional Magnetic Resonance Imaging Study. *PLOS ONE*, *10*(5), e0125337. https://doi.org/10.1371/journal.pone.0125337

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, *30*, 31–40.

Boyd, C. a. R. (2010). Cerebellar agenesis revisited. *Brain*, *133*(3), 941–944. https://doi.org/10.1093/brain/awp265

Brette, R. (2019). Is coding a relevant metaphor for the brain? *Behavioral and Brain Sciences*, *42*. https://doi.org/10.1017/S0140525X19000049

Bruner, J. S. (1973, April). *Beyond the information given: Studies in the psychology of knowing*. W. W. Norton, Incorporated.

Buzsáki, G. (2019). *The brain from inside out*. Oxford University Press.

Buzsáki, G. (2020). The Brain–Cognitive Behavior Problem: A Retrospective. *eNeuro*, *7*(4). https://doi.org/10.1523/ENEURO.0069-20.2020

Cao, R. (2020). New labels for old ideas: Predictive processing and the interpretation of neural signals. *Review of philosophy and psychology*, *11*(3), 517–546. https://doi.org/10.1007/s13164-020-00481-x

Carroll, S. (2021). Consciousness and the laws of physics. *Journal of Consciousness Studies*, *28*(9), 16–31. https://doi.org/10.53765/20512201.28.9.016

Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., Casarotto, S., Bruno, M.-A., Laureys, S., Tononi, G., & Massimini, M. (2013). A theoretically based index of consciousness independent of sensory processing and behavior. *Science translational medicine*, *5*(198), 198ra105. https://doi.org/10.1126/scitranslmed.3006294

Casarotto, S., Comanducci, A., Rosanova, M., Sarasso, S., Fecchio, M., Napolitani, M., Pigorini, A., Casali, A. G., Trimarchi, P. D., Boly, M., Gosseries, O., Bodart, O., Curto, F., Landi, C., Mariotti, M., Devalle, G., Laureys, S., Tononi, G., & Massimini, M. (2016). Stratification of unresponsive patients by an independently validated index of brain complexity. *Annals of Neurology*, *80*(5), 718–729. https://doi.org/10.1002/ana.24779

Castrillon, G., Epp, S., Bose, A., Fraticelli, L., Hechler, A., Belenya, R., Ranft, A., Yakushev, I., Utz, L., Sundar, L., Rauschecker, J. P., Preibisch, C., Kurcyus, K., & Riedl, V. (2023). An energy costly architecture of neuromodulators for human brain evolution and cognition. *bioRxiv : the preprint server for biology*. https://doi.org/10.1101/2023.04.25.538209

Chalmers, D. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, *2*(3), 200–219.

Chance, F. S., Nelson, S. B., & Abbott, L. F. (1999). Complex cells as cortically amplified simple cells. *Nature Neuroscience*, *2*(3), 277–282. https://doi.org/10.1038/6381

Chen, T.-W., Wardill, T. J., Sun, Y., Pulver, S. R., Renninger, S. L., Baohan, A., Schreiter, E. R., Kerr, R. A., Orger, M. B., Jayaraman, V., Looger, L. L., Svoboda, K., & Kim, D. S. (2013). Ultra-sensitive fluorescent proteins for imaging neuronal activity. *Nature*, *499*(7458), 295–300. https://doi.org/10.1038/nature12354

Citri, A., & Malenka, R. C. (2008). Synaptic plasticity: Multiple forms, functions, and mechanisms. *Neuropsychopharmacology: official publication of the American College of Neuropsychopharmacology*, *33*(1), 18–41. https://doi.org/10.1038/sj.npp.1301559

Clancy, K. B., & Mrsic-Flogel, T. D. (2020). The sensory representation of causally controlled objects. *Neuron*. https://doi.org/10.1016/j.neuron.2020.12.001

Cogitate Consortium, Ferrante, O., Gorska-Klimowska, U., Henin, S., Hirschhorn, R., Khalaf, A., Lepauvre, A., Liu, L., Richter, D., Vidal, Y., Bonacchi, N., Brown, T., Sripad, P., Armendariz, M., Bendtz, K., Ghafari, T., Hetenyi, D., Jeschke, J., Kozma, C., … Melloni, L. (2023, June 30). *An* *adversarial collaboration to critically evaluate theories of consciousness*. https://doi.org/10.1101/2023.06.23.546249

Comolatti, R., & Grasso, M. (in preparation). *Why does time feel flowing?*

Comolatti, R., Pigorini, A., Casarotto, S., Fecchio, M., Faria, G., Sarasso, S., Rosanova, M., Gosseries, O., Boly, M., Bodart, O., Ledoux, D., Brichant, J.-F., Nobili, L., Laureys, S., Tononi, G., Massimini, M., & Casali, A. G. (2019). A fast and general method to empirically estimate the complexity of brain responses to transcranial and intracranial stimulations: Supplemental text, figures and table. *bioRxiv*. https://doi.org/10.1101/445882

Cooper, J. (1997). *Plato: Complete works*. Hackett.

Cottaris, N. P., & De Valois, R. L. (1998). Temporal dynamics of chromatic tuning in macaque primary visual cortex. *Nature*, *395*(6705), 896–900. https://doi.org/10.1038/27666

Dadarlat, M. C., & Stryker, M. P. (2017). Locomotion Enhances Neural Encoding of Visual Stimuli in Mouse V1. *The Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.2728-16.2017

Dagan, I., Marcus, S., & Markovitch, S. (1993). Contextual word similarity and estimation from sparse data. *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*, 164–171. https://doi.org/10.3115/981574.981596

Dehaene, S., Changeux, J.-P., & Naccache, L. (2011). The global neuronal workspace model of conscious access: From neuronal architectures to clinical applications. In *Characterizing consciousness: From cognition to the clinic?* (pp. 55–84). https://doi.org/10.1007/978-3-642-18015-6_4

Deneux, T., Kaszas, A., Szalay, G., Katona, G., Lakner, T., Grinvald, A., Rózsa, B., & Vanzetta, I. (2016). Accurate spike estimation from noisy calcium signals for ultrafast three-dimensional imaging of large neuronal populations in vivo. *Nature Communications*, *7*(1), 12190. https://doi.org/10.1038/ncomms12190

de Vries, S. E. J., Lecoq, J. A., Buice, M. A., Groblewski, P. A., Ocker, G. K., Oliver, M., Feng, D., Cain, N., Ledochowitsch, P., Millman, D., Roll, K., Garrett, M., Keenan, T., Kuan, L., Mihalas, S., Olsen, S., Thompson, C., Wakeman, W., Waters, J., … Koch, C. (2020). A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nature Neuroscience*, *23*(1), 138–151. https://doi.org/10.1038/s41593-019-0550-9

Dijkstra, E. W. (1982). *Why numbering should start at zero (EWD 831)*. https://www.cs.utexas.edu/users/EWD/transcriptions/EWD08xx/EWD831.html

Ding, Z., Tran, D. T., Ponder, K., Cobos, E., Ding, Z., Fahey, P. G., Wang, E., Muhammad, T., Fu, J., Cadena, S. A., Papadopoulos, S., Patel, S., Franke, K., Reimer, J., Sinz, F. H., Ecker, A. S., Pitkow, X., & Tolias, A. S. (2023, March 16). *Bipartite invariance in mouse primary visual cortex*. https://doi.org/10.1101/2023.03.15.532836

Doerig, A., Schurger, A., & Herzog, M. H. (2021). Hard criteria for empirical theories of consciousness. *Cognitive Neuroscience*, *12*(2), 41–62. https://doi.org/10.1080/17588928.2020.1772214

Doerig, A., Schurger, A., Hess, K., & Herzog, M. H. (2019). The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness. *Consciousness and Cognition*, *72*, 49–59. https://doi.org/10.1016/j.concog.2019.04.002

Douglas, R. J., Koch, C., Mahowald, M., Martin, K. A. C., & Suarez, H. H. (1995). Recurrent Excitation in Neocortical Circuits. *Science*, *269*(5226), 981–985. https://doi.org/10.1126/science.7638624

Ebitz, R. B., & Hayden, B. Y. (2021). The population doctrine in cognitive neuroscience. *Neuron*, *109*(19), 3055–3068. https://doi.org/10.1016/j.neuron.2021.07.011

Ellia, F., Hendren, J., Grasso, M., Kozma, C., Mindt, G., P. Lang, J., M. Haun, A., Albantakis, L., Boly, M., & Tononi, G. (2021). Consciousness and the fallacy of misplaced objectivity. *Neuroscience of Consciousness*, *2021*(2), niab032. https://doi.org/10.1093/nc/niab032

Erlich, J. C., Brunton, B. W., Duan, C. A., Hanks, T. D., & Brody, C. D. (2015). Distinct effects of prefrontal and parietal cortex inactivations on an accumulation of evidence task in the rat. *eLife*, *4*, e05457. https://doi.org/10.7554/eLife.05457

Esteban, F. J., Galadí, J. A., Langa, J. A., Portillo, J. R., & Soler-Toscano, F. (2018). Informational structures: A dynamical system approach for integrated information. *PLOS Computational Biology*, *14*(9), e1006154. https://doi.org/10.1371/journal.pcbi.1006154

Fahrenfort, J. J., & van Gaal, S. (2021). Criteria for empirical theories of consciousness should focus on the explanatory power of mechanisms, not on functional equivalence. *Cognitive Neuroscience*, *12*(2), 93–94. https://doi.org/10.1080/17588928.2020.1838470

Favela, L. H. (2021). The dynamical renaissance in neuroscience. *Synthese*, *199*(1), 2103–2127. https://doi.org/10.1007/s11229-020-02874-y

Feldman, J. (2003). A catalog of Boolean concepts. *Journal of Mathematical Psychology*, *47*(1), 75–89. https://doi.org/10.1016/S0022-2496(02)00025-1

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America. A, Optics and image science*, *4*(12), 2379–2394. https://doi.org/10.1364/josaa.4.002379

Findlay, G., Marshall, W., Albantakis, L., Mayner, W. G. P., Koch, C., & Tononi, G. (2019). Dissociating intelligence from consciousness in artificial systems – implications of integrated information theory. *Proceedings of the 2019 towards Conscious AI Systems Symposium, AAAI SSS19*.

Findlay, G., Marshall, W., Albantakis, L., Mayner, W. G. P., Koch, C., & Tononi, G. (in preparation). *Dissociating intelligence from consciousness in artificial systems – implications of integrated information theory*.

Foster, N. N., Barry, J., Korobkova, L., Garcia, L., Gao, L., Becerra, M., Sherafat, Y., Peng, B., Li, X., Choi, J.-H., et al. (2021). The mouse cortico–basal ganglia–thalamic network. *Nature*, *598*(7879), 188–194.

Fu, J., Shrinivasan, S., Ponder, K., Muhammad, T., Ding, Z., Wang, E., Ding, Z., Tran, D. T., Fahey, P. G., Papadopoulos, S., Patel, S., Reimer, J., Ecker, A. S., Pitkow, X., Haefner, R. M., Sinz, F. H., Franke, K., & Tolias, A. S. (2023, March 14). *Pattern completion and disruption characterize contextual modulation in mouse visual cortex*. https://doi.org/10.1101/2023.03.13.532473

Gandhi, S. R., Mayner, W. G. P., Marshall, W., Billeh, Y. N., Bennett, C., Gale, S. D., Mochizuki, C., Siegle, J. H., Olsen, S., Tononi, G., Koch, C., & Arkhipov, A. (2023). A survey of neurophysiological differentiation across mouse visual brain areas and timescales. *Frontiers in Computational Neuroscience*, *17*. https://doi.org/10.3389/fncom.2023.1040629

Ganea, D. A., Bexter, A., Günther, M., Gardères, P.-M., Kampa, B. M., & Haiss, F. (2020). Pupillary Dilations of Mice Performing a Vibrotactile Discrimination Task Reflect Task Engagement and Response Confidence. *Frontiers in Behavioral Neuroscience*, *14*. https://doi.org/10.3389/fnbeh.2020.00159

Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. The MIT Press. https://doi.org/10.7551/mitpress/2076.001.0001

Glickfeld, L. L., & Olsen, S. R. (2017). Higher-Order Areas of the Mouse Visual Cortex. *Annual Review of Vision Science*, *3*(1), 251–273. https://doi.org/10.1146/annurev-vision-102016-061331

Goering, S., & Yuste, R. (2016). On the Necessity of Ethical Guidelines for Novel Neurotechnologies. *Cell*, *167*(4), 882–885. https://doi.org/10.1016/j.cell.2016.10.029

Goh, R. Z., Phillips, I. B., & Firestone, C. (2023). The perception of silence. *Proceedings of the National Academy of Sciences*, *120*(29), e2301463120. https://doi.org/10.1073/pnas.2301463120

Gomez, J. D., Mayner, W. G. P., Beheler-Amass, M., Tononi, G., & Albantakis, L. (2020). Computing Integrated Information (Φ) in Discrete Dynamical Systems with Multi-Valued Elements. *Entropy*, *23*(1), 6. https://doi.org/10.3390/e23010006

Gosseries, O., Schnakers, C., Ledoux, D., Vanhaudenhuyse, A., Bruno, M.-A., Demertzi, A., Noirhomme, Q., Lehembre, R., Damas, P., Goldman, S., Peeters, E., Moonen, G., & Laureys, S. (2011). Automated EEG entropy measurements in coma, vegetative state/unresponsive wakefulness syndrome and minimally conscious state. *Functional Neurology*, *26*(1), 25–30.

Grasso, M. (in preparation). *How do phenomenal objects bind general concepts with particular features?*

Grasso, M., Albantakis, L., Lang, J. P., & Tononi, G. (2021). Causal reductionism and causal structures. *Nature Neuroscience*, *24*(10), 1348–1355. https://doi.org/10.1038/s41593-021-00911-8

Groblewski, P. A., Sullivan, D., Lecoq, J., de Vries, S. E. J., Caldejon, S., L'Heureux, Q., Keenan, T., Roll, K., Slaughterback, C., Williford, A., & Farrell, C. (2020). A standardized head-fixation system for performing large-scale, in vivo physiological recordings in mice. *Journal of Neuroscience Methods*, *346*, 108922. https://doi.org/10.1016/j.jneumeth.2020.108922

Hanson, J. R., & Walker, S. I. (2021). Formalizing falsification for theories of consciousness across computational hierarchies. *Neuroscience of Consciousness*, *2021*(2), niab014. https://doi.org/10.1093/nc/niab014

Hanson, J. R., & Walker, S. I. (2023). On the non-uniqueness problem in integrated information theory. *Neuroscience of Consciousness*, *2023*(1), niad014. https://doi.org/10.1093/nc/niad014

Harris, C. R., Millman, K. J., Walt, S. J. van der, Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Kerkwijk, M. H. van, Brett, M., Haldane, A., R'ıo, J. F. del, Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Harris, J. A., Mihalas, S., Hirokawa, K. E., Whitesell, J. D., Choi, H., Bernard, A., Bohn, P., Caldejon, S., Casal, L., Cho, A., Feiner, A., Feng, D., Gaudreault, N., Gerfen, C. R., Graddis, N., Groblewski, P. A., Henry, A. M., Ho, A., Howard, R., … Zeng, H. (2019). Hierarchical organization of cortical and thalamic connectivity. *Nature*, *575*(7781), 195–202. https://doi.org/10.1038/s41586-019-1716-z

Hashmi, A., Nere, A., & Tononi, G. (2013). Sleep-Dependent Synaptic Down-Selection (II): Single-Neuron Level Benefits for Matching, Selectivity, and Specificity. *Frontiers in Neurology*, *4*. https://doi.org/10.3389/fneur.2013.00148

Haun, A., & Tononi, G. (2019). Why Does Space Feel the Way it Does? Towards a Principled Account of Spatial Experience. *Entropy*, *21*(12), 1160. https://doi.org/10.3390/e21121160

Hayden, B. Y., & Niv, Y. (2021a). The case against economic values in the orbitofrontal cortex (or anywhere else in the brain). *Behavioral Neuroscience*, *135*(2), 192–201. https://doi.org/10.1037/bne0000448

Hayden, B. Y., & Niv, Y. (2021b). The case against economic values in the orbitofrontal cortex (or anywhere else in the brain). *Behavioral Neuroscience*, *135*(2), 192–201. https://doi.org/10.1037/bne0000448

Hebb, D. (2002). *The organization of behavior: A neuropsychological theory (reprint, ebook)*. Psychology Press.

Hires, S. A., Gutnisky, D. A., Yu, J., O'Connor, D. H., & Svoboda, K. (2015). Low-noise encoding of active touch by layer 4 in the somatosensory cortex. *eLife*, *4*, e06619.

Hoel, E. P., Albantakis, L., Marshall, W., & Tononi, G. (2016). Can the macro beat the micro? Integrated information across spatiotemporal scales. *Neuroscience of Consciousness*, *2016*(1), niw012. https://doi.org/10.1093/nc/niw012

Hoel, E. P., Albantakis, L., & Tononi, G. (2013). Quantifying causal emergence shows that macro can beat micro. *PNAS*, *110*(49), 19790–19795. https://doi.org/10.1073/pnas.1314922110

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous Inference in General Parametric Models. *Biometrical Journal*, *50*(3), 346–363. https://doi.org/10.1002/bimj.200810425

Huang, L., Ledochowitsch, P., Knoblich, U., Lecoq, J., Murphy, G. J., Reid, R. C., de Vries, S. E., Koch, C., Zeng, H., Buice, M. A., Waters, J., & Li, L. (2021). Relationship between simultaneously recorded spiking activity and fluorescence signal in GCaMP6 transgenic mice. *eLife*, *10*, e51675. https://doi.org/10.7554/eLife.51675

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, *148*(3), 574–591. https://doi.org/10.1113/jphysiol.1959.sp006308

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, *160*(1), 106–154. https://doi.org/10.1113/jphysiol.1962.sp006837

Hudetz, A. G., Liu, X., & Pillay, S. (2014). Dynamic Repertoire of Intrinsic Brain States Is Reduced in Propofol-Induced Unconsciousness. *Brain Connectivity*, *5*(1), 10–22. https://doi.org/10.1089/brain.2014.0230

Hudetz, A. G., Liu, X., Pillay, S., Boly, M., & Tononi, G. (2016). Propofol anesthesia reduces Lempel-Ziv complexity of spontaneous brain activity in rats. *Neuroscience Letters*, *628*, 132–135. https://doi.org/10.1016/j.neulet.2016.06.017

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Intrinsic Ontology Wiki. (in preparation).

Ito, S., Oizumi, M., & Amari, S.-i. (2020). Unified framework for the entropy production and the stochastic interaction based on information geometry. *Physical Review Research*, *2*(3), 033048. https://doi.org/10.1103/PhysRevResearch.2.033048

Jacobs, E. A. K., Steinmetz, N. A., Peters, A. J., Carandini, M., & Harris, K. D. (2020). Cortical State Fluctuations during Sensory Decision Making. *Current Biology*, *30*(24), 4944–4955.e7. https://doi.org/10.1016/j.cub.2020.09.067

Janzing, D., Balduzzi, D., Grosse-Wentrup, M., & Schölkopf, B. (2013). Quantifying causal influences. *The Annals of Statistics*, *41*(5), 2324–2358. https://doi.org/10.1214/13-AOS1145

Jewell, S., & Witten, D. (2018). Exact spike train inference via L0 optimization. *The Annals of Applied Statistics*, *12*(4), 2457–2482. https://doi.org/10.1214/18-AOAS1162

Jewell, S. W., Hocking, T. D., Fearnhead, P., & Witten, D. M. (2020). Fast nonconvex deconvolution of calcium imaging data. *Biostatistics*, *21*(4), 709–726. https://doi.org/10.1093/biostatistics/kxy083

Jin, M., & Glickfeld, L. L. (2020). Mouse Higher Visual Areas Provide Both Distributed and Specialized Contributions to Visually Guided Behaviors. *Current Biology*, *30*(23), 4682–4692.e7. https://doi.org/10.1016/j.cub.2020.09.015

Kalita, P., Langa, J. A., & Soler-Toscano, F. (2019). Informational structures and informational fields as a prototype for the description of postulates of the integrated information theory. *Entropy. An*

International and Interdisciplinary Journal of Entropy and Information Studies, *21*(5), 493. https://doi.org/10.3390/e21050493

Kandel, E. R., & Tauc, L. (1965). Mechanism of heterosynaptic facilitation in the giant cell of the abdominal ganglion of Aplysia depilans. *The Journal of physiology*, *181*(1), 28–47. https://doi.org/10.1113/jphysiol.1965.sp007743

Kanwisher, N. (2010). Functional specificity in the human brain: A window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences*, *107*(25), 11163–11170.

Katz, L. N., Yates, J. L., Pillow, J. W., & Huk, A. C. (2016). Dissociated functional significance of decision-related activity in the primate dorsal stream. *Nature*, *535*(7611), 285–288. https://doi.org/10.1038/nature18617

Keller, A. J., Roth, M. M., & Scanziani, M. (2020). Feedback generates a second receptive field in neurons of the visual cortex. *Nature*, *582*(7813), 545–549. https://doi.org/10.1038/s41586-020-2319-4

Khosla, M., & Wehbe, L. (2022). High-level visual areas act like domain-general filters with strong selectivity and functional specialization. *bioRxiv : the preprint server for biology*. https://doi.org/10.1101/2022.03.16.484578

Kitazono, J., Kanai, R., & Oizumi, M. (2018). Efficient Algorithms for Searching the Minimum Information Partition in Integrated Information Theory. *Entropy*, *20*(3), 173. https://doi.org/10.3390/e20030173

Kleiner, J. (2020). Brain states matter. A reply to the unfolding argument. *Consciousness and Cognition*, *85*, 102981. https://doi.org/10.1016/j.concog.2020.102981

Kleiner, J., & Hoel, E. (2021). Falsification and consciousness. *Neuroscience of Consciousness*, *2021*(1), niab001. https://doi.org/10.1093/nc/niab001

Kleiner, J., & Tull, S. (2021). The mathematical structure of integrated information theory. *Frontiers in Applied Mathematics and Statistics*, *6*, 74. https://doi.org/10.3389/FAMS.2020.602973/BIBTEX

Koch, C. (2019). *The Feeling of Life Itself: Why Consciousness is Widespread But Can't be Computed*. Mit Press.

Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: Progress and problems. *Nature Reviews Neuroscience*, *17*, 307–321. https://doi.org/10.1038/nrn.2016.22

Korb, K. B., Nyberg, E. P., & Hope, L. (2011). A new causal power theory. In *Causality in the sciences*. https://doi.org/10.1093/acprof:oso/9780199574131.003.0030

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*, 4. https://doi.org/10.3389/neuro.06.004.2008

Krohn, K., & Rhodes, J. (1965). Algebraic theory of machines. I. Prime decomposition theorem for finite semigroups and machines. *Transactions of the American Mathematical Society*, *116*, 450. https://doi.org/10.2307/1994127

Krohn, S., & Ostwald, D. (2017). Computing integrated information. *Neuroscience of Consciousness*, *2017*. https://doi.org/10.1093/nc/nix017

Lamme, V. A., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in neurosciences*, *23*(11), 571–579. https://doi.org/10.1016/s0166-2236(00)01657-x

Larkum, M. E., Zhu, J. J., & Sakmann, B. (1999). A new cellular mechanism for coupling inputs arriving at different cortical layers. *Nature*, *398*(6725), 338–341. https://doi.org/10.1038/18686

Larkum, M. E., Nevian, T., Sandler, M., Polsky, A., & Schiller, J. (2009). Synaptic integration in tuft dendrites of layer 5 pyramidal neurons: A new unifying principle. *Science (New York, N.Y.)*, *325*(5941), 756–760. https://doi.org/10.1126/science.1171958

Larkum, M. E., Senn, W., & Lüscher, H.-R. (2004). Top-down dendritic input increases the gain of layer 5 pyramidal neurons. *Cerebral cortex*, *14*(10), 1059–1070. https://doi.org/10.1093/cercor/bhh065

Larsen, R. S., & Waters, J. (2018). Neuromodulatory Correlates of Pupil Dilation. *Frontiers in Neural Circuits*, *12*. https://doi.org/10.3389/fncir.2018.00021

Lavazza, A., & Massimini, M. (2018). Cerebral organoids: Ethical issues and consciousness assessment. *Journal of Medical Ethics*, *44*(9), 606–610. https://doi.org/10.1136/medethics-2017-104555

Ledochowitsch, P., Huang, L., Knoblich, U., Oliver, M., Lecoq, J., Reid, C., Li, L., Zeng, H., Koch, C., Waters, J., Vries, S. E. J. de, & Buice, M. A. (2019). On the correspondence of electrical and optical physiology in in vivo population-scale two-photon calcium imaging. *bioRxiv*, 800102. https://doi.org/10.1101/800102

Lemon, R. N., & Edgley, S. A. (2010). Life without a cerebellum. *Brain : a journal of neurology*, *133*(3), 652–654.

Lenth, R. (2020). *Emmeans: Estimated Marginal Means, aka Least-Squares Means, v1.5.1.*

Leung, A., & Tsuchiya, N. (2023). Separating weak integrated information theory into inspired and aspirational approaches. *Neuroscience of Consciousness*, *2023*(1), niad012. https://doi.org/10.1093/nc/niad012

Levine, J. (1983). Materialism and Qualia: The Explanatory Gap. *Pacific Philosophical Quarterly*, *64*(4), 354–361. https://doi.org/10.1111/j.1468-0114.1983.tb00207.x

Li, L.-y., Li, Y.-t., Zhou, M., Tao, H. W., & Zhang, L. I. (2013). Intracortical multiplication of thalamocortical signals in mouse auditory cortex. *Nature Neuroscience*, *16*(9), 1179–1181. https://doi.org/10.1038/nn.3493

Li, Y.-t., Ibrahim, L. A., Liu, B.-h., Zhang, L. I., & Tao, H. W. (2013). Linear transformation of thalamocortical input by intracortical excitation. *Nature Neuroscience*, *16*(9), 1324–1330. https://doi.org/10.1038/nn.3494

Lien, A. D., & Scanziani, M. (2013). Tuned thalamic excitation is amplified by visual cortical circuits. *Nature Neuroscience*, *16*(9), 1315–1323. https://doi.org/10.1038/nn.3488

Liu, L. D., & Pack, C. C. (2017). The Contribution of Area MT to Visual Motion Perception Depends on Training. *Neuron*, *95*(2), 436–446.e3. https://doi.org/10.1016/j.neuron.2017.06.024

Luppi, A. I., Mediano, P. A. M., Rosas, F. E., Allanson, J., Pickard, J. D., Carhart-Harris, R. L., Williams, G. B., Craig, M. M., Finoia, P., Owen, A. M., Naci, L., Menon, D. K., Bor, D., & Stamatakis, E. A. (2023, March 28). *A Synergistic Workspace for Human Consciousness Revealed by Integrated Information Decomposition*. https://doi.org/10.1101/2020.11.25.398081

Luppi, A. I., Mediano, P. A. M., Rosas, F. E., Harrison, D. J., Carhart-Harris, R. L., Bor, D., & Stamatakis, E. A. (2021). What it is like to be a bit: An integrated information decomposition account of emergent mental phenomena. *Neuroscience of Consciousness*, *2021*(2), niab027. https://doi.org/10.1093/nc/niab027

Madisen, L., Zwingman, T. A., Sunkin, S. M., Oh, S. W., Zariwala, H. A., Gu, H., Ng, L. L., Palmiter, R. D., Hawrylycz, M. J., Jones, A. R., Lein, E. S., & Zeng, H. (2010). A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. *Nature Neuroscience*, *13*(1), 133–140. https://doi.org/10.1038/nn.2467

Mainen, Z. F., & Sejnowski, T. J. (1995). Reliability of spike timing in neocortical neurons. *Science (New York, N.Y.)*, *268*(5216), 1503–1506.

Markov, N. T., Misery, P., Falchier, A., Lamy, C., Vezoli, J., Quilodran, R., Gariel, M. A., Giroud, P., Ercsey-Ravasz, M., Pilaz, L. J., Huissoud, C., Barone, P., Dehay, C., Toroczkai, Z., Van Essen, D. C., Kennedy, H., & Knoblauch, K. (2011). Weight consistency specifies regularities of macaque cortical networks. *Cerebral cortex*, *21*(6), 1254–1272. https://doi.org/10.1093/cercor/bhq201

Marshall, W., Albantakis, L., & Tononi, G. (2018). Black-boxing and cause-effect power. *PLOS Computational Biology*, *14*(4), e1006114. https://doi.org/10.1371/journal.pcbi.1006114

Marshall, W., Gomez-Ramirez, J., & Tononi, G. (2016). Integrated Information and State Differentiation. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00926

Marshall, W., Grasso, M., Mayner, W. G. P., Zaeemzadeh, A., Barbosa, L. S., Chastain, E., Findlay, G., Sasai, S., Albantakis, L., & Tononi, G. (2023). System Integrated Information. *Entropy*, *25*(2), 334. https://doi.org/10.3390/e25020334

Marshall, W., Kim, H., Walker, S. I., Tononi, G., & Albantakis, L. (2017). How causal analysis can reveal autonomy in models of biological systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*.

Marshel, J. H., Garrett, M. E., Nauhaus, I., & Callaway, E. M. (2011). Functional specialization of seven mouse visual cortical areas. *Neuron*. https://doi.org/10.1016/j.neuron.2011.12.004

Mashour, G. A., Roelfsema, P., Changeux, J.-P., & Dehaene, S. (2020). Conscious Processing and the Global Neuronal Workspace Hypothesis. *Neuron*, *105*(5), 776–798. https://doi.org/10.1016/j.neuron.2020.01.026

Massimini, M., Ferrarelli, F., Sarasso, S., & Tononi, G. (2012). Cortical mechanisms of loss of consciousness: Insight from TMS/EEG studies. *Archives Italiennes De Biologie*, *150*(2-3), 44–55. https://doi.org/10.4449/aib.v150i2.1361

Massimini, M., Ferrarelli, F., Huber, R., Esser, S. K., Singh, H., & Tononi, G. (2005). Breakdown of cortical effective connectivity during sleep. *Science (New York, N.Y.)*, *309*(5744), 2228–2232.

Mayner, W. G. P., Marshall, W., Albantakis, L., Findlay, G., Marchman, R., & Tononi, G. (2018). PyPhi: A toolbox for integrated information theory. *PLoS computational biology*, *14*(7), e1006343.

Mayner, W. G. P., Marshall, W., Billeh, Y. N., Gandhi, S. R., Caldejon, S., Cho, A., Griffin, F., Hancock, N., Lambert, S., Lee, E. K., Luviano, J. A., Mace, K., Nayan, C., Nguyen, T. V., North, K., Seid, S., Williford, A., Cirelli, C., Groblewski, P. A., … Arkhipov, A. (2021, December 14). *Dataset: Measuring stimulus-evoked neurophysiological differentiation in distinct populations of neurons in mouse visual cortex*. https://doi.org/10.5281/zenodo.5781567

Mayner, W. G. P., Marshall, W., Billeh, Y. N., Gandhi, S. R., Caldejon, S., Cho, A., Griffin, F., Hancock, N., Lambert, S., Lee, E. K., Luviano, J. A., Mace, K., Nayan, C., Nguyen, T. V., North, K., Seid, S., Williford, A., Cirelli, C., Groblewski, P. A., … Arkhipov, A. (2022). Measuring Stimulus-Evoked Neurophysiological Differentiation in Distinct Populations of Neurons in Mouse Visual Cortex. *eNeuro*, *9*(1). https://doi.org/10.1523/ENEURO.0280-21.2021

McGinley, M. J., David, S. V., & McCormick, D. A. (2015). Cortical Membrane Potential Signature of Optimal States for Sensory Signal Detection. *Neuron*, *87*(1), 179–192. https://doi.org/10.1016/j.neuron.2015.05.038

McGinley, M. J., Vinck, M., Reimer, J., Batista-Brito, R., Zagha, E., Cadwell, C. R., Tolias, A. S., Cardin, J. A., & McCormick, D. A. (2015). Waking State: Rapid Variations Modulate Neural and Behavioral Responses. *Neuron*, *87*(6), 1143–1161. https://doi.org/10.1016/j.neuron.2015.09.012

McInnes, L., Healy, J., & Melville, J. (2020, September 17). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*.

McQueen, K. (2019). Interpretation- Neutral Integrated Information Theory. *Journal of Consciousness Studies*, *26*(1-2), 76–106.

Mediano, P. A. M., Rosas, F., Carhart-Harris, R. L., Seth, A. K., & Barrett, A. B. (2019, September 5). *Beyond integrated information: A taxonomy of information dynamics phenomena*. arXiv: 1909.02297 [physics, q-bio]. https://doi.org/10.48550/arXiv.1909.02297

Mediano, P. A. M., Rosas, F. E., Bor, D., Seth, A. K., & Barrett, A. B. (2022). The strength of weak integrated information theory. *Trends in Cognitive Sciences*, *26*(8), 646–655. `https://doi.org/10.1016/j.tics.2022.04.008`

Mediano, P. A. M., Rosas, F. E., Luppi, A. I., Carhart-Harris, R. L., Bor, D., Seth, A. K., & Barrett, A. B. (2021, September 27). *Towards an extended taxonomy of information dynamics via Integrated Information Decomposition*. arXiv: `2109.13186 [physics, q-bio]`. `https://doi.org/10.48550/arXiv.2109.13186`

Mediano, P. A. M., Rosas, F. E., Luppi, A. I., Jensen, H. J., Seth, A. K., Barrett, A. B., Carhart-Harris, R. L., & Bor, D. (2022). Greater than the parts: A review of the information decomposition approach to causal emergence. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *380*(2227), 20210246. `https://doi.org/10.1098/rsta.2021.0246`

Melloni, L., Mudrik, L., Pitts, M., & Koch, C. (2021). Making the hard problem of consciousness easier. *Science*, *372*(6545), 911–912. `https://doi.org/10.1126/science.abj3259`

Mensen, A., Marshall, W., Sasai, S., & Tononi, G. (2018). Differentiation analysis of continuous electroencephalographic activity triggered by video clip contents. *Journal of cognitive neuroscience*, *30*(8), 1108–1118. `https://doi.org/10.1162/jocn\_a\_01278`

Mensen, A., Marshall, W., & Tononi, G. (2017). EEG Differentiation Analysis and Stimulus Set Meaningfulness. *Frontiers in Psychology*, *8*, 1748. `https://doi.org/10.3389/fpsyg.2017.01748`

Merker, B., Williford, K., & Rudrauf, D. (2022). The integrated information theory of consciousness: A case of mistaken identity. *Behavioral and Brain Sciences*, *45*, e41. `https://doi.org/10.1017/S0140525X21000881`

Michel, M., & Lau, H. (2020). On the dangers of conflating strong and weak versions of a theory of consciousness. *Philosophy and the Mind Sciences*, *1*(2). `https://doi.org/10.33735/phimisci.2020.II.54`

Middleton, F. A., & Strick, P. L. (2000). Basal ganglia and cerebellar loops: Motor and cognitive circuits. *Brain Research Reviews*, *31*(2-3), 236–250.

Moon, K. (2019). Exclusion and Underdetermined Qualia. *Entropy*, *21*(4), 405. `https://doi.org/10.3390/e21040405`

Morch, H. (2019). Is Consciousness Intrinsic?: A Problem for the Integrated Information Theory. *Journal of Consciousness Studies*, *26*(1-2), 133–162.

Moyal, R., Fekete, T., & Edelman, S. (2020). Dynamical emergence theory (DET): A computational account of phenomenal consciousness. *Minds and Machines*, *30*(1), 1–21. `https://doi.org/10.1007/s11023-020-09516-9`

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

Negro, N. (2020). Phenomenology-first versus third-person approaches in the science of consciousness: The case of the integrated information theory and the unfolding argument. *Phenomenology and the Cognitive Sciences*, *19*(5), 979–996. https://doi.org/10.1007/s11097-020-09681-3

Negro, N. (2022a). Axioms and postulates: Finding the right match through logical inference. *Behavioral and Brain Sciences*, *45*, e56. https://doi.org/10.1017/S0140525X2100193X

Negro, N. (2022b). Can the Integrated Information Theory Explain Consciousness from Consciousness Itself? *Review of Philosophy and Psychology*. https://doi.org/10.1007/s13164-022-00653-x

Nere, A., Hashmi, A., Cirelli, C., & Tononi, G. (2013). Sleep-Dependent Synaptic Down-Selection (I): Modeling the Benefits of Sleep on Memory Consolidation and Integration. *Frontiers in Neurology*, *4*. https://doi.org/10.3389/fneur.2013.00143

Niell, C. M., & Stryker, M. P. (2010). Modulation of Visual Responses by Behavioral State in Mouse Visual Cortex. *Neuron*. https://doi.org/10.1016/j.neuron.2010.01.033

Nielsen, T. A. (1999). Mentation during sleep: The NREM/REM distinction. *Handbook of behavioral state control: Cellular and molecular mechanisms*, 101–128.

Nieminen, J. O., Gosseries, O., Massimini, M., Saad, E., Sheldon, A. D., Boly, M., Siclari, F., Postle, B. R., & Tononi, G. (2016). Consciousness and cortical responsiveness: A within-state study during non-rapid eye movement sleep. *Scientific Reports*, *6*, 30932. https://doi.org/10.1038/srep30932

Nir, Y., & Tononi, G. (2010). Dreaming and the brain: From phenomenology to neurophysiology. *Trends in cognitive sciences*, *14*(2), 88–100. https://doi.org/10.1016/j.tics.2009.12.001

Nolte, M., Reimann, M. W., King, J. G., Markram, H., & Muller, E. B. (2019). Cortical reliability amid noise and chaos. *Nature Communications*, *10*(1), 1–15.

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated information theory 3.0. *PLoS computational biology*, *10*(5), e1003588. https://doi.org/10.1371/journal.pcbi.1003588

Oizumi, M., Amari, S.-i., Yanagawa, T., Fujii, N., & Tsuchiya, N. (2016). Measuring Integrated Information from the Decoding Perspective. *PLOS Computational Biology*, *12*(1), e1004654. https://doi.org/10.1371/journal.pcbi.1004654

Oizumi, M., Tsuchiya, N., & Amari, S.-i. (2016). A unified framework for information integration based on information geometry. *Proceedings of the National Academy of Sciences*, *113*(51), 14817–14822. https://doi.org/10.1073/pnas.1603583113

Pachitariu, M., Stringer, C., & Harris, K. D. (2018). Robustness of Spike Deconvolution for Neuronal Calcium Imaging. *The Journal of Neuroscience*, *38*(37), 7976–7985. https://doi.org/10.1523/JNEUROSCI.3339-17.2018

Pearl, J. (2000). *Causality: Models, reasoning and inference* (1st ed.). Cambridge Univ Press.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pearson, J. (2019). The human imagination: The cognitive neuroscience of visual mental imagery. *Nature Reviews Neuroscience*, *20*(10), 624–634. https://doi.org/10.1038/s41583-019-0202-9

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Pele, O., & Werman, M. (2009). Fast and robust Earth Mover's Distances. *2009 IEEE 12th International Conference on Computer Vision*, 460–467. https://doi.org/10.1109/ICCV.2009.5459199

Peron, S., Pancholi, R., Voelcker, B., Wittenbach, J. D., Ólafsdóttir, H. F., Freeman, J., & Svoboda, K. (2020). Recurrent interactions in local cortical circuits. *Nature*, *579*(7798), 256–259. https://doi.org/10.1038/s41586-020-2062-x

Pigorini, A., Sarasso, S., Proserpio, P., Szymanski, C., Arnulfo, G., Casarotto, S., Fecchio, M., Rosanova, M., Mariotti, M., Lo Russo, G., Palva, J. M., Nobili, L., & Massimini, M. (2015). Bistability breaks-off deterministic responses to intracortical stimulation during non-REM sleep. *NeuroImage*, *112*, 105–113. https://doi.org/10.1016/j.neuroimage.2015.02.056

Polack, P.-O., Friedman, J., & Golshani, P. (2013). Cellular mechanisms of brain state–dependent gain modulation in visual cortex. *Nature Neuroscience*, *16*(9), 1331–1339. https://doi.org/10.1038/nn.3464

Ponce, C. R., Xiao, W., Schade, P. F., Hartmann, T. S., Kreiman, G., & Livingstone, M. S. (2019). Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell*, *177*(4), 999–1009.

Putnam, H. (1973). Meaning and reference. *The journal of philosophy*, *70*(19), 699–711. https://doi.org/10.2307/2025079

Quiroga, R. Q., & Panzeri, S. (2009). Extracting information from neuronal populations: Information theory and decoding approaches. *Nature Reviews Neuroscience*, *10*(3), 173–185. https://doi.org/10.1038/nrn2578

Reback, J., McKinney, W., jbrockmendel, Bossche, J. V. den, Augspurger, T., Cloud, P., gfyoung, Sinhrks, Hawkins, S., Klein, A., Roeschke, M., Tratner, J., Petersen, T., She, C., Ayd, W., MomIsBestFriend, Garcia, M., Schendel, J., Hayden, A., … Winkel, M. (2020, October). *Pandas-dev/pandas: Pandas 1.1.3* (Version v1.1.3). https://doi.org/10.5281/zenodo.4067057

Reimer, J., Froudarakis, E., Cadwell, C. R., Yatsenko, D., Denfield, G. H., & Tolias, A. S. (2014). Pupil Fluctuations Track Fast Switching of Cortical States during Quiet Wakefulness. *Neuron*, *84*(2), 355–362. https://doi.org/10.1016/j.neuron.2014.09.033

Ringach, D. L., Hawken, M. J., & Shapley, R. (2003). Dynamics of orientation tuning in macaque V1: The role of global and tuned suppression. *Journal of neurophysiology*, *90*(1), 342–352. https://doi.org/10.1152/jn.01018.2002

Salkoff, D. B., Zagha, E., McCarthy, E., & McCormick, D. A. (2020). Movement and Performance Explain Widespread Cortical Activity in a Visual Detection Task. *Cerebral Cortex*, *30*(1), 421–437. https://doi.org/10.1093/cercor/bhz206

Sarà, M., Pistoia, F., Pasqualetti, P., Sebastiano, F., Onorati, P., & Rossini, P. M. (2011). Functional Isolation Within the Cerebral Cortex in the Vegetative State: A Nonlinear Method to Predict Clinical Outcomes. *Neurorehabilitation and Neural Repair*, *25*(1), 35–42. https://doi.org/10.1177/1545968310378508

Sarasso, S., Casali, A. G., Casarotto, S., Rosanova, M., Sinigaglia, C., & Massimini, M. (2021). Consciousness and complexity: A consilience of evidence. *Neuroscience of Consciousness*. https://doi.org/10.1093/nc/niab023

Sarasso, S., D'Ambrosio, S., Fecchio, M., Casarotto, S., Viganò, A., Landi, C., Mattavelli, G., Gosseries, O., Quarenghi, M., Laureys, S., et al. (2020). Local sleep-like cortical reactivity in the awake brain after focal injury. *Brain : a journal of neurology*, *143*(12), 3672–3684.

Seth, A. K., & Bayne, T. (2022). Theories of consciousness. *Nature Reviews Neuroscience*, *23*(7), 439–452. https://doi.org/10.1038/s41583-022-00587-4

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shirdhonkar, S., & Jacobs, D. W. (2008). Approximate earth mover's distance in linear time. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8. https://doi.org/10.1109/CVPR.2008.4587662

Siclari, F., Baird, B., Perogamvros, L., Bernardi, G., LaRocque, J. J., Riedner, B., Boly, M., Postle, B. R., & Tononi, G. (2017). The neural correlates of dreaming. *Nature Neuroscience*, *20*(6), 872–878. https://doi.org/10.1038/nn.4545

Siegle, J. H., Jia, X., Durand, S., Gale, S., Bennett, C., Graddis, N., Heller, G., Ramirez, T. K., Choi, H., Luviano, J. A., Groblewski, P. A., Ahmed, R., Arkhipov, A., Bernard, A., Billeh, Y. N., Brown, D., Buice, M. A., Cain, N., Caldejon, S., … Koch, C. (2021). Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature*, *592*(7852), 86–92. https://doi.org/10.1038/s41586-020-03171-x
Bandiera_abtest: a Cg_type: Nature Research Journals Primary_atype: Research Subject_term: Neural circuits;Sensory processing;Visual system Subject_term_id: neural-circuit;sensory-processing;visual-system.

Siegle, J. H., Ledochowitsch, P., Jia, X., Millman, D. J., Ocker, G. K., Caldejon, S., Casal, L., Cho, A., Denman, D. J., Durand, S., Groblewski, P. A., Heller, G., Kato, I., Kivikas, S., Lecoq, J., Nayan, C., Ngo, K., Nicovich, P. R., North, K., … de Vries, S. E. (2021). Reconciling functional differences in populations of neurons recorded with two-photon imaging and electrophysiology. *eLife*, *10*, e69068. https://doi.org/10.7554/eLife.69068

Song, C., Haun, A. M., & Tononi, G. (2017). Plasticity in the structure of visual space. *Eneuro*, *4*(3).

Sorrell, E., Rule, M. E., & O'Leary, T. (2021). Brain–Machine Interfaces: Closed-Loop Control in an Adaptive System. *Annual Review of Control, Robotics, and Autonomous Systems*, *4*(1), 167–189. https://doi.org/10.1146/annurev-control-061720-012348

Sporns, O., Tononi, G., & Kötter, R. (2005). The Human Connectome: A Structural Description of the Human Brain. *PLOS Computational Biology*, *1*(4), e42. https://doi.org/10.1371/journal.pcbi.0010042

Steriade, M., Nunez, A., & Amzica, F. (1993). A novel slow ($<$ 1 Hz) oscillation of neocortical neurons in vivo: Depolarizing and hyperpolarizing components. *Journal of Neuroscience*, *13*(8), 3252–3265.

Tegmark, M. (2015). Consciousness as a state of matter. *Chaos, Solitons & Fractals*, *76*, 238–270. https://doi.org/10.1016/j.chaos.2015.03.014

Tegmark, M. (2016). Improved measures of integrated information. *PLoS Computational Biology*, *12*(11), e1005123. https://doi.org/10.1371/journal.pcbi.1005123

Tillemans, T. (2021). Dharmakīrti. In *The Stanford encyclopedia of philosophy* (Spring 2021).

Tingley, D., Yamamoto, T., Hirose, L., Imai, K., Trinh, M., & Wong, W. (2019, September 13). *Mediation: R Package for Causal Mediation Analysis* (Version 4.5.0).

Tononi, G. (2012). Integrated information theory of consciousness: An updated account. *Archives italiennes de biologie*, *150*, 56–90.

Tononi, G. (forthcoming). *On being*.

Tononi, G., Sporns, O., & Edelman, G. M. (1994). A measure for brain complexity: Relating functional segregation and integration in the nervous system. *Proceedings of the National Academy of Sciences*, *91*(11), 5033–5037. https://doi.org/10.1073/pnas.91.11.5033

Tononi, G., Sporns, O., & Edelman, G. M. (1996). A complexity measure for selective matching of signals by the brain. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(8), 3422–3427. https://doi.org/10.1073/pnas.93.8.3422

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, *5*(1), 42. https://doi.org/10.1186/1471-2202-5-42

Tononi, G. (2008). Consciousness as integrated information: A provisional manifesto. *The Biological bulletin*, *215*(3), 216–42. https://doi.org/10.2307/25470707

Tononi, G. (2013). On the irreducibility of consciousness and its relevance to free will. https://doi.org/10.1007/978-1-4614-5212-6_11

Tononi, G. (2014, May 30). *Why Scott Should Stare at a Blank Wall and Reconsider (Or, the Conscious Grid)*. Shtetl-Optimized. `https://scottaaronson.blog/?p=1823`

Tononi, G. (2015). Integrated information theory. *Scholarpedia*, *10*(1), 4164. `https://doi.org/10.4249/scholarpedia.4164`

Tononi, G., Albantakis, L., Boly, M., Cirelli, C., & Koch, C. (2022, June). *Only what exists can cause: An intrinsic view of free will*. `https://doi.org/10.48550/arxiv.2206.02069`

Tononi, G., Boly, M., Grasso, M., Hendren, J., Juel, B. E., Mayner, W. G. P., Marshall, W., & Koch, C. (2022). IIT, half masked and half disfigured. *Behavioral and Brain Sciences*, *45*, e60. `https://doi.org/10.1017/S0140525X21001990`

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience*, *17*(7), 450–461. `https://doi.org/10.1038/nrn.2016.44`

Tononi, G., & Cirelli, C. (2014). Sleep and the Price of Plasticity: From Synaptic and Cellular Homeostasis to Memory Consolidation and Integration. *Neuron*, *81*(1), 12–34. `https://doi.org/10.1016/j.neuron.2013.12.025`

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

Tononi, G., & Koch, C. (2015). Consciousness: Here, there and everywhere? *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *370*(1668), 20140167–. `https://doi.org/10.1098/rstb.2014.0167`

Tononi, G., & Massimini, M. (2008). Why does consciousness fade in early sleep? *Annals of the New York Academy of Sciences*, *1129*, 330–334. `https://doi.org/10.1196/annals.1417.024`

Tononi, G., McIntosh, A. R., Russell, D. P., & Edelman, G. M. (1998). Functional Clustering: Identifying Strongly Interactive Brain Regions in Neuroimaging Data. *NeuroImage*, *7*(2), 133–149. `https://doi.org/10.1006/nimg.1997.0313`

Tononi, G., & Sporns, O. (2003). Measuring information integration. *BMC Neuroscience*, *4*(1), 31. `https://doi.org/10.1186/1471-2202-4-31`

Tsuchiya, N., Andrillon, T., & Haun, A. (2020). A reply to “the unfolding argument”: Beyond functionalism/behaviorism and towards a science of causal structure theories of consciousness. *Consciousness and Cognition*, *79*, 102877. `https://doi.org/10.1016/j.concog.2020.102877`

Tsunada, J., Liu, A. S. K., Gold, J. I., & Cohen, Y. E. (2016). Causal contribution of primate auditory cortex to auditory perceptual decision-making. *Nature Neuroscience*, *19*(1), 135–142. `https://doi.org/10.1038/nn.4195`

Vezoli, J., Magrou, L., Goebel, R., Wang, X.-J., Knoblauch, K., Vinck, M., & Kennedy, H. (2021). Cortical hierarchy, dual counterstream architecture and the importance of top-down generative networks. *NeuroImage*, *225*, 117479. https://doi.org/10.1016/j.neuroimage.2020.117479

Vinck, M., Batista-Brito, R., Knoblich, U., & Cardin, J. A. (2015). Arousal and Locomotion Make Distinct Contributions to Cortical Activity Patterns and Visual Encoding. *Neuron*, *86*(3), 740–754. https://doi.org/10.1016/j.neuron.2015.03.028

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Walker, E. Y., Sinz, F. H., Cobos, E., Muhammad, T., Froudarakis, E., Fahey, P. G., Ecker, A. S., Reimer, J., Pitkow, X., & Tolias, A. S. (2019). Inception loops discover what excites neurons most using deep predictive models. *Nature Neuroscience*, *22*(12), 2060–2065. https://doi.org/10.1038/s41593-019-0517-x

Walt, S. van der, Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science & Engineering*, *13*(2), 22–30. https://doi.org/10.1109/MCSE.2011.37

Wamsley, E., Donjacour, C. E. H. M., Scammell, T. E., Lammers, G. J., & Stickgold, R. (2014). Delusional confusion of dreaming and reality in narcolepsy. *Sleep*, *37*(2), 419–422. https://doi.org/10.5665/sleep.3428

Wang, Q., Ding, S.-L., Li, Y., Royall, J., Feng, D., Lesnar, P., Graddis, N., Naeemi, M., Facer, B., Ho, A., Dolbeare, T., Blanchard, B., Dee, N., Wakeman, W., Hirokawa, K. E., Szafer, A., Sunkin, S. M., Oh, S. W., Bernard, A., … Ng, L. (2020). The Allen Mouse Brain Common Coordinate Framework: A 3D Reference Atlas. *Cell*, *181*(4), 936–953.e20. https://doi.org/10.1016/j.cell.2020.04.007

Wang, Q., Sporns, O., & Burkhalter, A. (2012). Network Analysis of Corticocortical Connections Reveals Ventral and Dorsal Processing Streams in Mouse Visual Cortex. *Journal of Neuroscience*, *32*(13), 4386–4399. https://doi.org/10.1523/JNEUROSCI.6063-11.2012

Waskom, M., & the seaborn development team. (2020, September). *Mwaskom/seaborn* (Version latest). https://doi.org/10.5281/zenodo.592845

Watakabe, A., Skibbe, H., Nakae, K., Abe, H., Ichinohe, N., Rachmadi, M. F., Wang, J., Takaji, M., Mizukami, H., Woodward, A., Gong, R., Hata, J., Van Essen, D. C., Okano, H., Ishii, S., & Yamamori, T. (2023). Local and long-distance organization of prefrontal cortex circuits in the marmoset brain. *Neuron*, *111*(14), 2258–2273.e10. https://doi.org/10.1016/j.neuron.2023.04.028

Wei, Z., Lin, B.-J., Chen, T.-W., Daie, K., Svoboda, K., & Druckmann, S. (2020). A comparison of neuronal population dynamics measured with calcium imaging and electrophysiology. *PLOS Computational Biology*, *16*(9), e1008198. https://doi.org/10.1371/journal.pcbi.1008198

Wenzel, M., Han, S., Smith, E. H., Hoel, E., Greger, B., House, P. A., & Yuste, R. (2019). Reduced Repertoire of Cortical Microstates and Neuronal Ensembles in Medically Induced Loss of Consciousness. *Cell Systems*, *8*(5), 467–474. `https://doi.org/10.1016/j.cels.2019.03.007`

Yang, C.-M., & Wu, C.-S. (2007). The effects of sleep stages and time of night on NREM sleep ERPs. *International journal of psychophysiology: official journal of the International Organization of Psychophysiology*, *63*(1), 87–97. `https://doi.org/10.1016/j.ijpsycho.2006.08.006`

Yu, F., Jiang, Q.-j., Sun, X.-y., & Zhang, R.-w. (2015). A new case of complete primary cerebellar agenesis: Clinical and imaging findings in a living patient. *Brain*, *138*(6), e353–e353. `https://doi.org/10.1093/brain/awu239`

Yuste, R., Goering, S., Arcas, B. A. y, Bi, G., Carmena, J. M., Carter, A., Fins, J. J., Friesen, P., Gallant, J., Huggins, J. E., Illes, J., Kellmeyer, P., Klein, E., Marblestone, A., Mitchell, C., Parens, E., Pham, M., Rubel, A., Sadato, N., … Wolpaw, J. (2017). Four ethical priorities for neurotechnologies and AI. *Nature*, *551*(7679), 159–163. `https://doi.org/10.1038/551159a`

Zaeemzadeh, A. (in preparation). *Shannon information and integrated information*.

Zaeemzadeh, A., & Tononi, G. (2023, May). *Upper bounds for integrated information*.

Zanardi, P., Tomka, M., & Venuti, L. C. (2018, June 4). *Towards Quantum Integrated Information Theory*.

Zatka-Haas, P., Steinmetz, N. A., Carandini, M., & Harris, K. D. (2021). Sensory coding and the causal impact of mouse cortex in a visual decision. *eLife*, *10*, e63163. `https://doi.org/10.7554/eLife.63163`

Zerlaut, Y., Zucca, S., Panzeri, S., & Fellin, T. (2019). The Spectrum of Asynchronous Dynamics in Spiking Networks as a Model for the Diversity of Non-rhythmic Waking States in the Neocortex. *Cell Reports*, *27*(4), 1119–1132.e7. `https://doi.org/10.1016/j.celrep.2019.03.102`

Zilles, K., & Palomero-Gallagher, N. (2017). Multiple transmitter receptors in regions and layers of the human cerebral cortex. *Frontiers in neuroanatomy*, *11*, 78. `https://doi.org/10.3389/fnana.2017.00078`

A Chapter 2 supporting information

A.1 Calculating Φ

Available at: `https://doi.org/10.1371/journal.pcbi.1006343.s001`

A.2 PyPhi v1.1.0 source code

Available at: `https://doi.org/10.1371/journal.pcbi.1006343.s006`
Note that installing PyPhi via '`pip`' or downloading the source code from GitHub (`https://github.com/wmayner/pyphi`) is recommended in order to obtain the most up-to-date version of the software.

A.3 PyPhi v1.1.0 documentation.

Available at: `https://doi.org/10.1371/journal.pcbi.1006343.s007`
Note that accessing the documentation online at `https://pyphi.readthedocs.io` is recommended, as it is updated for each new version of the software.

A.4 Memoization and caching optimizations

During the course of computing a `SystemIrreducibilityAnalysis`, several functions in PyPhi are called multiple times with the same input. For example, calculating `cause_repertoire((A,B), (A,B,C)` and `cause_repertoire((A,C), (A,B,C))` both require calculating `cause_repertoire((A,), (A,B,C))`. Similarly, `cause_repertoire((A,), (B,C))` is both the unpartitioned repertoire of the candidate MIP of mechanism A over purview BC and the first term in the expression for the partitioned repertoire of the candidate MIP of mechanism AB over purview ABC under the partition

$$\frac{A}{BC} \times \frac{B}{A}.$$

In such situations, a natural optimization technique to reduce expensive re-computation of these functions is *memoization*: when a function is computed for a given input, the input-output pair is stored in a lookup table; if the function is called again with that input, the output is simply looked up in the table and returned, without computing the function again.

In PyPhi, memoization is applied to various functions at different levels of the algorithm, listed here:

- `Subsystem._single_node_cause_repertoire()` and `Subsystem._single_node_effect_repertoire()`, the un-normalized cause repertoire of a single-node mechanism and the effect repertoire over a single-node purview, respectively (note that these functions are meant to be called internally by other PyPhi functions and not by the user, as indicated by the leading underscore);

- `Subsystem.cause_repertoire()` and `Subsystem.effect_repertoire()`, the cause and effect repertoires of arbitrary mechanism-purview pairs;

- `Subsystem.mic()` and `Subsystem.mie()`, the MIC and MIE of a mechanism;

- `Network.potential_purviews()`, the purviews which are not necessarily reducible based on the CM;
- Various utility functions such as `pyphi.distribution.max_entropy_distribution()`; and
- `pyphi.compute.sia()`, the full IIT analysis of a `Subsystem`.

At the highest level, `pyphi.compute.sia()` is memoized such that the `SystemIrreducibilityAnalysis` object is stored persistently on the filesystem, rather than in memory, in a directory named `__pyphi_cache__` (which is automatically created in the directory where the Python session was started). This means that the `SystemIrreducibilityAnalysis` objects are automatically saved across Python sessions and can be quickly retrieved simply by running the same code, which is useful for interactive, exploratory work. This behavior can be controlled with the `pyphi.config.CACHE_SIAS` configuration setting. Note, however, that this feature is primarily for convenience and is not intended to replace explicit data management. Additionally, care must be taken to erase or disable the cache when upgrading to new versions of PyPhi, as changes to the algorithm may invalidate previously computed output. A final caveat: because the results are stored on the filesystem, they can accumulate and occupy a large amount of disk space if the `__pyphi_cache__` directory is not periodically removed.

A.5 Proof of the strong connectivity optimization

Theorem A.1 (strong connectivity). *If $G = (V, E)$ is a directed graph that is not strongly connected, then $\Phi(G) = 0$.*

Proof. Since G is not strongly connected, G contains $n \geq 2$ strongly connected components, which we arbitrarily label

$$G_1, G_2, \ldots, G_n \;=\; (V_1, E_1), (V_2, E_2), \ldots, (V_n, E_n).$$

Let $E_{j,k}$ denote the set of directed edges from nodes in component G_j to those in component G_k, $\{(a, b) \in E \mid a \in V_j \text{ and } b \in V_k\}$.

Consider the first component G_1. For every other component $G_i,\ i \neq 1$, either $E_{1,i} = \varnothing$ or $E_{i,1} = \varnothing$, because otherwise G_1 and G_i would not be distinct connected components. Now let $\overline{G_1}$ be the indices of components that receive no edges from G_1, $\{i \in 1, \ldots, n \mid E_{1,i} = \varnothing\}$. Then let Y be the union of the nodes in these components,

$$Y = \bigcup_{i \,\in\, \overline{G_1}} V_i \,,$$

and let $X = V \setminus Y$. Then X and Y form a partition of V such that there are no edges from any nodes in X to any nodes in Y.

Now consider the system cut $c(X, Y)$ that cuts edges from nodes in X to nodes in Y. Because there are no such edges, none of the node TPMs are changed after applying the cut, and thus the subsystem TPM is unchanged because it is the product of the node TPMs. Since the cause-effect structure of a system is a function of the subsystem's TPM, the cause-effect structure $C_{c(X,Y)}$ of the partitioned subsystem and the cause-effect structure C of the unpartitioned subsystem are identical.

Let $\mathit{\Phi}_{c(X,Y)}(G)$ be the $\mathit{\Phi}$ value of G with respect to $c(X, Y)$. By definition, this is the `ces_distance` between C and $C_{c(X,Y)}$. The `ces_distance` function is a metric, so since $C_{c(X,Y)} = C$ we have that

$$\texttt{ces\_distance}(C, C_{c(X,Y)}) \;=\; 0$$

by non-negativity of metrics, and thus $\Phi_{c(X,Y)}(G) = 0$. Now, by definition,

$$\Phi(G) \;=\; \min_{c\,\in\,\mathbb{C}} \;\Phi_c(G)$$

where $\mathbb{C}$ is the set of all system cuts. Since $\Phi_{c(X,Y)}(G) = 0$, by non-negativity we have $\Phi(G) = 0$. □

A.6 Proof of the block-factorable optimization.

Definition A.2 (block diagonal). *An $n \times m$ matrix $\mathbf{A}_{n,m}$ is said to be* block diagonal *if it can be written as*

$$\mathbf{A}_{n,m} \;=\; \begin{bmatrix} \mathbf{B}_{s,t} & \mathbf{0}_{s,m-t} \\ \mathbf{0}_{n-s,t} & \mathbf{C}_{n-s,m-t} \end{bmatrix},$$

where $1 \leq s < n$ and $1 \leq t < m$.

We consider sub-matrices of the connectivity matrix CM of the form $\mathrm{CM}(\pi_{\mathrm{row}}, \pi_{\mathrm{col}})$, where

$$[\mathrm{CM}(\pi_{\mathrm{row}}, \pi_{\mathrm{col}})]_{i,j} \;=\; [\mathrm{CM}]_{\pi_{\mathrm{row}}(i), \pi_{\mathrm{col}}(j)}.$$

Definition A.3 (block factorable). *A mechanism-purview pair (M, P) is said to be* block factorable *if there exists a permutation π_M of the mechanism indices and a permutation π_P of the purview indices such that $CM(\pi_M, \pi_P)$ is block diagonal (for effect purviews) or $CM(\pi_P, \pi_M)$ is block diagonal (for cause purviews).*

Theorem A.4 (block reducibility). *If a mechanism-purview pair is block factorable, then it is reducible $(\varphi = 0)$.*

Proof. Consider a mechanism M constituted of n elements and a purview P constituted of m elements. Assume without loss of generality that P is an effect purview. Since (M, P) is block factorable, there exist permutations π_M and π_P such that $\mathrm{CM}(\pi_M, \pi_P)$ is block diagonal, *i.e.*,

$$\mathrm{CM}(\pi_M, \pi_P) \;=\; \begin{bmatrix} \mathbf{B}_{s,t} & \mathbf{0}_{s,m-t} \\ \mathbf{0}_{n-s,t} & \mathbf{C}_{n-s,m-t} \end{bmatrix},$$

where $1 \leq s < n$ and $1 \leq t < m$.

We define a mechanism-purview partition

$$c \;:=\; \frac{M_1}{P_1} \times \frac{M_2}{P_2}$$

that cuts edges from M_1 to P_2 and from M_2 to P_1, where

$$\begin{aligned} M_1 &= \{\, \pi_M(i) \mid \quad\; 1 \leq i < \; s+1 \,\} \\ M_2 &= \{\, \pi_M(i) \mid s+1 \leq i < \; n+1 \,\} \\ P_1 &= \{\; \pi_P(i) \mid \quad\; 1 \leq i < \; t+1 \,\} \\ P_2 &= \{\; \pi_P(i) \mid t+1 \leq i < m+1 \,\} \end{aligned}$$

Note that $[\mathrm{CM}(\pi_M, \pi_P)]_{i,j} \;=\; 0$ if either $i \in M_1$ and $j \in P_2$ or $i \in M_2$ and $j \in P_1$. Thus there are no edges cut by c, and it leaves the subsystem's TPM unchanged. Since the effect repertoire of a mechanism-purview combination is a function of the subsystem's TPM, the unpartitioned effect repertoire, $\mathrm{ER}(M, P)$ and the partitioned repertoire $\mathrm{ER}_c(M, P)$ are identical.

By definition, $\varphi_c(M,P)$ is the distance between $\mathrm{ER}(M,P)$ and $\mathrm{ER}_c(M,P)$, so $\varphi_c(M,P) = 0$. Now, by definition,

$$\varphi(M,P) = \min_{p \,\in\, \mathbb{P}} \varphi_p(M,P),$$

where $\mathbb{P}$ is the set of all partitions of (M,P). Since $\varphi_c(M,P) = 0$, by the non-negativity of metrics we have $\varphi(M,P) = 0$.

□

A.7 Proof of an analytical solution to the EMD between effect repertoires

Theorem A.5 (analytical EMD). *Consider two random variables X_1, X_2 with corresponding state spaces Ω_1, Ω_2 and an 'additive' metric D,*

$$D((i_1,i_2),\,(j_1,j_2)) \;=\; D(i_1,j_1) + D(i_2,j_2) \quad \forall\,(i_1,i_2) \text{ and } (j_1,j_2) \in \Omega_1 \times \Omega_2.$$

Let p_1 and q_1 be two probability distributions on X_1, and let p_2 and q_2 be probability distributions on X_2. If X_1 and X_2 are independent, then the EMD between the joint distributions $p = p_1p_2$ and $q = q_1q_2$, with D as the ground metric, is equal to the sum of the EMDs between the marginal distributions:

$$EMD(p,q) \;=\; EMD(p_1,q_1) + EMD(p_2,q_2).$$

Proof. First, we demonstrate that

$$\mathrm{EMD}(p,q) \;\leq\; \mathrm{EMD}(p_1,q_1) + \mathrm{EMD}(p_2,q_2).$$

To do this, we define a third probability distribution as an intermediate point,

$$r \;:=\; q_1p_2.$$

We define the following flow from p to r,

$$f_{p,r}(i_1,i_2,j_1,j_2) \;:=\; \begin{cases} p_2(i_2) f^*_{p_1,q_1}(i_1,j_1) & \text{if } i_2 = j_2 \\ 0 & \text{otherwise,} \end{cases}$$

where $f^*_{p_1,q_1}$ is the optimal flow for the EMD between p_1 and q_1. With this flow, we have

$$\begin{aligned} \mathrm{EMD}(p,r) &\leq \sum_{i_1,i_2,j_1,j_2} f_{p,r}(i_1,i_2,j_1,j_2) D((i_1,i_2),\,(j_1,j_2)) \\ &= \sum_{i_1,i_2,j_1} p_2(i_2) f^*_{p_1,q_1}(i_1,j_1) D(i_1,j_1) \\ &= \sum_{i_2} p_2(i_2) \sum_{i_1,j_1} f^*_{p_1,q_1}(i_1,j_1) D(i_1,j_1) \\ &= \mathrm{EMD}(p_1,q_1) \end{aligned}$$

We next define a flow from r to q,

$$f_{r,q}(i_1,i_2,j_1,j_2) \;:=\; \begin{cases} q_1(i_1) f^*_{p_2,q_2}(i_2,j_2) & \text{if } i_1 = j_1 \\ 0 & \text{otherwise,} \end{cases}$$

where $f^*_{p_2,q_2}$ is the optimal flow for the EMD between p_2 and q_2. With this flow, we have

$$\begin{aligned}
\mathrm{EMD}(r,q) &\le \sum_{i_1,i_2,j_1,j_2} f_{r,q}(i_1,i_2,j_1,j_2)D((i_1,i_2),\,(j_1,j_2)) \\
&= \sum_{i_1,i_2,j_2} q_1(i_1)f^*_{p_2,q_2}(i_2,j_2)D(i_2,j_2) \\
&= \sum_{i_1} q_1(i_1)\sum_{i_2,j_2} f^*_{p_2,q_2}(i_2,j_2)D(i_2,j_2) \\
&= \mathrm{EMD}(p_2,q_2)
\end{aligned}$$

Then using the triangle inequality (the EMD is a metric), we have

$$\mathrm{EMD}(p,q) \;\le\; \mathrm{EMD}(p,r)+\mathrm{EMD}(r,q) \;\le\; \mathrm{EMD}(p_1,q_1)+\mathrm{EMD}(p_2,q_2).$$

To complete the proof, we next demonstrate that

$$\mathrm{EMD}(p,q) \;\ge\; \mathrm{EMD}(p_1,q_1)+\mathrm{EMD}(p_2,q_2).$$

If $f^*_{p,q}$ is the optimal flow for $\mathrm{EMD}(p,q)$, then define a flow between p_1 and q_1,

$$f_1(i_1,j_1) \;:=\; \sum_{i_2,j_2} f^*_{p,q}(i_1,i_2,j_1,j_2),$$

and a flow between p_2 and q_2

$$f_2(i_2,j_2) \;:=\; \sum_{i_1,j_1} f^*_{p,q}(i_1,i_2,j_1,j_2).$$

Then using the additive property of the ground metric D,

$$\begin{aligned}
\mathrm{EMD}(p,q) &= \sum_{i_1,i_2,j_1,j_2} f^*_{p,q}(i_1,i_2,j_1,j_2)D((i_1,i_2),\,(j_1,j_2)) \\
&= \sum_{i_1,i_2,j_1,j_2} f^*_{p,q}(i_1,i_2,j_1,j_2)D(i_1,j_1)+\sum_{i_1,i_2,j_1,j_2} f^*_{p,q}(i_1,i_2,j_1,j_2)D(i_2,j_2) \\
&= \sum_{i_1,j_1}\left(\sum_{i_2,j_2} f^*_{p,q}(i_1,i_2,j_1,j_2)\right)D(i_1,j_1) \\
&\qquad +\sum_{i_2,j_2}\left(\sum_{i_1,j_1} f^*_{p,q}(i_1,i_2,j_1,j_2)\right)D(i_2,j_2) \\
&= \sum_{i_1,j_1} f_1(i_1,j_1)D(i_1,j_1)+\sum_{i_2,j_2} f_2(i_2,j_2)D(i_2,j_2) \\
&\ge \mathrm{EMD}(p_1,q_1)+\mathrm{EMD}(p_2,q_2).
\end{aligned}$$

Therefore $\mathrm{EMD}(p,q)=\mathrm{EMD}(p_1,q_1)+\mathrm{EMD}(p_2,q_2)$.

□

Perhaps it is worth demonstrating that the flows f_1, f_2, $f_{p,r}$ and $f_{r,q}$ satisfy the EMD requirements. Consider $f_{p,r}$,

$$f_{p,r}(i_1,i_2,j_1,j_2) \;=\; \begin{cases} p_2(i_2)f^*_{p_1,q_1}(i_1,j_1) & \text{if } i_2=j_2 \\ 0 & \text{otherwise,} \end{cases}$$

where $f^*_{p_1,q_1}$ is the optimal flow for the EMD between p_1 and q_1.

Since $p_2(i_2) \geq 0$ (probability) and $f^*_{p_1,q_1}(i_1, j_1) \geq 0$ (definition of optimal flow),

$$f_{p,r}(i_1, i_2, j_1, j_2) \geq 0.$$

Next,

$$\begin{aligned} \sum_{j_1,j_2} f_{p,r}(i_1, i_2, j_1, j_2) &= \sum_{j_1} p_2(i_2) f^*_{p_1,q_1}(i_1, j_1) \\ &= q_1(i_1)p_2(i_2) \\ &= r(i_1, i_2), \end{aligned}$$

and

$$\begin{aligned} \sum_{i_1,i_2} f_{p,r}(i_1, i_2, j_1, j_2) &= \sum_{i_1} p_2(j_2) f^*_{p_1,q_1}(i_1, j_1) \\ &= q_1(j_1)p_2(i_2) \\ &= r(j_1, j_2). \end{aligned}$$

Finally,

$$\begin{aligned} \sum_{i_1,i_2,j_1,j_2} f_{p,r}(i_1, i_2, j_1, j_2) &= \sum_{i_2} p_2(i_2) \sum_{i_1,j_1} f^*_{p_1,q_1}(i_1, j_1) \\ &= \sum_{i_1,j_1} f^*_{p_1,q_1}(i_1, j_1) \\ &= 1 \end{aligned}$$

Thus $f_{p,r}$ satisfies the criteria for a potential flow. The others follow similarly.

A.8 Time complexity

Here I analyze the time complexity of the IIT 3.0 algorithm for finding the major complex in a substrate of n binary elements. (Note that this does not apply to the IIT 4.0 algorithm, because it uses different partitions, a different distance metric, includes explicit calculation of relations, etc.)

The time complexity of computing the EMD on the cause side is $O(N^3 \log N)$, where N is the number of bins of "earth" (Pele & Werman, 2009; Shirdhonkar & Jacobs, 2008). For a given mechanism M with $|M| = m$ and purview P with $|P| = p$, the repertoire over the purview is an array of size 2^p. Evaluating a mechanism-purview partition using the EMD is thus $O(p2^{p^3})$ for the cause side. On the effect side, Theorem A.5 reduces this to $O(p)$.

We can obtain the desired expression by counting the number of EMD calculations[3] as

$$O\left(\underbrace{\sum_{s=0}^{n} \binom{n}{s}}_{\substack{\text{Candidate} \\ \text{systems}}} \underbrace{\sum_{\psi_s=0}^{s} \binom{s}{\psi_s}}_{\substack{\text{System} \\ \text{bipartitions}}} \left(\left[\underbrace{\sum_{m=0}^{s} \binom{s}{m}}_{\text{Mechanisms}} \underbrace{\sum_{\psi_m=0}^{m} \binom{m}{\psi_m}}_{\substack{\text{Mechanism} \\ \text{bipartitions}}} \right] \left[\underbrace{\sum_{p=0}^{s} \binom{s}{p}}_{\text{Purviews}} \underbrace{2 \sum_{\psi_p=0}^{p} \binom{p}{\psi_p}}_{\substack{\text{Purview} \\ \text{bipartitions} \\ \text{(directional)}}} \underbrace{O(p2^{p^3})}_{\substack{\text{Cause} \\ \text{EMD}}} \underbrace{O(p)}_{\substack{\text{Effect} \\ \text{EMD}}} \right] \right) \right),$$

which simplifies to $O(192\, n\, 103^{(n-1)}) = O(n103^n)$.

[3] I use 0 as the lower bound throughout to yield a simpler expression. Using 1 (reflecting that systems, mechanisms, etc. cannot be empty) changes the expression by only a constant factor.

Comparison to the analysis in Hanson and Walker, 2023

Hanson and Walker (2023) analyze the time complexity as $O(13^n)$ and note the discrepancy with Mayner et al. (2018), suggesting that perhaps my analysis "resolves the elementary computation in terms of some more fundamental operation". In some sense, this is indeed the reason for the discrepancy. Their expression is given in terms of "elementary distance calculations," *i.e.*, it assumes the EMD computation is $O(1)$. However, the distance calculations cannot be considered "elementary," because the EMD is between two repertoires over purviews, and purview sizes grow with the size of the system n. This means that the repertoires are of size $O(2^n)$, not $O(1)$, so their assumption leads to an incorrect analysis.

B Chapter 3 supporting information

B.1 Resolving ties in the IIT algorithm

The information postulate requires that a system's cause-effect power is specific: the system in its current state must select a specific cause-effect state for its units. Likewise, mechanisms within the system must select a specific cause and effect state over their purviews. The exclusion postulate requires that a complex must be constituted of a definite set of units, and that mechanisms within the complex specify a definite cause and effect. By the principle of maximal existence, cause-effect states, complexes, and the cause-effect states of a mechanism within the system are identified as those with maximal cause-effect power.

However, for systems with built-in symmetries in their architecture or in the input–output functions of their units, multiple sets of units or cause-effect states may "tie" for maximal intrinsic information or integrated information (Barbosa et al., 2021; Hanson & Walker, 2023; S. Krohn & Ostwald, 2017; Moon, 2019). While such ties are frequently encountered in small, deterministic toy models, they are unlikely to occur in realistic systems. In fact, IIT requires a degree of indeterminism at the micro-level from first principles (Tononi, forthcoming). Ties are a consequence of symmetries in the substrate TPMs, which also become increasingly unlikely for larger, more realistic systems and are thus unlikely to play any role in the empirical application and evaluation of IIT. Nevertheless, such ties should be resolved in line with IIT's postulates and principles, as we outline in the following for algorithmic purposes. In general, ties that occur at an intermediate step in the algorithm are resolved based on the principle of maximal existence by considering the subsequent postulates (essential requirements for existence) in order.

Maximal substrates can be identified using an iterative algorithm (3.26). A maximal substrate excludes all overlapping systems with lower φ_s from existing as complexes. If overlapping systems tie for $\max_{S \subseteq U_k} \varphi_s(\mathcal{T}_e, \mathcal{T}_c, s)$, we apply the maximum existence principle taking their respective Φ values (composition) into consideration and choose the system with maximal structure integrated information Φ as the complex. In the rare case that two or more such systems also tie in Φ, these systems do not comply with the exclusion postulate. For this reason, they do not qualify as complexes and we choose the next best system (based on φ_s) that is unique (Marshall et al., 2023; see also Moon, 2019 for a similar line of reasoning).

For a system, the maximal cause-effect state $s' = \{s'_c, s'_e\}$ is the one that maximizes the system's intrinsic cause and effect information. However, if multiple states comply with equation (3.13), we select the one for which the system specifies the maximal integrated information $\varphi_s(\mathcal{T}_e, \mathcal{T}_c, s, \theta')$ (3.22) over its minimum partition θ'. This is again in line with the principle of maximal existence: being tied for intrinsic information, it is the integrated information φ_s that determines in which cause-effect state the system "exists the most."

Remaining ties in which multiple cause-effect states specify the same φ_s rarely matter for system selection, but need to be resolved in order to determine the system's cause-effect structure. By the maximum existence principle, we choose the cause-effect state that maximizes the system's structure integrated information Φ. As above, in the rare case that two or more states also tie in Φ, the system does not comply with the information postulate and thus does not qualify as a complex (unless the cause-effect structures are actually identical from the intrinsic perspective, in which case the tie would be extrinsic and not a violation of the information postulate).

The cause or effect state of a mechanism within the system for a candidate purview is first selected based on its intrinsic information $\text{ii}(m, z)$ (3.37). Next, we compare the integrated information $\varphi(m, Z)$ (3.43) of all maximal cause or effect states across all possible purviews (including all possible ties in $\text{ii}(m, z)$ within a candidate purview) to identify the maximally irreducible cause or effect $z^*_{c/e}$ of the mechanism within the system (3.46). By the maximum existence postulate, potential ties in max $\varphi_d(m, Z)$ and thus in the cause-effect state $z^*_{c/e}$ of a distinction may be resolved at the level of the cause-effect structure, by selecting the $z^*_{c/e}$ that maximizes the system's structure integrated information Φ. Accordingly, in case of state ties within the same purview, we select the state that is congruent with the system's cause-effect state s'. In case of ties across different purviews, the maximal cause-effect state will generally correspond to the one that supports the most relations with other distinctions, which typically favors larger purviews.

B.2 Comparison to IIT 1.0–3.0 and subsequent publications

As highlighted in the main text, IIT is a work in progress. While the core theory has remained the same, its formal framework has been progressively refined and extended (Balduzzi & Tononi, 2008; Oizumi et al., 2014; Tononi, 2004; Tononi & Sporns, 2003). Compared to prior versions (IIT 1.0, Tononi, 2004; Tononi and Sporns, 2003; IIT 2.0, Balduzzi and Tononi, 2008, 2009; Tononi, 2008; and IIT 3.0, Kleiner and Tull, 2021; Mayner et al., 2018; Oizumi et al., 2014; Tononi, 2012), IIT 4.0 presents a more complete, self-consistent formulation. The most notable advances in IIT 4.0 include the introduction of an Intrinsic Difference (ID) measure (Barbosa et al., 2020, 2021) that is uniquely consistent with IIT's postulates, the explicit assessment of causal relations (Haun & Tononi, 2019), and a more exact translation of the axioms into postulates. Because IIT 3.0 already included a comparison to IIT 1.0 and 2.0 (see Oizumi et al., 2014, Supporting Information Text S1), we mainly focus on subsequent developments.

Axioms and postulates

The starting point of IIT has always been phenomenology, but the axioms and postulates of the theory were first explicitly presented in IIT 3.0 (Oizumi et al., 2014; Tononi et al., 2016). The updated 4.0 exposition of IIT's axioms explicitly separates phenomenal existence, which is not a property, from intrinsicality, which is one of the essential properties of phenomenal existence. Accordingly, the existence of experience is introduced as IIT's foundational, zeroth axiom. The remaining five axioms (intrinsicality, information, integration, exclusion, and composition) capture the essential properties that are immediate and irrefutably true of every conceivable experience.

Compared to IIT 3.0, the formulation of the axioms has been refined to avoid misunderstandings (Bayne, 2018; Merker et al., 2022) and to highlight their immediacy and irrefutability. The formulation of IIT's postulates has been updated accordingly with the objective of tracking the phenomenal axioms as closely as possible. For example, conforming more closely to the information axiom, the information postulate requires that the system must select a specific cause-effect state over its units. The composition axiom now highlights both phenomenal distinctions and their relations. By the composition postulate, phenomenal distinctions and relations are accounted for in physical terms by causal distinctions and relations. Because only an experience that exists intrinsically, in a way that is specific, irreducible, and definite can also be structured, composition takes the final position in the ordering of the axioms and postulates.

Background Conditions

The IIT 4.0 mathematical framework updates the treatment of background conditions (units $W = U \setminus S$). In IIT 3.0 (Oizumi et al., 2014), the background units were fixed (conditioned) in their current state for the evaluation of effects, and fixed (conditioned) in their actual past state for the evaluation of causes. In publications since then, the actual past state of background units was considered to be unavailable from the intrinsic perspective of the system, so instead the background units were fixed (conditioned) in their current state for evaluating both causes and effects (Barbosa et al., 2021; Haun & Tononi, 2019; Hoel et al., 2016; Marshall et al., 2023). However, fixing the background conditions in the current state for evaluating causes leads to situations where the current state is unreachable (no cause). In IIT 4.0, the treatment of background conditions is updated to causally marginalize the background units, conditional on the current state of the universe (see Identifying substrates of consciousness). This formulation avoids the problem of unreachable states, while also only requiring knowledge of the current state of background units (the 'context' for the causal powers analysis).

Identifying maximal substrates

The IIT 4.0 formalism to identify maximal substrates was first described in detail in Marshall et al. (2023). Maximal substrates, or complexes, are identified based on their system integrated information φ_s (as in IIT 2.0 but unlike IIT 3.0, which evaluated integration after composition). System partitions remain directional (as in IIT 3.0). In IIT 4.0, the minimum partition (MIP) is identified as the partition with minimal integrated information (φ_s), normalized by the maximal possible value of φ_s across this partition for an arbitrary TPM of the same dimensions (3.24). In this way, the MIP is sensitive to the fault lines of a candidate system, rather than defaulting to partitions of individual system units. The IIT 4.0 analysis is state-dependent (as in IIT 2.0 and 3.0) and requires positive cause and effect power for a system to exist (as in IIT 1.0 and 3.0).

Measuring intrinsic information

Supplanting prior measures such as the Kullback-Leibler divergence (KLD, IIT 2.0 Balduzzi and Tononi, 2008), or the (extended) Earth Mover's Distance (EMD, IIT 3.0, Oizumi et al., 2014), the IIT 4.0 formalism features a newly developed Intrinsic Difference (ID) measure (Barbosa et al., 2020), which uniquely complies with the postulates of IIT. Formally:

Theorem B.1. *Let $(p, q) \in \mathcal{T}_U$ be two probability distributions, and $D\colon (\mathcal{T}_U, \mathcal{T}_U) \to \mathbb{R}$, where D satisfies Properties I, II and III defined in Barbosa et al. (2020). Then*

$$D(p,q) = \max_{u \in \Omega_U} \{f(p(u), q(u))\},$$

where

$$f(p(u), q(u)) = k\, p(u) \log\left(\frac{p(u)}{q(u)}\right), \quad k \in \mathbb{R}^+. \tag{B.1}$$

The proof of the Theorem can be found in Barbosa et al. (2020). Note that ID is related to the KLD, which can be viewed as an average of the point-wise mutual information across states and is an additive measure. By contrast, the ID is defined based on the state that maximizes the difference between distributions (specificity property). Accordingly, the intrinsic information specified by a system (or mechanism) over a cause or effect state is evaluated as a product of informativeness and selectivity, which makes it subadditive if $p < 1$, that is, if cause-effect power is spread over more than one state. As intrinsic information is evaluated over specific cause or effect states, its maximal value over a state distribution identifies the specific cause or effect state selected by the system (or mechanism), in line with the information postulate. The intrinsic effect information ii_e is equivalent to the ID between the constrained and unconstrained effect probability distributions, but the intrinsic cause information ii_c is not because of the use of backwards cause probabilities for selectivity (see main text).

Causal distinctions

Distinctions capture how the various system subsets specify system subsets as their cause and effect within the system. In IIT 3.0, distinctions were called "concepts" (which composed to a "conceptual" structure), a term that could generate unnecessary misunderstandings (Haun & Tononi, 2019). An updated formalism to identify causal distinctions was first presented in Barbosa et al. (2021). As in IIT 3.0, causes and effects must be specified in a way that is irreducible ($\varphi_d > 0$). Unlike IIT 3.0, in IIT 4.0 distinctions select a specific state as a cause and an effect. The definition of cause and effect probabilities $\pi(z \mid m)$ ((3.31), (3.34) and (3.30)) remains unchanged, but is now presented more formally in terms of product probabilities, rather than referring to "virtual elements"; Albantakis et al., 2019; S. Krohn and Ostwald, 2017). In IIT 4.0, a mechanism selects a specific cause and effect state based on the Intrinsic Difference (ID) measure introduced in Barbosa et al. (2020) (see above). The set of permissible partitions $\theta \in \Theta(M, Z)$ (3.39) has also been updated to ensure that partitions are always "disintegrating" the mechanism (Albantakis et al., 2019; Barbosa et al., 2021).

The present formulation includes several updates compared to Barbosa et al. (2021). First, the cause-effect state z' is selected based on the intrinsic information $\text{ii}_e(m, Z)$ (3.37), before evaluating its integrated information $\varphi(m, Z)$ (3.43) for z'. This is because the cause and effect of m (its cause-effect state) should be determined by the mechanism as a whole, independent of how it can be partitioned. Second, we evaluate intrinsic information without the absolute value, as in Barbosa et al. (2020), because, to comply with the existence postulate, a mechanism's cause state should be one that would increase the probability of its current state, and its effect state one whose probability would be increased by the mechanism being in its current state. Third, to correctly capture this increase in probability on the cause side, the informativeness term is expressed in terms of forward probabilities (as opposed to backward probabilities employing Bayes rule) also on the cause side for ii_c and φ_c, evaluating the increase in probability of the current state due to the cause state. Fourth, we have updated the resolution of ties at the level of distinctions according to the principle of maximal existence (see B.1). Finally, a candidate distinction only contributes to the system's cause-effect structure if its maximal cause-effect state z^* is congruent with the maximal cause-effect state s' of the system.

Causal Relations

Relations bind together a set of causal distinctions over a congruent overlap between their causes and/or effects. Developing an explicit account of phenomenal relations in terms of causal relations was a main goal of IIT since IIT 1.0. IIT 3.0 employed a distance metric—the Earth Mover's Distance—that was sensitive to whether different distinctions ("concepts") had similar cause-effect repertoires, but relations did not figure explicitly in the formalism despite their central role in characterizing experience. An explicit account was first described in Haun and Tononi (2019). The IIT 4.0 formalism further distinguishes between relations (which bind a set of distinctions with overlapping causes and/or effects) and the faces of a relation (which specify the maximal overlap of a set of purviews and jointly characterize the type of the relation). Moreover, the amount of information specified by a distinction over the overlap and the way relation partitions are assessed differs from the original account (Haun & Tononi, 2019). Because distinctions are irreducible components within the cause-effect structure upon which relations are built, a distinction involved in a relation contributes its entire φ_d, weighted by the extent of its joint overlap (3.56). For this reason, we do not recompute the irreducible information of the mechanism m of a distinction $d(m, z^*, \varphi_d)$ over the candidate overlap o. In Haun and Tononi (2019), distinctions contributing to a relation were partitioned by "noising" the interactions among distinction units.

Φ-structures

In IIT 4.0, a system is a substrate of consciousness (a complex) if it corresponds to a maximum of system integrated information φ_s, as determined through information, integration, and exclusion. This is similar to IIT 2.0 (Balduzzi & Tononi, 2008), although in that case only causes (and not effects) were evaluated. The quality of the experience is identical to the Φ-structure of distinctions and relations unfolded from the complex. The quantity of experience corresponds to the Φ value, the sum of the integrated information of the distinctions and relations that compose the Φ-structure. In IIT 3.0, the determination of the complex through information, integration, and exclusion took into account its compositional structure, although without explicit relations. However, the Φ value corresponding to the quantity of consciousness only captured the distinctions affected by the minimum partition, as opposed to all the distinctions (and relations) unfolded from a maximally irreducible substrate. Because Φ is not evaluated with respect to a MIP, there is no normalization involved in determining Φ (in contrast to φ_S). Instead, Φ captures the complete structure integrated information of a complex in the form of a sum over the integrated information φ of all components of the complex (its distinctions and relations), where the sum was chosen as the simplest option that captures all that exists within the complex. As such, Φ in IIT 4.0 is more aligned with the common-sense notion that the quantity of consciousness relates to the "richness" or "vividness" of an experience considering all its contents. However, whether a system exists as one integrated entity is evaluated by its system integrated information φ_s, which is not compositional.

B.3 Analytical solution for $\sum \varphi_r$ and the number of causal relations

Here, we show how the sum of the relation integrated information over all the causal relations ($\sum \varphi_r$) and the number of relations can be computed without assessing the relations individually. We only need the set of causal distinctions:

$$D(\mathcal{T}_e, \mathcal{T}_c, s) = \{d(m) \; : \; m \subseteq s, \; \varphi_d(m) > 0, \; z_c^*(m) \subseteq s_c', \; z_e^*(m) \subseteq s_e'\},$$

where $d(m) = (m, z^*(m), \varphi_d(m))$ and $z^*(m) = \{z_c^*(m), z_e^*(m)\}$.

Analytical computation of $\sum \varphi_r$

Given a subset of distinctions $\boldsymbol{d} \subseteq D(\mathcal{T}_e, \mathcal{T}_c, s)$ with $|\boldsymbol{d}| \geq 2$, any subset $\boldsymbol{z}$ of purviews that contains either the cause, or effect, or both the cause and effect of each distinction $d \in \boldsymbol{d}$ and overlap congruently defines a relation face f with face overlap $o_f^* = \bigcap_{z \in \boldsymbol{z}} z$. The relation overlap is further defined as the union of the face overlaps $\bigcup_{f \in \boldsymbol{f}(\boldsymbol{d})} o_f^*$, where $\boldsymbol{f}(\boldsymbol{d})$ represents the set of all the faces over the distinction set $\boldsymbol{d}$. Here, intersection and union take into account both the units and their states.

First, we can show:

$$\bigcup_{f \in \boldsymbol{f}(\boldsymbol{d})} o_f^* = \bigcap_{d \in \boldsymbol{d}} \left(z_c^*(d) \cup z_e^*(d)\right),$$

by proving any unit n in $\bigcup_{f \in \boldsymbol{f}(\boldsymbol{d})} o_f^*$ is in $\bigcap_{d \in \boldsymbol{d}} \left(z_c^*(d) \cup z_e^*(d)\right)$ and vice versa:

$$\begin{aligned} n \in \bigcup_{f \in \boldsymbol{f}(\boldsymbol{d})} o_f^* &\iff \exists f \in \boldsymbol{f}(\boldsymbol{d}), n \in o_f^* \iff \forall d \in \boldsymbol{d}, n \in z_c^*(d) \text{ or } n \in z_e^*(d) \\ &\iff \forall d \in \boldsymbol{d}, n \in z_c^*(d) \cup z_e^*(d) \iff n \in \bigcap_{d \in \boldsymbol{d}} \left(z_c^*(d) \cup z_e^*(d)\right) \end{aligned}$$

This helps us to rewrite the relation integrated information of a set of distinctions $\boldsymbol{d} \subseteq D(\mathcal{T}_e, \mathcal{T}_c, s)$ with $|\boldsymbol{d}| \geq 2$ as:

$$\left|\bigcap_{d\in\boldsymbol{d}}\left(z_c^*(d)\cup z_e^*(d)\right)\right| \min_{(z_d,\varphi_d)\in\boldsymbol{d}} \frac{\varphi_d}{|z_c^*(d)\cup z_e^*(d)|}.$$

We further define the set of $z_c^*(d)\cup z_e^*(d)$ of all distinctions in D and their corresponding distinction integrated information as:

$$\mathcal{Z}(\mathcal{T}_e,\mathcal{T}_c,s)=\left\{\left(z_c^*(m)\cup z_e^*(m),\varphi(m)\right):(m,z^*(m),\varphi_d(m))\in D(\mathcal{T}_e,\mathcal{T}_c,s)\right\}.$$

Now, given a single node n in a specific state, we can find all the distinctions that contain n in that state in their cause, or effect, or both purviews as:

$$\mathcal{Z}(n)=\{(z,\varphi):(z,\varphi)\in\mathcal{Z}(\mathcal{T}_e,\mathcal{T}_c,s),n\in z\}. \tag{B.2}$$

Any subset of $\mathcal{Z}(n)$ of size 2 or larger defines a relation whose overlap contains at least n. Formally, for $\boldsymbol{r}\subseteq\mathcal{Z}(n)$, $|\boldsymbol{r}|\geq 2$, there exists a relation with relation purview $\bigcap_{(z_d,\varphi_d)\in\boldsymbol{r}} z_d$ and integrated information value of:

$$\left|\bigcap_{(z_d,\varphi_d)\in\boldsymbol{r}} z_d\right| \min_{(z_d,\varphi_d)\in\boldsymbol{r}} \frac{\varphi_d}{|z_d|}.$$

Note that, by definition of $\mathcal{Z}(n)$ and $\mathcal{Z}(\mathcal{T}_e,\mathcal{T}_c,s)$, z_d is the union of cause and effect purviews. Using the definition of $\mathcal{Z}(\mathcal{T}_e,\mathcal{T}_c,s)$ and $\mathcal{Z}(n)$, we can write the sum of the integrated information of relations, except self-relations, as

$$\sum_{\substack{\boldsymbol{r}\subseteq\mathcal{Z}(\mathcal{T}_e,\mathcal{T}_c,s)\\ \boldsymbol{r}\geq 2}}\left|\bigcap_{(z_d,\varphi_d)\in\boldsymbol{r}} z_d\right| \min_{(z_d,\varphi_d)\in\boldsymbol{r}} \frac{\varphi_d}{|z_d|}=\sum_{n\in s_c'\cup s_e'}\sum_{\substack{\boldsymbol{r}\subseteq\mathcal{Z}(n)\\ |\boldsymbol{r}|\geq 2}} \min_{(z_d,\varphi_d)\in\boldsymbol{r}} \frac{\varphi_d}{|z_d|},$$

By factoring the sum over $\boldsymbol{r}\subseteq\mathcal{Z}(\mathcal{T}_e,\mathcal{T}_c,s)$ into two sums over the nodes n and the relations whose purview contains *at least* n, $\boldsymbol{r}\subseteq\mathcal{Z}(n)$, $|\boldsymbol{r}|\geq 2$, we are overcounting each relation by a factor of its joint purview size $\left|\bigcap_{(z_d,\varphi_d)\in\boldsymbol{r}} z_d\right|$. For example, if a set of distinctions make up a relation $\boldsymbol{r}$ over two units n_1 and n_2, they all are members of both $\mathcal{Z}(n_1)$ and $\mathcal{Z}(n_2)$. Therefore, $\boldsymbol{r}\subseteq\mathcal{Z}(n_1)$ and $\boldsymbol{r}\subseteq\mathcal{Z}(n_2)$. This simplifies the summand to just $\min_{(z_d,\varphi_d)\in\boldsymbol{r}}\frac{\varphi_d}{|z_d|}$. To compute the inner sum, we can sort the distinctions in $\mathcal{Z}(n)$ by their $\frac{\varphi}{|z|}$ value in a non-decreasing order, such that $(z_{(1)},\varphi_{(1)})$ has the summary est $\frac{\varphi}{|z|}$ ratio, $(z_{(2)},\varphi_{(2)})$ has the second smallest $\frac{\varphi}{|z|}$ ratio, and so on. Then, we can compute the sum as:

$$\sum_{\substack{\boldsymbol{r}\subseteq\mathcal{Z}(n)\\ |\boldsymbol{r}|\geq 2}} \min_{(z_d,\varphi_d)\in\boldsymbol{r}} \frac{\varphi_d}{|z_d|}=\sum_{j=1}^{|\mathcal{Z}(n)|}\frac{\varphi_{(j)}}{|z_{(j)}|}(2^{|\mathcal{Z}(n)|-j}-1).$$

In words, any subset $\boldsymbol{r}\subseteq\mathcal{Z}(n)$, $|\boldsymbol{r}|\geq 2$, that contains $(z_{(1)},\varphi_{(1)})$ will have $\min_{(z_d,\varphi_d)\in\boldsymbol{r}}\frac{\varphi_d}{|z_d|}=\frac{\varphi_{(1)}}{|z_{(1)}|}$. There are $2^{|\mathcal{Z}(n)|-1}-1$ of such subsets. Similarly, there are $2^{|\mathcal{Z}(n)|-2}-1$ subsets that contain $(z_{(2)},\varphi_{(2)})$, but not $(z_{(1)},\varphi_{(1)})$, etc. This helps us arrive at our final results:

$$\sum_{\substack{\boldsymbol{r}\subseteq\mathcal{Z}(\mathcal{T}_e,\mathcal{T}_c,s)\\ |\boldsymbol{r}|\geq 2}}\left|\bigcap_{(z_d,\varphi_d)\in\boldsymbol{r}} z_d\right| \min_{(z_d,\varphi_d)\in\boldsymbol{r}} \frac{\varphi_d}{|z_d|}=\sum_{n\in s_c'\cup s_e'}\sum_{j=1}^{|\mathcal{Z}(n)|}\frac{\varphi_{(j)}}{|z_{(j)}|}(2^{|\mathcal{Z}(n)|-j}-1).$$

This gives us the sum of the relation integrated information of all the relations, except the self-relations, *i.e.* $|\boldsymbol{r}| = 1$. The self-relations can be assessed individually without combinatorial explosion.

Analytical count of the number of relations

We can also count all the causal relations among all the distinctions in $D(\mathcal{T}_e, \mathcal{T}_c, s)$ by generalizing the definition of $\mathcal{Z}(n)$ in (B.2) to all the subsets $o \subseteq s'_c \cup s'_e$:

$$\mathcal{Z}(o) = \{(z, \varphi) : (z, \varphi) \in \mathcal{Z}(\mathcal{T}_e, \mathcal{T}_c, s), z \supseteq o\}.$$

For each distinction $d \in D(\mathcal{T}_e, \mathcal{T}_e, s)$, there is a corresponding element $\big(\big(z^*_c(d) \cup z^*_e(d), \varphi(d)\big)$ in $\mathcal{Z}(o)$ if $o \subseteq z^*_c(d) \cup z^*_e(d)$. Any subset of $\mathcal{Z}(o)$ of size 2 or larger defines a relation whose overlap contains at least o. The number of such subsets is:

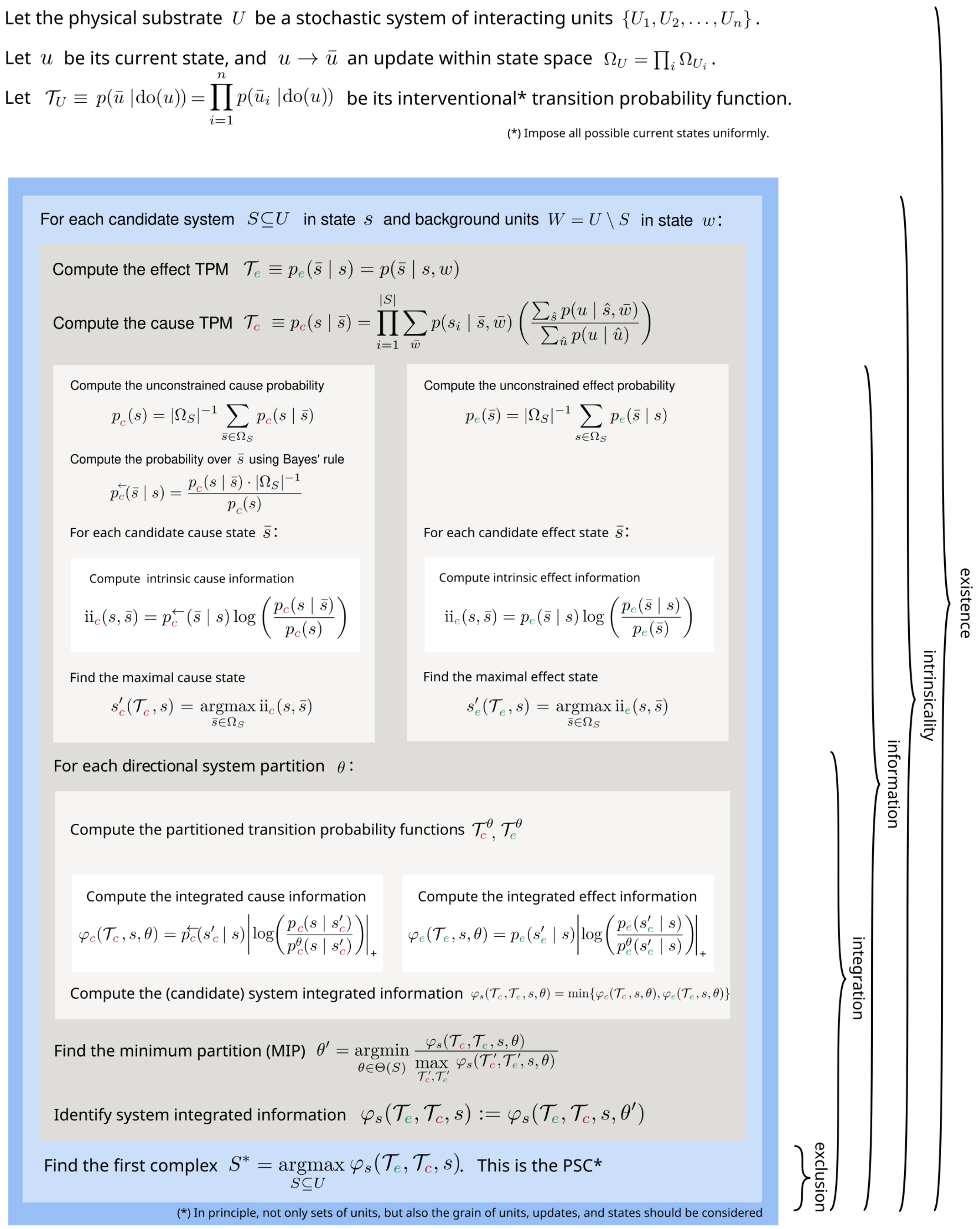
$$2^{|\mathcal{Z}(o)|} - |\mathcal{Z}(o)| - 1.$$

We can count all the relations by applying the inclusion-exclusion principle (from combinatorics) as:

$$\sum_{o \subseteq s'_c \cup s'_e} (-1)^{|o|-1} \left(2^{|\mathcal{Z}(o)|} - |\mathcal{Z}(o)| - 1\right).$$

This is the number of all the causal relations among the causal distinctions in $D(\mathcal{T}_e, \mathcal{T}_c, s)$, except the self-relations. Again, the self-relations can be counted individually without combinatorial explosion.

B.4 IIT Algorithm



Let the physical substrate U be a stochastic system of interacting units $\{U_1, U_2, \ldots, U_n\}$.

Let u be its current state, and $u \to \bar{u}$ an update within state space $\Omega_U = \prod_i \Omega_{U_i}$.

Let $\mathcal{T}_U \equiv p(\bar{u} \,|\mathrm{do}(u)) = \prod_{i=1}^{n} p(\bar{u}_i \,|\mathrm{do}(u))$ be its interventional* transition probability function.

(*) Impose all possible current states uniformly.

For each candidate system $S \subseteq U$ in state s and background units $W = U \setminus S$ in state w:

Compute the effect TPM $\mathcal{T}_e \equiv p_e(\bar{s} \mid s) = p(\bar{s} \mid s, w)$

Compute the cause TPM $\mathcal{T}_c \equiv p_c(s \mid \bar{s}) = \prod_{i=1}^{|S|} \sum_{\bar{w}} p(s_i \mid \bar{s}, \bar{w}) \left(\frac{\sum_{\hat{s}} p(u \mid \hat{s}, \bar{w})}{\sum_{\hat{u}} p(u \mid \hat{u})} \right)$

Compute the unconstrained cause probability

$$p_c(s) = |\Omega_S|^{-1} \sum_{\bar{s} \in \Omega_S} p_c(s \mid \bar{s})$$

Compute the probability over $\bar{s}$ using Bayes' rule

$$\overleftarrow{p}_c(\bar{s} \mid s) = \frac{p_c(s \mid \bar{s}) \cdot |\Omega_S|^{-1}}{p_c(s)}$$

For each candidate cause state $\bar{s}$:

Compute intrinsic cause information

$$\mathrm{ii}_c(s, \bar{s}) = \overleftarrow{p}_c(\bar{s} \mid s) \log \left(\frac{p_c(s \mid \bar{s})}{p_c(s)} \right)$$

Find the maximal cause state

$$s'_c(\mathcal{T}_c, s) = \operatorname*{argmax}_{\bar{s} \in \Omega_S} \mathrm{ii}_c(s, \bar{s})$$

Compute the unconstrained effect probability

$$p_e(\bar{s}) = |\Omega_S|^{-1} \sum_{s \in \Omega_S} p_e(\bar{s} \mid s)$$

For each candidate effect state $\bar{s}$:

Compute intrinsic effect information

$$\mathrm{ii}_e(s, \bar{s}) = p_e(\bar{s} \mid s) \log \left(\frac{p_e(\bar{s} \mid s)}{p_e(\bar{s})} \right)$$

Find the maximal effect state

$$s'_e(\mathcal{T}_e, s) = \operatorname*{argmax}_{\bar{s} \in \Omega_S} \mathrm{ii}_e(s, \bar{s})$$

For each directional system partition θ:

Compute the partitioned transition probability functions $\mathcal{T}_c^\theta, \mathcal{T}_e^\theta$

Compute the integrated cause information

$$\varphi_c(\mathcal{T}_c, s, \theta) = \overleftarrow{p}_c(s'_c \mid s) \left| \log \left(\frac{p_c(s \mid s'_c)}{p_c^\theta(s \mid s'_c)} \right) \right|_+$$

Compute the integrated effect information

$$\varphi_e(\mathcal{T}_e, s, \theta) = p_e(s'_e \mid s) \left| \log \left(\frac{p_e(s'_e \mid s)}{p_e^\theta(s'_e \mid s)} \right) \right|_+$$

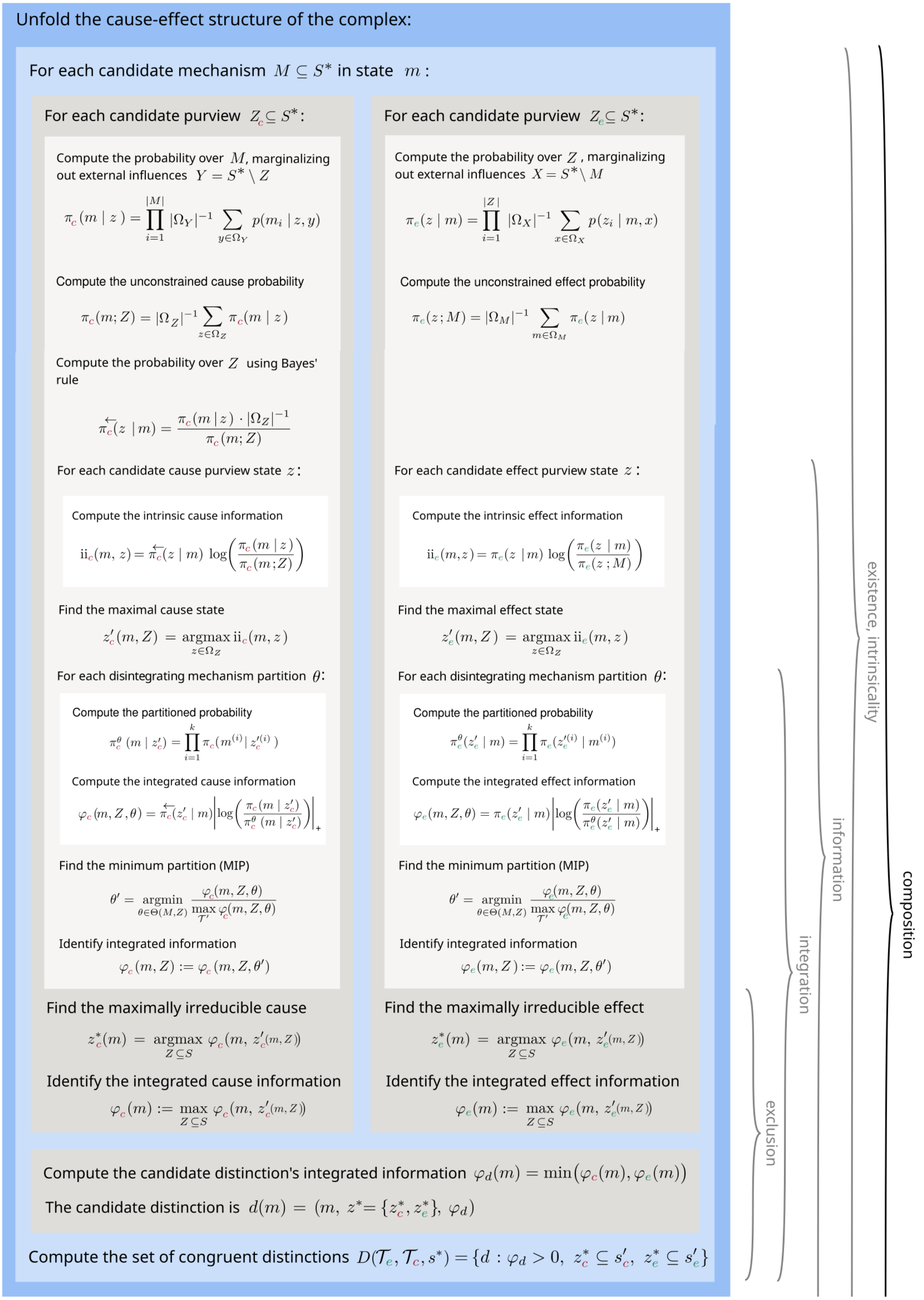
Compute the (candidate) system integrated information $\varphi_s(\mathcal{T}_c, \mathcal{T}_e, s, \theta) = \min\{\varphi_c(\mathcal{T}_c, s, \theta), \varphi_e(\mathcal{T}_e, s, \theta)\}$

Find the minimum partition (MIP) $\theta' = \operatorname*{argmin}_{\theta \in \Theta(S)} \frac{\varphi_s(\mathcal{T}_c, \mathcal{T}_e, s, \theta)}{\max_{\mathcal{T}'_c, \mathcal{T}'_e} \varphi_s(\mathcal{T}'_c, \mathcal{T}'_e, s, \theta)}$

Identify system integrated information $\varphi_s(\mathcal{T}_e, \mathcal{T}_c, s) := \varphi_s(\mathcal{T}_e, \mathcal{T}_c, s, \theta')$

Find the first complex $S^* = \operatorname*{argmax}_{S \subseteq U} \varphi_s(\mathcal{T}_e, \mathcal{T}_c, s)$. This is the PSC*

(*) In principle, not only sets of units, but also the grain of units, updates, and states should be considered



Unfold the cause-effect structure of the complex:

For each candidate mechanism $M \subseteq S^*$ in state m:

For each candidate purview $Z_c \subseteq S^*$:

Compute the probability over M, marginalizing out external influences $Y = S^* \setminus Z$

$$\pi_c(m \mid z) = \prod_{i=1}^{|M|} |\Omega_Y|^{-1} \sum_{y \in \Omega_Y} p(m_i \mid z, y)$$

Compute the unconstrained cause probability

$$\pi_c(m; Z) = |\Omega_Z|^{-1} \sum_{z \in \Omega_Z} \pi_c(m \mid z)$$

Compute the probability over Z using Bayes' rule

$$\overleftarrow{\pi}_c(z \mid m) = \frac{\pi_c(m \mid z) \cdot |\Omega_Z|^{-1}}{\pi_c(m; Z)}$$

For each candidate cause purview state z:

Compute the intrinsic cause information

$$\mathrm{ii}_c(m, z) = \overleftarrow{\pi}_c(z \mid m) \log\left(\frac{\pi_c(m \mid z)}{\pi_c(m; Z)}\right)$$

Find the maximal cause state

$$z'_c(m, Z) = \operatorname*{argmax}_{z \in \Omega_Z} \mathrm{ii}_c(m, z)$$

For each disintegrating mechanism partition θ:

Compute the partitioned probability

$$\pi_c^\theta(m \mid z'_c) = \prod_{i=1}^{k} \pi_c(m^{(i)} \mid z'^{(i)}_c)$$

Compute the integrated cause information

$$\varphi_c(m, Z, \theta) = \overleftarrow{\pi}_c(z'_c \mid m) \left| \log\left(\frac{\pi_c(m \mid z'_c)}{\pi_c^\theta(m \mid z'_c)}\right)\right|_+$$

Find the minimum partition (MIP)

$$\theta' = \operatorname*{argmin}_{\theta \in \Theta(M,Z)} \frac{\varphi_c(m, Z, \theta)}{\max_{\mathcal{T}'} \varphi_c(m, Z, \theta)}$$

Identify integrated information

$$\varphi_c(m, Z) := \varphi_c(m, Z, \theta')$$

Find the maximally irreducible cause

$$z^*_c(m) = \operatorname*{argmax}_{Z \subseteq S} \varphi_c(m, z'_c(m, Z))$$

Identify the integrated cause information

$$\varphi_c(m) := \max_{Z \subseteq S} \varphi_c(m, z'_c(m, Z))$$

For each candidate purview $Z_e \subseteq S^*$:

Compute the probability over Z, marginalizing out external influences $X = S^* \setminus M$

$$\pi_e(z \mid m) = \prod_{i=1}^{|Z|} |\Omega_X|^{-1} \sum_{x \in \Omega_X} p(z_i \mid m, x)$$

Compute the unconstrained effect probability

$$\pi_e(z; M) = |\Omega_M|^{-1} \sum_{m \in \Omega_M} \pi_e(z \mid m)$$

For each candidate effect purview state z:

Compute the intrinsic effect information

$$\mathrm{ii}_e(m, z) = \pi_e(z \mid m) \log\left(\frac{\pi_e(z \mid m)}{\pi_e(z; M)}\right)$$

Find the maximal effect state

$$z'_e(m, Z) = \operatorname*{argmax}_{z \in \Omega_Z} \mathrm{ii}_e(m, z)$$

For each disintegrating mechanism partition θ:

Compute the partitioned probability

$$\pi_e^\theta(z'_e \mid m) = \prod_{i=1}^{k} \pi_e(z'^{(i)}_e \mid m^{(i)})$$

Compute the integrated effect information

$$\varphi_e(m, Z, \theta) = \pi_e(z'_e \mid m) \left| \log\left(\frac{\pi_e(z'_e \mid m)}{\pi_e^\theta(z'_e \mid m)}\right)\right|_+$$

Find the minimum partition (MIP)

$$\theta' = \operatorname*{argmin}_{\theta \in \Theta(M,Z)} \frac{\varphi_e(m, Z, \theta)}{\max_{\mathcal{T}'} \varphi_e(m, Z, \theta)}$$

Identify integrated information

$$\varphi_e(m, Z) := \varphi_e(m, Z, \theta')$$

Find the maximally irreducible effect

$$z^*_e(m) = \operatorname*{argmax}_{Z \subseteq S} \varphi_e(m, z'_e(m, Z))$$

Identify the integrated effect information

$$\varphi_e(m) := \max_{Z \subseteq S} \varphi_e(m, z'_e(m, Z))$$

Compute the candidate distinction's integrated information $\varphi_d(m) = \min\big(\varphi_c(m), \varphi_e(m)\big)$

The candidate distinction is $d(m) = (m,\ z^* = \{z^*_c, z^*_e\},\ \varphi_d)$

Compute the set of congruent distinctions $D(\mathcal{T}_e, \mathcal{T}_c, s^*) = \{d : \varphi_d > 0,\ z^*_c \subseteq s'_c,\ z^*_e \subseteq s'_e\}$

```
For each candidate set of distincitons $\boldsymbol{d} \subseteq D(\mathcal{T}_e, \mathcal{T}_c, s^*)$ :
    For each set of causes and/or effects $\boldsymbol{z}$ such that:
        $\boldsymbol{z} \;:\; \boldsymbol{z} \cap \{z_c^*(d), z_e^*(d)\} \neq \varnothing \;\; \forall d \in \boldsymbol{d}, \;\; \bigcap_{z^* \in \boldsymbol{z}} z^* \neq \varnothing, \;\; |\boldsymbol{z}| > 1$
            Compute the maximal overlap $o^*(\boldsymbol{z}) = \bigcap_{z \in \boldsymbol{z}} z \neq \varnothing$
            The relation face is $f(\boldsymbol{z}) = \big(\boldsymbol{z}, o^*(\boldsymbol{z})\big)$
        The set of relation faces is $\boldsymbol{f}(\boldsymbol{d}) = \{f(\boldsymbol{z})\}_{\boldsymbol{d}}$
        Compute the integrated information of the candidate relation
            $\varphi_r(\boldsymbol{d}) = \min_{d \in \boldsymbol{d}} \left| \bigcup_{f \in \boldsymbol{f}(\boldsymbol{d})} o_f^* \right| \frac{\varphi_d}{|z_c^*(d) \cup z_e^*(d)|}$
        The candidate relation is $r(\boldsymbol{d}) = \big(\boldsymbol{d}, \boldsymbol{f}(\boldsymbol{d}), \varphi_r\big)$
    Compute the set of relations $R(D) = \{r(\boldsymbol{d}) : \varphi_r(\boldsymbol{d}) > 0\}$
The cause-effect structure of the complex $S^*$ (its Φ-structure) is
    $C(\mathcal{T}_e, \mathcal{T}_c, s^*) = D \cup R(D) = \{D(\mathcal{T}_e, \mathcal{T}_c, s^*) \cup R(D(\mathcal{T}_e, \mathcal{T}_c, s^*))\}$
Compute $\Phi(\mathcal{T}_e, \mathcal{T}_c, s^*) = \sum_{C(\mathcal{T}_e, \mathcal{T}_c, s^*)} \varphi.$
```

integration

exclusion

C Chapter 4 supporting information

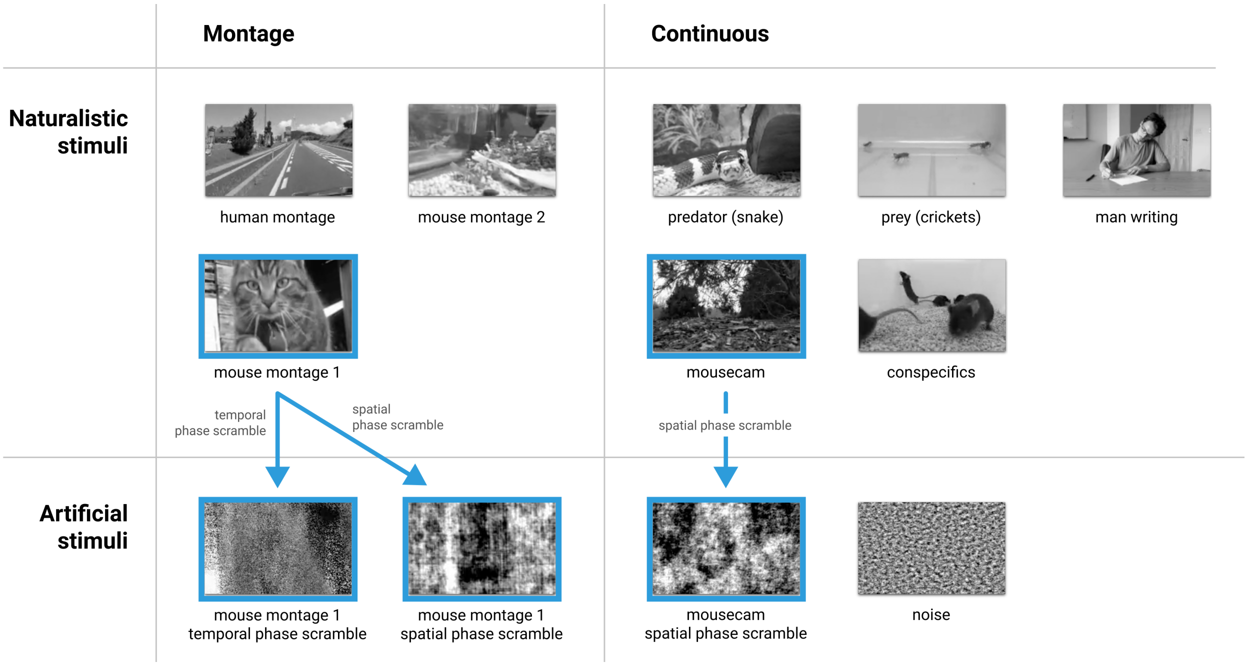


Figure C.1. Stimuli. Twelve 30 s long greyscale naturalistic (*top*) and artificial (*bottom*) movie stimuli were presented. *Left:* montages of six 5 s clips; *right:* continuous 30 s clips. Stimuli used in the main analysis are outlined in blue. Arrows indicate the phase-scrambling procedures.

Table C.1. Number of cells recorded per layer and area.

Layer	Area	Cells
L2/3	V1	451
L2/3	LM	353
L2/3	AL	336
L2/3	PM	259
L2/3	AM	284
L4	V1	649
L4	LM	456
L4	AL	479
L4	PM	135
L4	AM	62
L5	V1	126
L5	LM	185
L5	AL	86
L5	PM	76
L5	AM	81

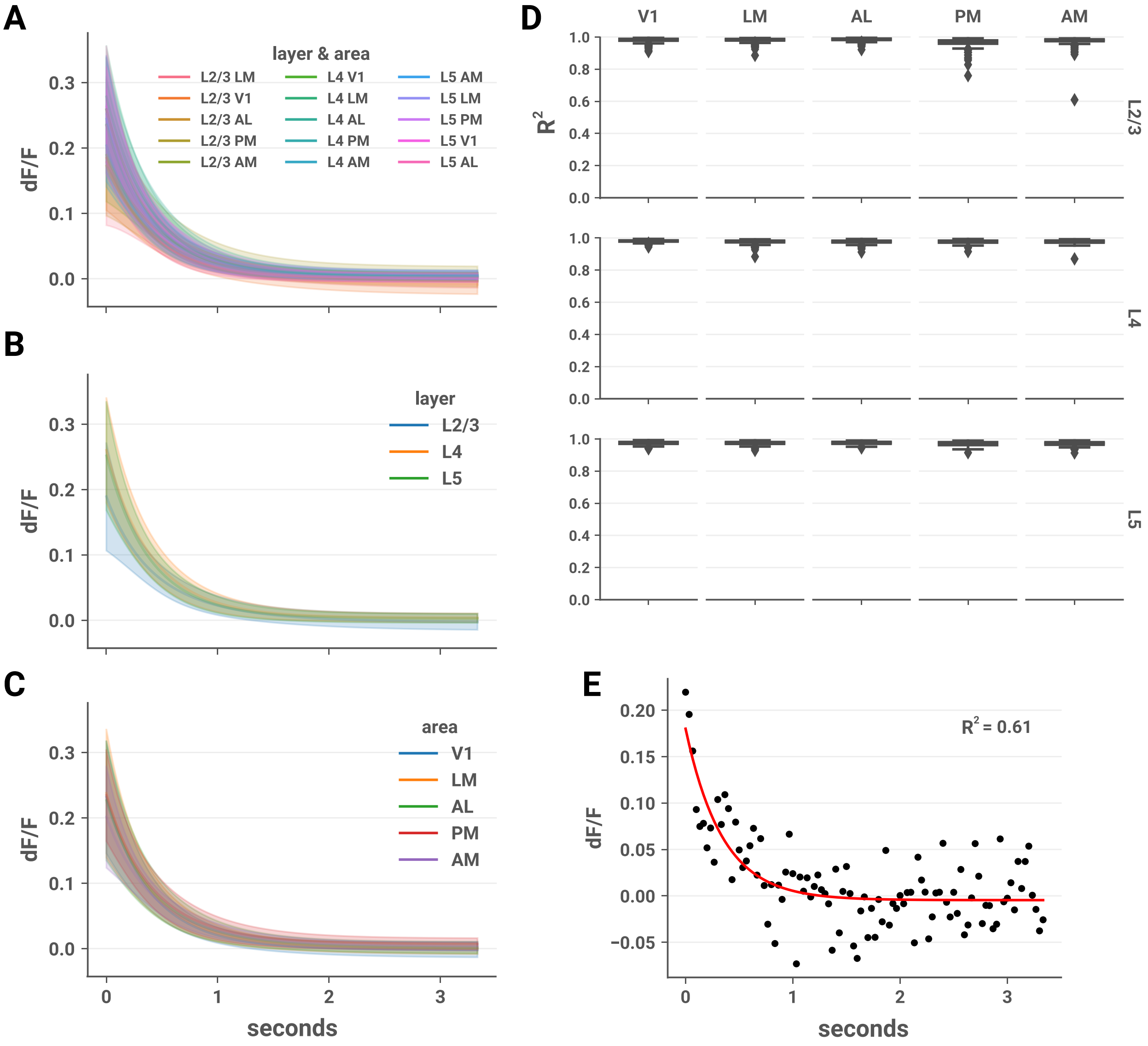


Figure C.2. Calcium indicator kinetics did not differ across cell populations. Mean (solid line) ± standard deviation (shaded region) calcium response averaged by **(A)** layer and area, **(B)** layer, and **(C)** area. Calcium responses were obtained for each cell by selecting isolated events (those without any other events occurring in the preceding 50 ms or the following 100 ms) and computing the mean event-locked trace (§ 4.2, Event detection). **(D)** Responses were well-fit by an exponential decay function; R^2 values of the fit for each cell are plotted by layer and area. **(E)** The fit with the lowest R^2 value, 0.61 (cell 5 in session 718673398, L2/3 AM; fit in red, data in black). We tested for a relationship between layer, area, and response half-life by fitting a linear mixed effects model with layer, area, and their interaction as fixed effects and experimental session as a random effect, and comparing this to a model without the interaction term. We found no layer × area interaction (likelihood ratio test; $\chi^2(8) = 1.293$, $p = 0.996$). We tested for main effects of layer and area in two further models and likewise found none (layer: $\chi^2(2) = 1.143$, $p = 0.565$; area: $\chi^2(4) = 0.2288$, $p = 0.994$).

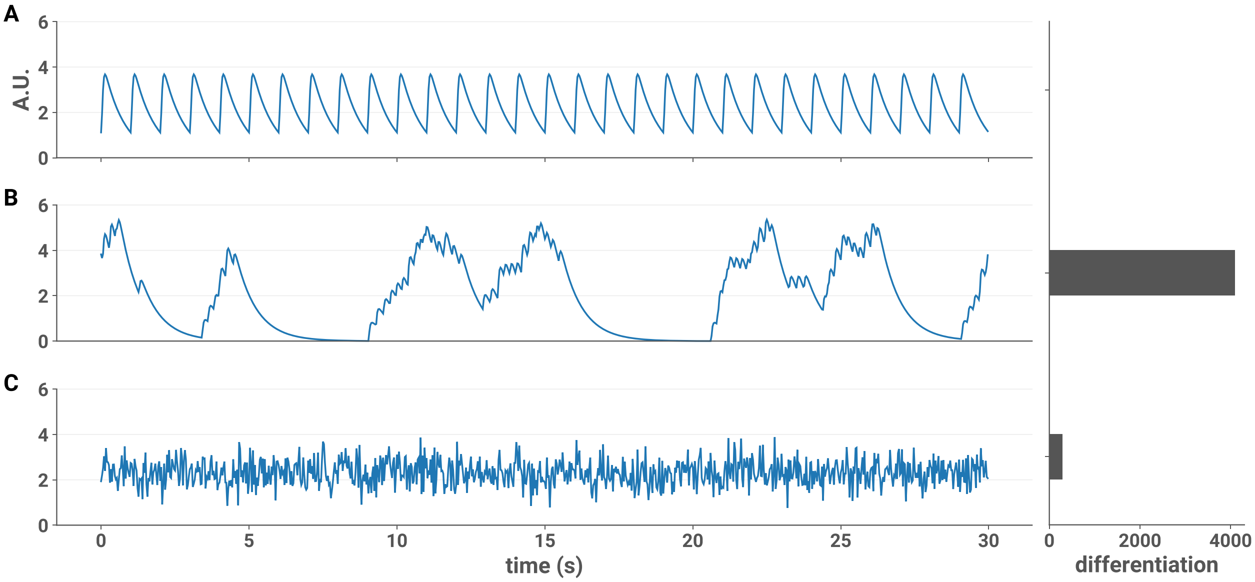


Figure C.3. Differentiation for simulated signals. To illustrate how the ND measure behaves, we generated three artificial signals and computed ND for each. Signals were normalized to have the same energy. **(A, B)** Artificial spike trains were convolved with an idealized GCaMP6f response kernel (difference of exponentials; decay time constant 0.6 s, rise time constant 0.05 s; Chen et al., 2013; Pachitariu et al., 2018) and downsampled to 30 Hz. **(A)** Periodic bursting at 1 Hz. Because the period is the same as the window length used in the spectral estimation step (Figure 4.2A), the estimated spectrum of each window is identical, and differentiation is zero. **(B)** An irregular firing pattern has high differentiation. **(C)** Gaussian noise. The theoretical spectrum is identical for each window, but differentiation is nonzero due to the spectral estimation error resulting from the finite window length.

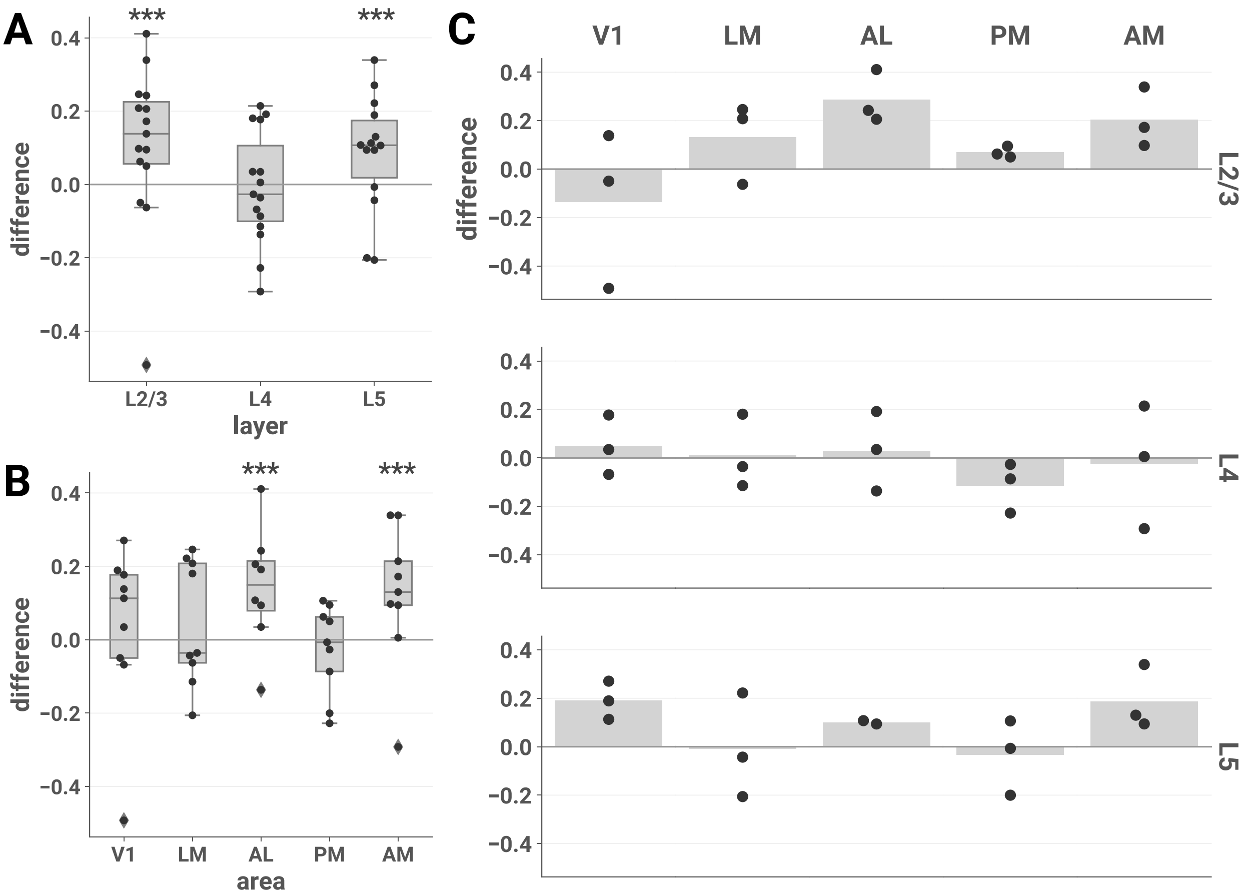


Figure C.4. Naturalistic *vs.* artificial differences in ND across the entire stimulus set. The mean difference in ND of responses to all 8 naturalistic *vs.* all 4 artificial stimuli is plotted for each session by layer **(A)**, area **(B)**, and layer-area pair **(C)**. Results are similar to the unscrambled *vs.* scrambled contrast shown in Figure 4.3. In this analysis, *post hoc* tests showed a significant effect also in L5; however, this contrast does not control for low-level stimulus characteristics and is thus harder to interpret. **(A)** We fit an LME model with stimulus category (naturalistic or artificial), layer, and their interaction as fixed effects and found a significant interaction (likelihood ratio test, $\chi^2(2) = 16.343$, $p = 2.83 \times 10^{-4}$). *Post hoc* one-sided z-tests (adjusted for multiple comparisons): L2/3, $z = 4.974$, $p = 9.82 \times 10^{-7}$, Cohen's $d = 0.153$, 95% CI $[0.064, \infty)$; L4, $z = -0.450$, $p = 0.965$, Cohen's $d = -0.019$, 95% CI $[-0.057, \infty)$; L5, $z = 3.745$, $p = 2.71e - 4$, Cohen's $d = 0.144$, 95% CI $[0.037, \infty)$. **(B)** We fit an LME model with stimulus category (naturalistic or artificial), area, and their interaction as fixed effects and found a significant interaction (likelihood ratio test, $\chi^2(2) = 16.343$, $p = 0.000283$). *Post hoc* one-sided z-tests (adjusted for multiple comparisons): V1, $z = 1.207$, $p = 0.453$, Cohen's $d = 0.066$, 95% CI $[-0.032, \infty)$; LM, $z = 1.523$, $p = 0.281$, Cohen's $d = 0.074$, 95% CI $[-0.023, \infty)$; AL, $z = 4.715$, $p = 6.04e - 6$, Cohen's $d = 0.222$, 95% CI $[0.073, \infty)$; PM, $z = -0.907$, $p > 0.999$, Cohen's $d = -0.040$, 95% CI $[-0.093, \infty)$; AM, $z = 4.249$, $p = 5.37e - 05$, Cohen's $d = 0.156$, 95% CI $[0.056, \infty)$. **(A)** and **(B)**: asterisks indicate significant *post hoc* tests in the layer (A) and area (B) interaction LME models (***, $p < 0.001$). Boxes indicate quartiles; whiskers indicate the minimum and maximum of data lying within 1.5 times the inter-quartile range of the 25% or 75% quartiles; diamonds indicate observations outside this range. **(C)** Mean values are indicated by bars.

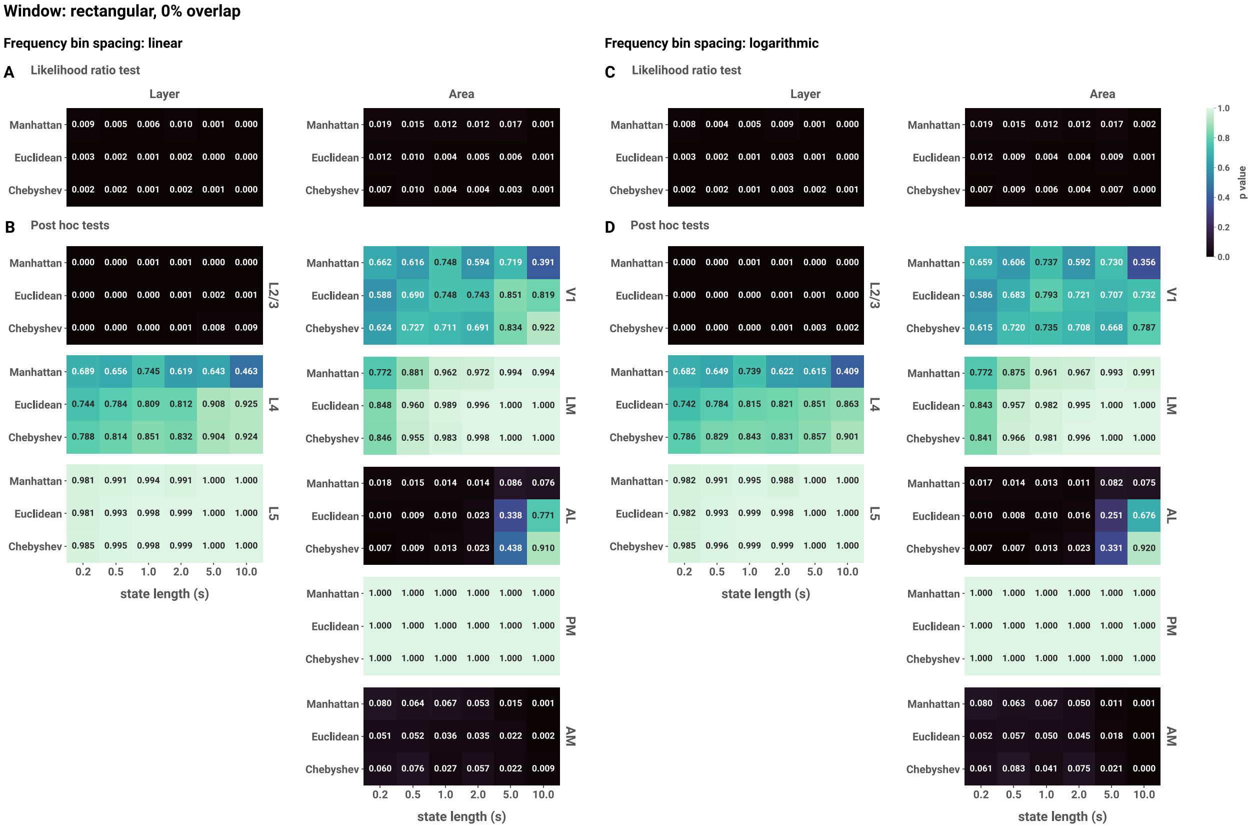


Figure C.5. Sensitivity analysis of main ND results. We investigated the sensitivity of our results to changes in various parameters of the ND calculation. We systematically varied (1) the distance metric used to compare population state vectors (vertical axis of heatmaps), (2) the length of the window that defines a single state, in which the spectrum is estimated (horizontal axis of heatmaps), (3) the spacing of the frequency bins in the estimated spectrum (linear, **A** and **B**; logarithmic, **C** and **D**); and (4) the window type and amount of overlap used in estimating the spectra across the stimulus presentation (shown in the following two figures). For each combination of these parameters, we computed ND values and performed the same statistical analysis as described in the main text. Each cell in the heatmaps in **A** and **C** shows the p value of the likelihood ratio test for the stimulus category × layer interaction (*left*) and stimulus category × area interaction (*right*); cells in **B** and **D** show p values for the associated *post hoc* tests. The results reported in the main text correspond to the second row and third column of the heatmaps in **A** and **B.** For nearly all other combinations of parameters, we likewise find that unscrambled stimuli elicit increased ND specifically in L2/3 of AL & AM.

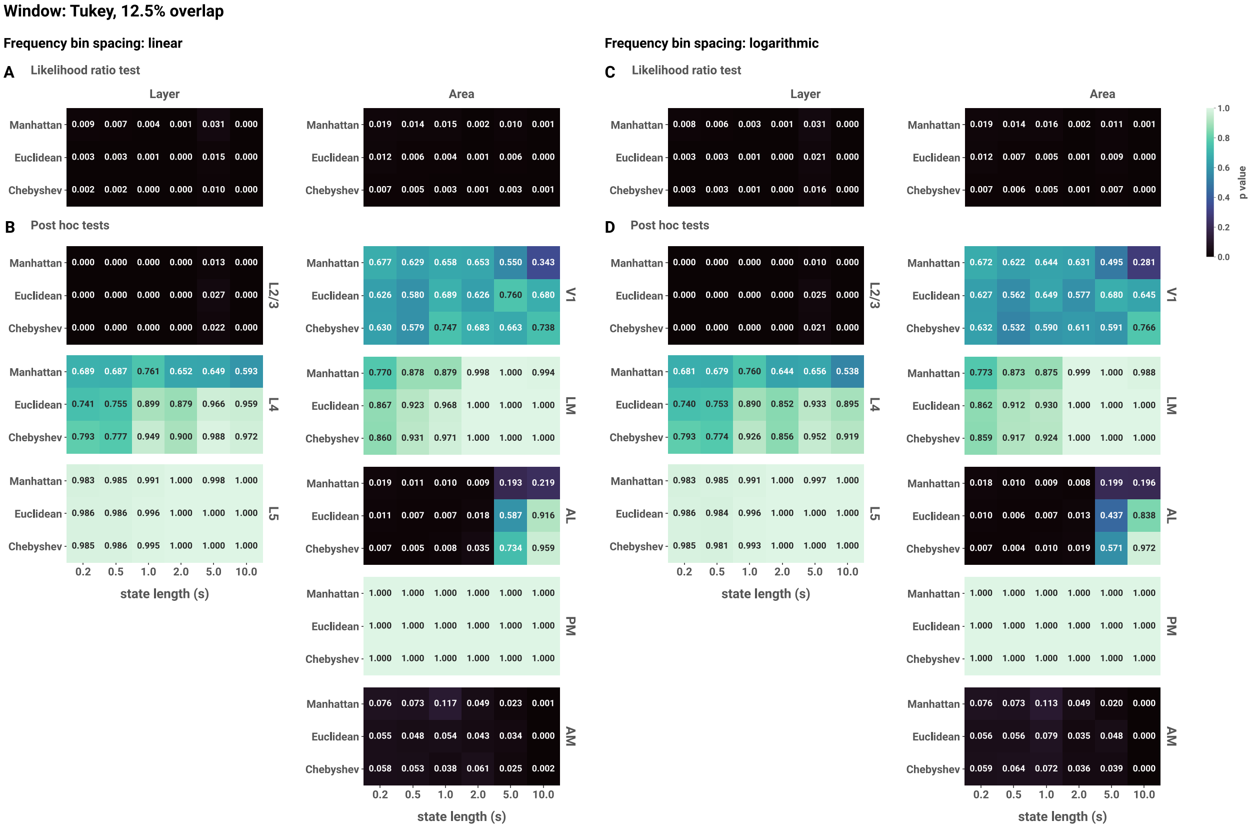


Figure C.6. Sensitivity analysis was performed as described in Figure C.5, except that the time-frequency analysis step of computing ND (Figure 4.2A) was performed using a Tukey window with 12.5% overlap.

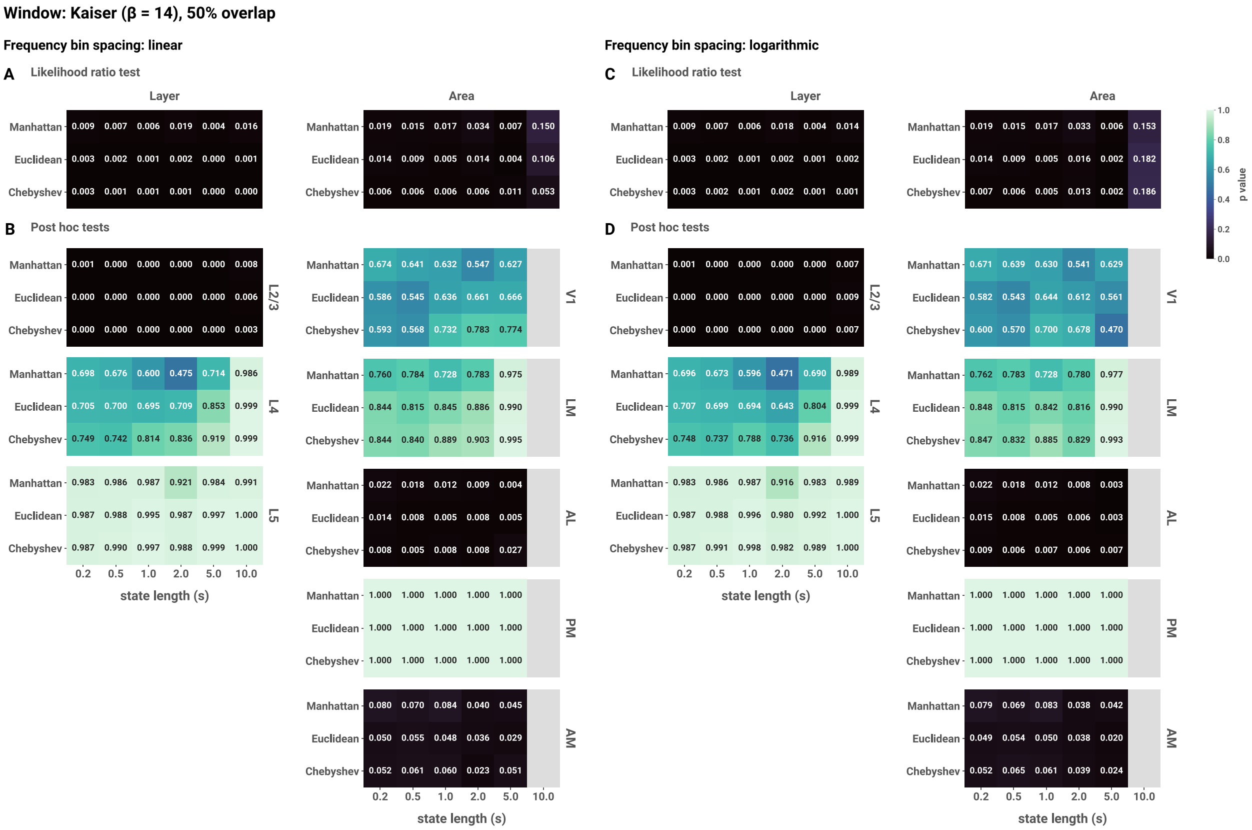


Figure C.7. Sensitivity analysis was performed as described in Figure C.5, except with a Kaiser window ($\beta = 14$) with 50% overlap.

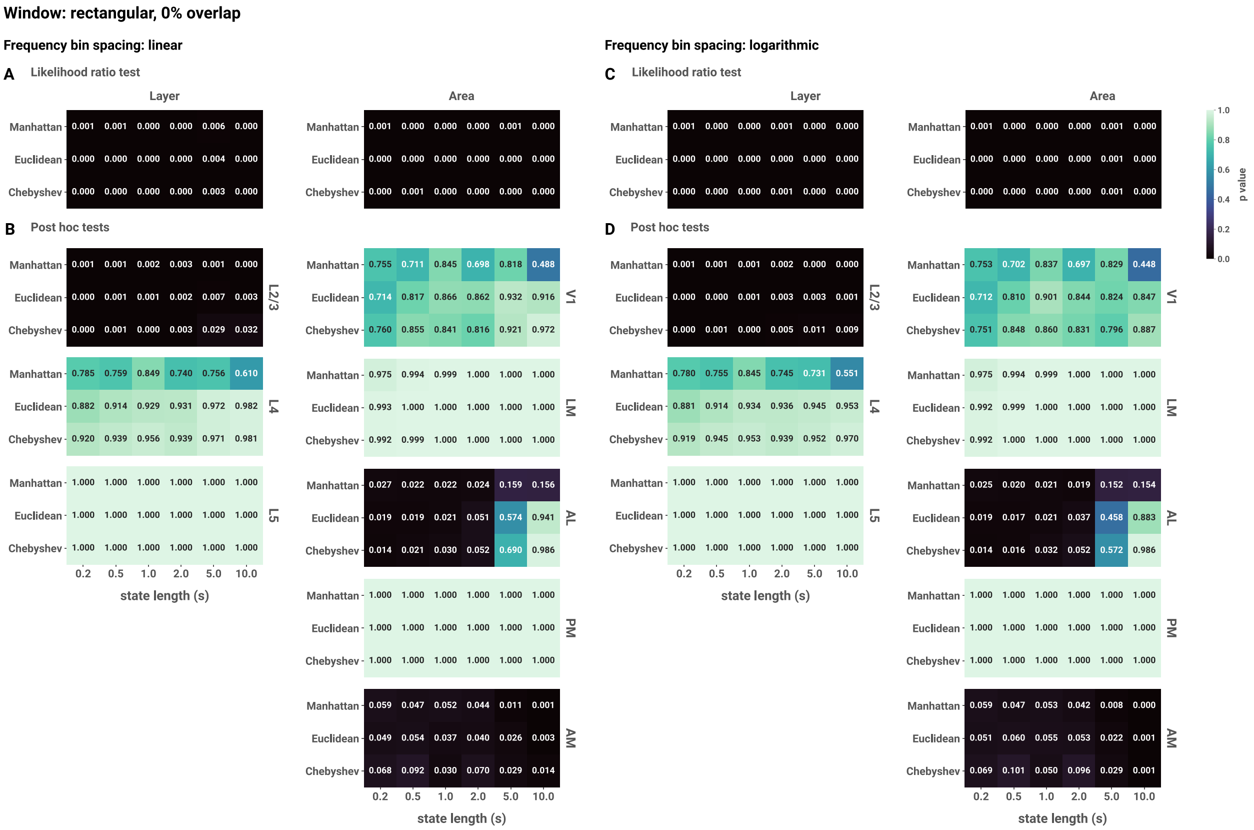


Figure C.8. Sensitivity analysis for LME models including arousal variables (locomotion and pupil diameter) as covariates. Consistent with results from the simpler models, L2/3 of AL & AM emerge as the cell populations in which ND is greater for unscrambled *vs.* scrambled stimuli for nearly all parameter combinations.

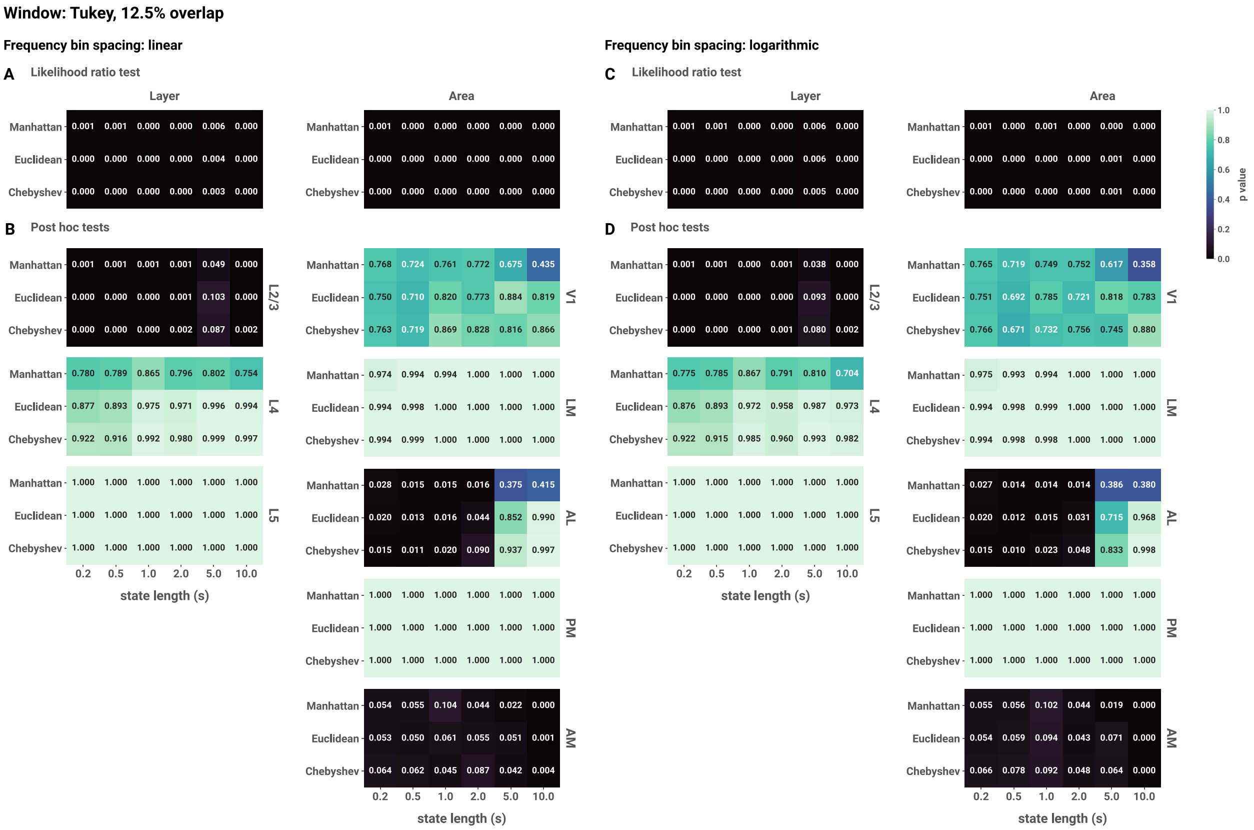


Figure C.9. Sensitivity analysis as in Figure C.8, using a Tukey window with 12.5% overlap.

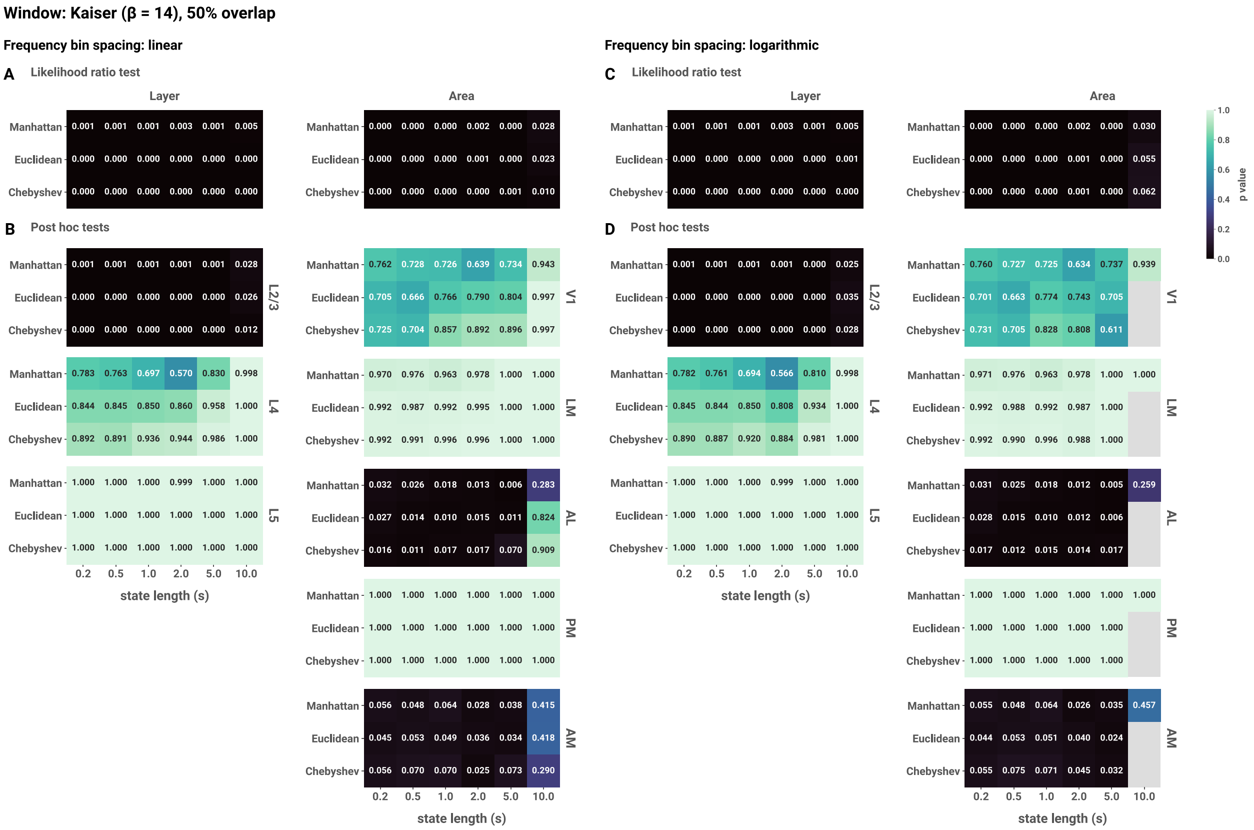


Figure C.10. Sensitivity analysis as in Figure C.8, using a Kaiser window ($\beta = 14$) with 50% overlap.

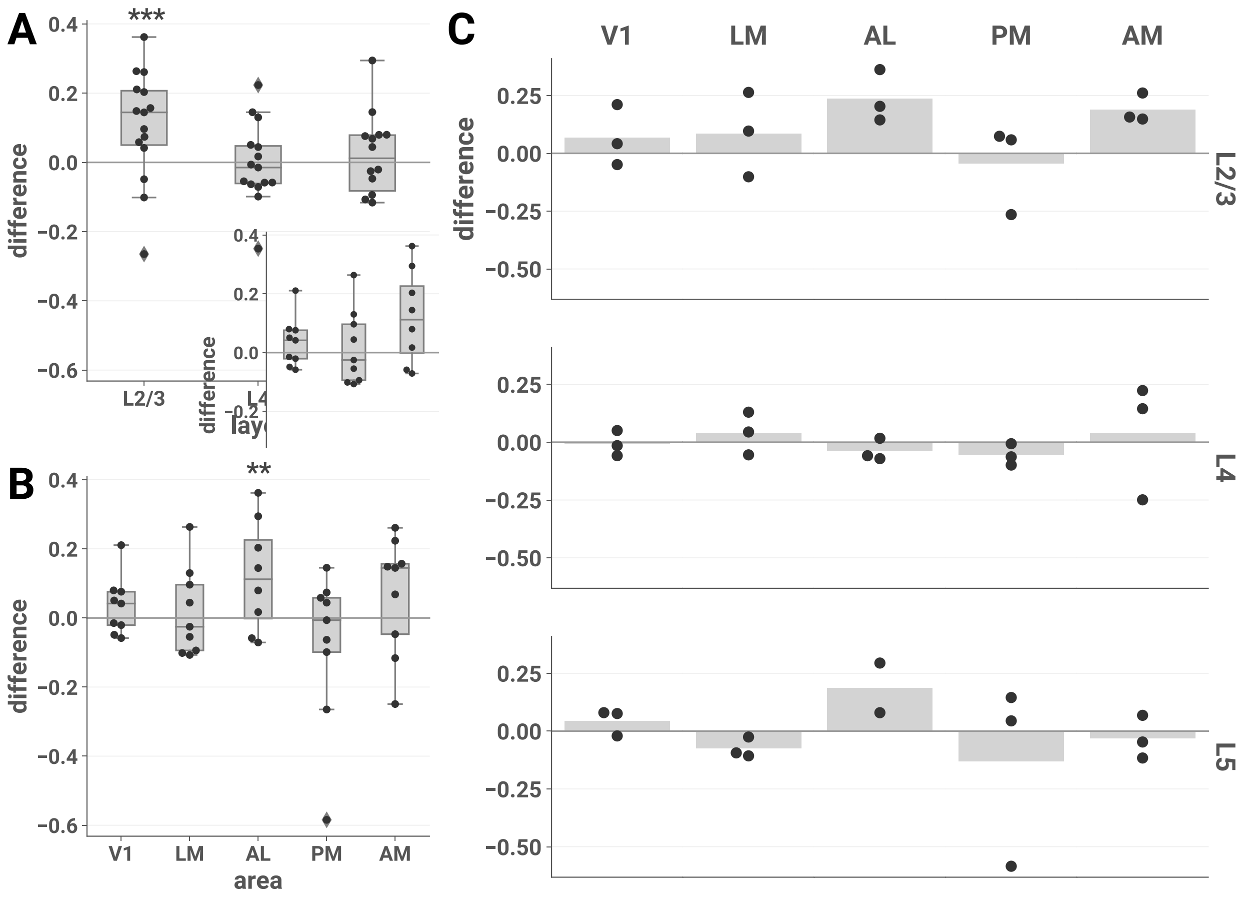


Figure C.11. Spectral differentiation analysis of discrete L_0 calcium events. The difference in ND of responses to unscrambled *vs.* scrambled stimuli is plotted for each session by layer **(A)**, area **(B)**, and layer-area pair **(C)**. Results are similar to the main analysis on $\Delta F / F_0$ traces shown in Figure 4.3. This indicates that the differences we observed in ND are driven by differences in the large-timescale patterns of responses rather than small-timescale spectral differences within the windows, consistent with the sparsity of calcium responses in our dataset. **(A)** We fit an LME model with stimulus category (naturalistic or artificial), layer, and their interaction as fixed effects and found a significant interaction (likelihood ratio test, $\chi^2(2) = 11.481$, $p = 0.00321$. *Post hoc* one-sided z-tests (adjusted for multiple comparisons): L2/3, $z = 3.835$, $p = 0.000188$, Cohen's $d = 0.175$, 95% CI $[0.0478, \infty)$; L4, $z = -0.154$, $p = 0.916$, Cohen's $d = -0.001$, 95% CI $[-0.0635, \infty)$; L5, $z = -0.507$, $p = 0.971$, Cohen's $d = -0.0318$, 95% CI $[-0.0761, \infty)$. **(B)** We fit an LME model with stimulus category (naturalistic or artificial), area, and their interaction as fixed effects and found a significant interaction (likelihood ratio test, $\chi^2(4) = 15.102$, $p = 0.00445$). *Post hoc* one-sided z-tests (adjusted for multiple comparisons): V1, $z = 0.612$, $p = 0.793$, Cohen's $d = 0.076$, 95% CI $[-0.0483, \infty)$; LM, $z = -0.136$, $p = 0.982$, Cohen's $d = 0.0270$, 95% CI $[-0.0665, \infty)$; AL, $z = 2.879$, $p = 0.00995$, Cohen's $d = 0.226$, 95% CI $[0.0329, \infty)$; PM, $z = -1.929$, $p > 0.999$, Cohen's $d = -0.151$, 95% CI $[-0.161, \infty)$; AM, $z = 2.318$, $p = 0.05005$, Cohen's $d = 0.0984$, 95% CI $[-0.0179, \infty)$. **(A)** and **(B)**: asterisks indicate significant *post hoc* one-sided z-tests in the layer **(A)** and area **(B)** interaction LME models (**, $p < 0.01$; ***, $p < 0.001$). Boxes indicate quartiles; whiskers indicate the minimum and maximum of data lying within 1.5 times the inter-quartile range of the 25% or 75% quartiles; diamonds indicate observations outside this range. **(C)** Mean values are indicated by bars.

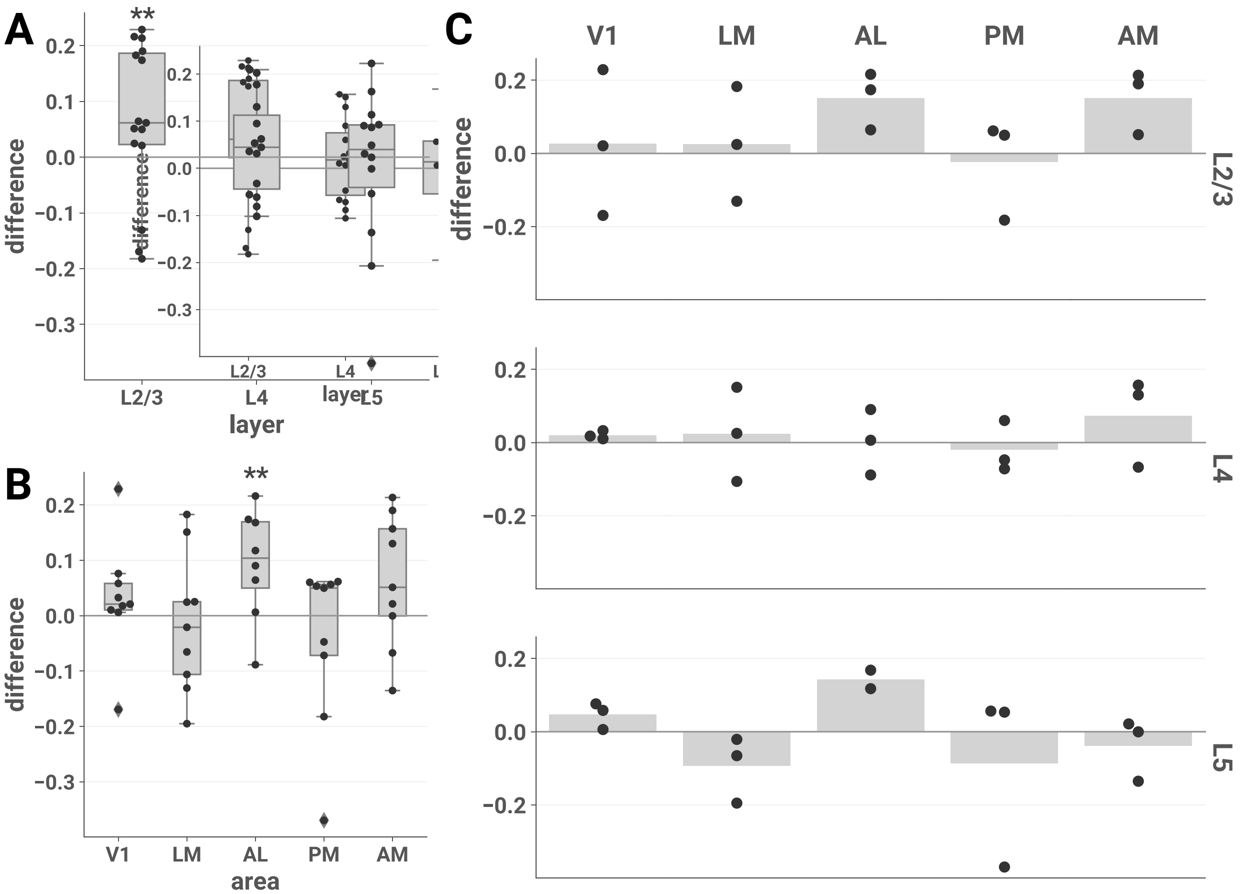


Figure C.12. Spectral differentiation analysis on $\Delta F/F_0$ traces with initial calcium transients removed. Transients were defined as the first 200 ms (6 imaging samples) of the signal after each L_0 calcium event. These samples were replaced with linearly interpolated values and ND was calculated for the resulting signal. The difference in ND of responses to unscrambled *vs.* scrambled stimuli is plotted for each session by layer **(A)**, area **(B)**, and layer-area pair **(C)**. Results are similar to the main analysis shown in Figure 4.3, indicating that our ND results are not driven solely by initial transients in the calcium response. **(A)** We fit an LME model with stimulus category (naturalistic or artificial), layer, and their interaction as fixed effects and found a significant interaction (likelihood ratio test, $\chi^2(2) = 6.024$, $p = 0.0492$. *Post hoc* one-sided z-tests (adjusted for multiple comparisons): L2/3, $z = 2.823$, $p = 0.00713$, Cohen's $d = 0.175$, 95% CI $[0.0165, \infty)$; L4, $z = 0.850$, $p = 0.483$, Cohen's $d = 0.0568$, 95% CI $[-0.0299, \infty)$; L5, $z = -0.674$, $p = 0.984$, Cohen's $d = -0.0357$, 95% CI $[-0.0680, \infty)$. **(B)** We fit an LME model with stimulus category (naturalistic or artificial), area, and their interaction as fixed effects and found a significant interaction (likelihood ratio test, $\chi^2(4) = 12.886$, $p = 0.0119$). *Post hoc* one-sided z-tests (adjusted for multiple comparisons): V1, $z = 1.032$, $p = 0.559$, Cohen's $d = 0.0946$, 95% CI $[-0.0390, \infty)$; LM, $z = -0.495$, $p = 0.997$, Cohen's $d = -0.0354$, 95% CI $[-0.092, \infty)$; AL, $z = 2.911$, $p = 0.00899$, Cohen's $d = 0.257$, 95% CI $[0.0190, \infty)$; PM, $z = -1.429$, $p > 0.999$, Cohen's $d = -0.0966$, 95% CI $[-0.114, \infty)$; AM, $z = 2.055$, $p = 0.0958$, Cohen's $d = 0.125$, 95% CI $[-0.00800, \infty)$. **(A)** and **(B)**: asterisks indicate significant *post hoc* one-sided z-tests in the layer **(A)** and area **(B)** interaction LME models (**, $p < 0.01$). Boxes indicate quartiles; whiskers indicate the minimum and maximum of data lying within 1.5 times the inter-quartile range of the 25% or 75% quartiles; diamonds indicate observations outside this range. **(C)** Mean values are indicated by bars.

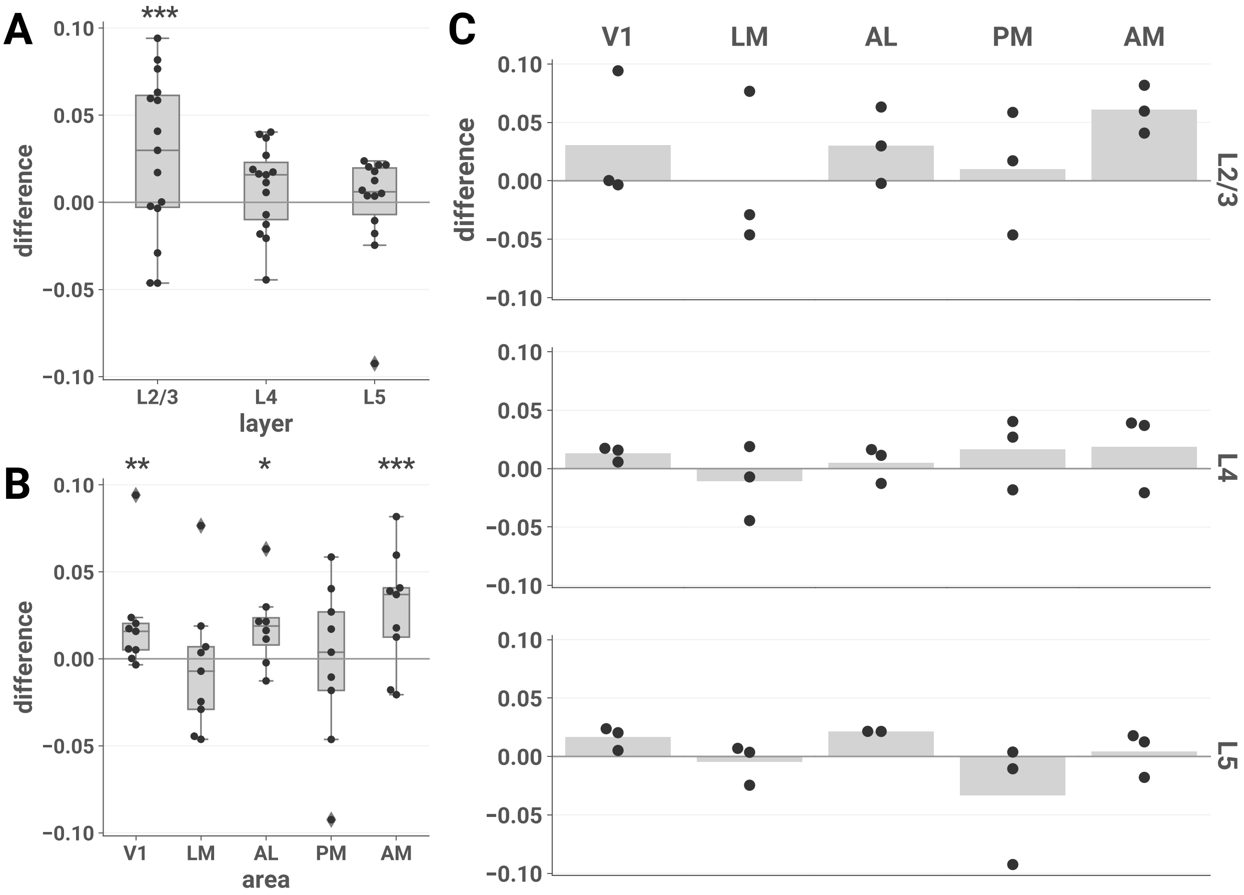


Figure C.13. Multivariate differentiation analysis of detected calcium events. The mean difference in the mean centroid distance of detected-event responses to unscrambled *vs.* scrambled stimuli is plotted for each session by layer **(A)**, area **(B)**, and layer-area pair **(C)**. Multivariate differentiation elicited by unscrambled *vs.* scrambled stimuli is higher in L2/3 of AL and AM, as well as the additional finding of higher differentiation in L2/3 of V1. **(A)** and **(B)**: asterisks indicate significant *post hoc* one-sided *z*-tests in the layer (A) and area (B) interaction LME models (*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$). Boxes indicate quartiles; whiskers indicate the minimum and maximum of data lying within 1.5 times the inter-quartile range of the 25% or 75% quartiles; diamonds indicate observations outside this range. **(C)** Mean values are indicated by bars.

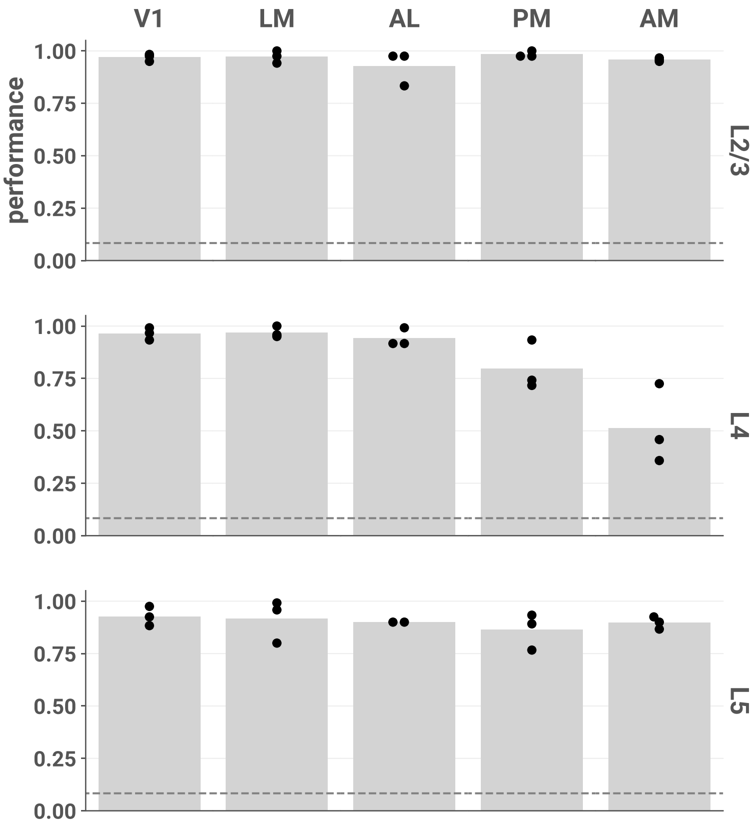


Figure C.14. Stimulus identity can be accurately decoded from most layers and areas using responses to all 12 stimuli. Each point represents the mean fivefold cross-validated balanced accuracy score of linear discriminant analysis performed on a single session (§ 4.2, Decoding analyses). Chance performance is 1/12, indicated by the dotted line.

Table C.2. Mediation analysis of the effect of stimulus. To disentangle the effects of stimulus and arousal on ND, we performed a causal mediation analysis (§ 4.2, Mediation analyses). For each of the 4 stimulus pairs identified as having significantly different levels of neurophysiological differentiation in our *post hoc* analysis, we asked whether each of the arousal variables (locomotion or pupil diameter) was a mediator of the effect of stimulus on differentiation, in each case including the other arousal variable as a covariate. The analysis revealed a mixture of direct and mediated effects. Notably, for the largest contrast (predator *vs.* man writing), for both locomotion and pupil diameter we found evidence only for a direct effect. Overall, we conclude that arousal alone cannot account for differences in ND among continuous, naturalistic stimuli.

Mediator	**Stimulus 1**	**Stimulus 2**	**Effect**	**Estimate**	**95% CI (lower)**	**95% CI (upper)**	***p* value**
Locomotion	Snake (predator)	Man writing	Mediated	−0.002 74	−0.007 07	0.00	0.094
			Direct	0.107	0.0531	0.16	$<2 \times 10^{-16}$
	Snake (predator)	Crickets	Mediated	−0.0078	−0.0170	0.00	0.072
			Direct	0.0293	−0.0265	0.08	0.340
	Mousecam	Man writing	Mediated	−0.012 21	−0.0265	0.00	0.082
			Direct	0.009 12	−0.0487	0.07	0.754
	Snake (predator)	Conspecifics	Mediated	0.006 536	−0.000 658	0.01	0.088
			Direct	0.128 162	0.0721	0.18	$<2 \times 10^{-16}$
Pupil diameter	Snake (predator)	Man writing	Mediated	0.0253	0.0134	0.04	$<2 \times 10^{-16}$
			Direct	0.1073	0.0529	0.16	$<2 \times 10^{-16}$
	Snake (predator)	Crickets	Mediated	0.0575	0.0427	0.07	$<2 \times 10^{-16}$
			Direct	0.0292	−0.0236	0.08	0.270
	Mousecam	Man writing	Mediated	0.079 88	0.0600	0.10	$<2 \times 10^{-16}$
			Direct	0.007 13	−0.0498	0.06	0.802
	Snake (predator)	Conspecifics	Mediated	−0.0385	−0.0527	−0.03	$<2 \times 10^{-16}$
			Direct	0.1247	0.0709	0.18	$<2 \times 10^{-16}$

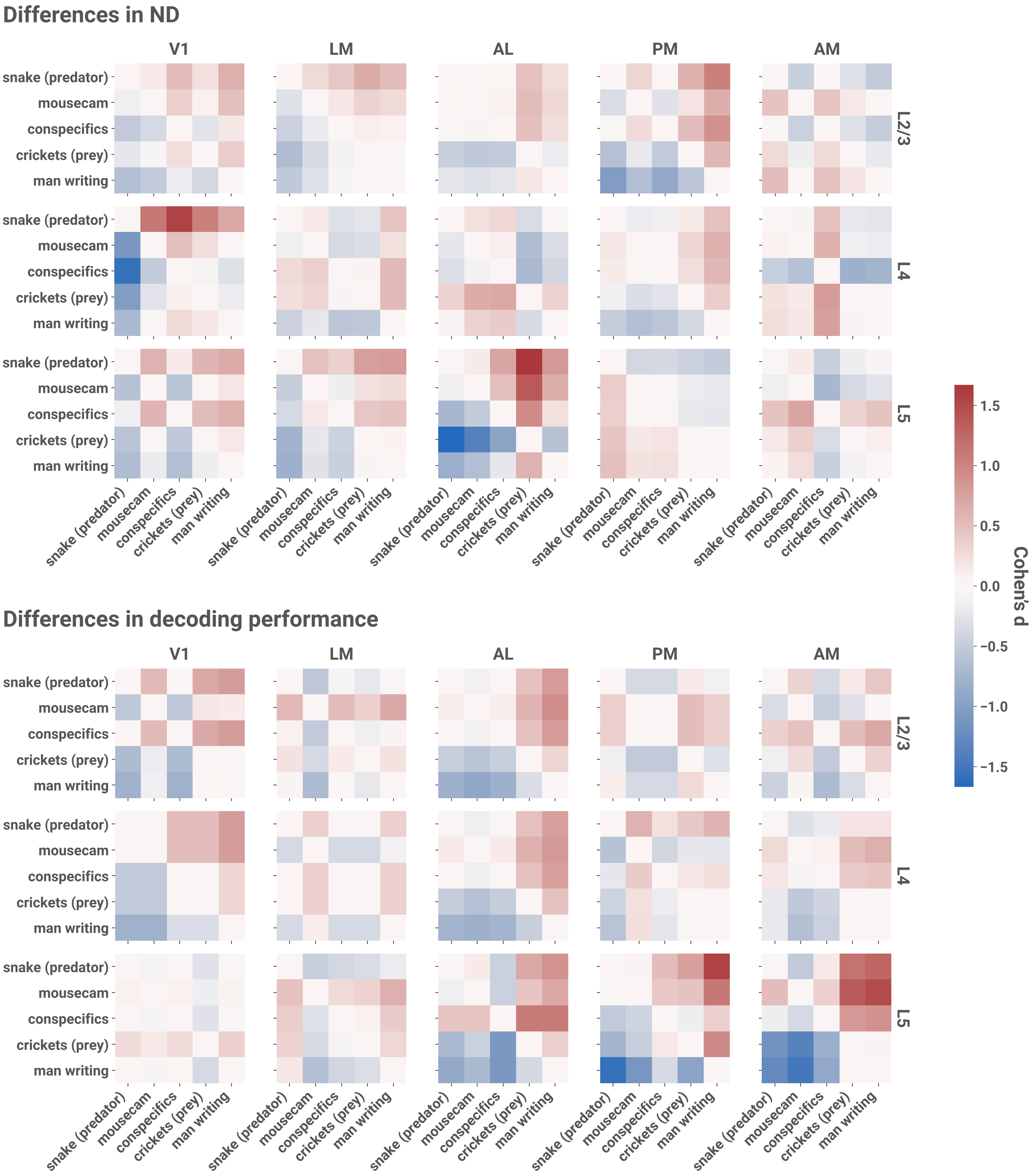


Figure C.15. Within-category differences in ND *vs.* within-category differences in decoding performance, by layer and area. *Top:* Cohen's d for pairwise mean differences in ND among naturalistic stimuli without jump cuts. *Bottom:* Cohen's d for pairwise mean differences in stimulus identity decoding performance. For each session, we trained a linear discriminant analysis classifier using only responses to these 5 stimuli; classification performance was evaluated as the mean fivefold cross-validated F_1 score for each stimulus (§ 4.2, Decoding analyses).

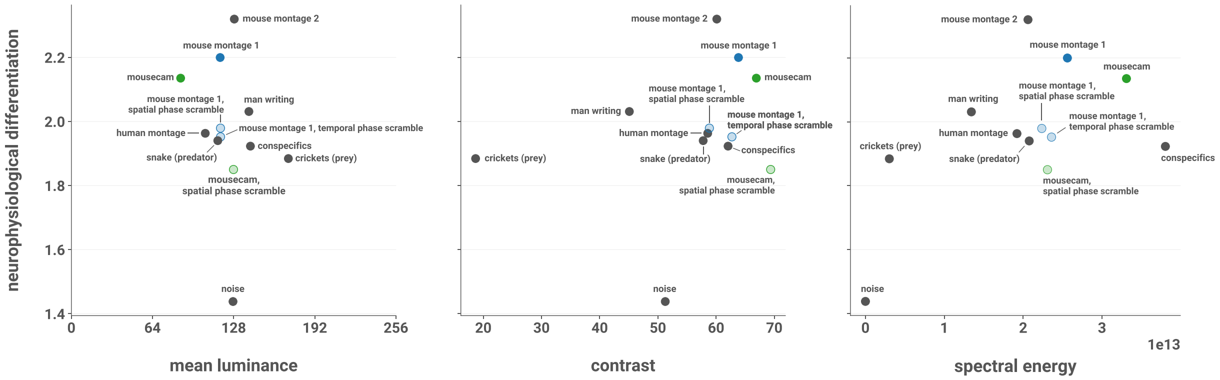


Figure C.16. ND *vs.* low-level stimulus characteristics. ND is plotted against the mean luminance, contrast, and spectral energy of the stimuli. Mean luminance was computed as the average pixel intensity. Contrast was calculated as the standard deviation of pixel intensities. Spectral energy of the blurred stimuli was computed as the sum of the energy spectral density of each pixel's intensity timeseries after removing the DC component.

D Chapter 5 supporting information

D.1 The substrate model and its units

The substrate model used in this work comprises 21 units, organized in four hierarchical levels: a sensory interface of eight units, a lattice level of eight units, a configuration detector level of four units, and a segment level of a single unit. Each level has an associated activation function. Within each level, the units are mechanistically identical, differing only in the identity of their inputs and outputs (Figure 5.2A). The substrate was built using the `substrate_modeler` package available on GitHub at `https://github.com/bjorneju/substrate_modeler.git`. Here, we give a narrative description of the substrate functionality. Further details can be found in the code repositories accompanying the paper (`https://github.com/wmayner/matching.git`).

While the units are inspired by neural mechanisms, they are not models of individual neurons. Rather, they are more analogous to small neural circuits consisting of both inhibitory and excitatory neurons, *i.e.* they are 'macro' units (Hoel et al., 2016; Marshall et al., 2018). Furthermore, the time scale of a single dynamical update in the model is a 'macro' time scale long enough to allow the constituent parts of each unit to adapt to the input before the resulting state is produced and transmitted to the unit's outputs. Thus, the activation functions are state-dependent, both with respect to the state of the unit itself and the state of its inputs. Finally, the final transmitted state of each unit is binary: it is 'ON' if the 'micro' unit corresponding to its output reaches a certain activity threshold (analogous to *e.g.* bursting of primary cells), and 'OFF' otherwise.

Sensory interface units

Each unit in the sensory interface receives a single input from itself, and its state is determined by a sigmoidal activation function. The particulars of the activation function are unimportant, as the sensory interface state is always held constant in our analyses. It is defined explicitly only in order to include the sensory interface units in the TPM of the substrate.

Lattice units

Each unit k in the lattice level receives a single bottom-up input x_k from the sensory interface. In addition, each lattice unit receives inputs I from within the system: one from itself and two from its nearest neighbors within the lattice level (or only one if k is on the boundary, *i.e.*, units A and H), and a single top-down input from the top-level unit M. The activation function is a combination of two sub-functions.

The first function $f_1(x_k, s_k)$ compares the unit's current state s_k with the bottom-up input x_k. If the input state differs from the unit's current state, the unit is driven to flip its state.

The second function implements the *intrinsic connectivity endorsement* (ICE) of the lateral, self, and top-down inputs by the unit k. This is a sigmoid function

$$\sigma_{\text{ICE}}(I) \;=\; 1/(1+\exp\left[-D/(w_{\text{in}}(I)-T)\right]), \tag{D.1}$$

where the determinism $D = 1$, the threshold $T = 0$, and the total input weight $w_{\text{in}}(I)$ is a function of the unit's own state s_k, the state of the inputs $s_j \in I$, and the coupling strength $w_{j,k}$ from unit j to unit k:

$$w_{\text{in}} \;=\; \sum_{s_j \in I} g(s_j, s_k)\, w_{j,k}\, s_j \tag{D.2}$$

Here, an 'OFF' state is counted as -1 and an 'ON' state is counted as 1, similar to the Ising model.

The factor $g(s_j, s_k)$ implements the ICE state-dependency of the coupling strength as follows:

$$g(s_j, s_k) \;=\; \begin{cases} 4 & \text{if } s_j = \text{OFF},\;\; s_k = \text{OFF} \\ -2 & \text{if } s_j = \text{OFF},\;\; s_k = \text{ON} \\ -3 & \text{if } s_j = \text{ON},\;\; s_k = \text{OFF} \\ 6 & \text{if } s_j = \text{ON},\;\; s_k = \text{ON}. \end{cases} \tag{D.3}$$

This function is inspired by the adaptive nature of neural mechanisms, and is analogous to a form of short-term plasticity.

These functions are combined to obtain the probability of activation by taking the 'maximally selective' one, *i.e.* the one that deviates maximally from chance:

$$\Pr(k = \text{ON}) \;=\; \underset{p \,\in\, \{f_1(x_k, s_k),\; \sigma_{\text{ICE}}(w_{\text{in}})\}}{\text{argmax}} \;\; |p - 0.5| \tag{D.4}$$

Configuration detectors

Each unit d in the configuration detector level receives bottom-up input from a unique tuple X of five lattice units (*e.g.*, unit I receives inputs from (A, B, C, D, E)). Like the lattice units, each configuration detector also has a self-connection and receives a top-down input from the top level unit M. The activation function is also a combination of two sub-functions.

The first function activates strongly in response to the segment configuration:

$$f_2(X) \;=\; \begin{cases} 0.99 & \text{if } X = (0, 1, 1, 1, 0) \\ 0.01 & \text{otherwise.} \end{cases} \tag{D.5}$$

Similar to the lattice units, the second function is an ICE sigmoid function of the self-connection and top-down input, $\sigma_{\text{ICE}}(I)$.

The probability of activation is then obtained by combining the functions 'in series':

$$\Pr(d = \text{ON}) \;=\; f_2(X) + (1 - f_2(X))\, \sigma_{\text{ICE}}(I). \tag{D.6}$$

Segment unit

The segment unit at the top of the hierarchy works similarly to the configuration detectors, except that it receives bottom-up input X from the four configuration detectors and intra-level input I from itself. Its bottom-up function f_3 activates strongly when one of the detectors is 'ON':

$$f_3(X) \;=\; \begin{cases} 0.99 & \text{if } \sum\limits_{x \in X} x = 1 \\ 0.01 & \text{otherwise.} \end{cases} \tag{D.7}$$

The activation function is then defined as in (D.6), substituting f_3 in place of f_2.

D.2 Distinctions and relations

Our chief concern in designing the model system was to ensure it had enough recognizable functionality to allow a useful illustration of the formalism. This resulted in a system of 13 units, and the exponential time complexity of the IIT analysis makes exhaustive unfolding of the entire Φ-structure impractical in a system of this size. Consequently, we chose a representative sample of distinctions and relations to compute. We chose mechanisms that feature connectivity motifs of interest based on prior work on how IIT may account for the experience of visual space (Haun & Tononi, 2019) and ongoing work on how IIT may account for the experience of objects (Grasso, in preparation). We further restricted the mechanisms to those whose units are directly connected to one another.

Specifically, we chose to compute distinctions specified by

(1) all subsets of size 1 (first-order mechanisms);
(2) all contiguous subsets of lattice units in level 1;
(3) all subsets of the level 2 configuration detGectors; and
(4) subsets up to and including size 3 that span one or more levels.

Furthermore, for each of the mechanisms, we restricted the set of possible purviews considered in the distinction calculation. We excluded any units not directly connected to the mechanism. We further restricted purviews to those that were likely to yield maximal φ_d for the mechanism, which was determined by consideration of the activation functions involved and a preliminary exhaustive search over all possible purviews for certain example mechanisms from each class in a representative selection of states.

Since the number of relations grows much faster than the number of distinctions, we computed all relations up to and including degree 3 (*i.e.*, relations among up to three distinctions).

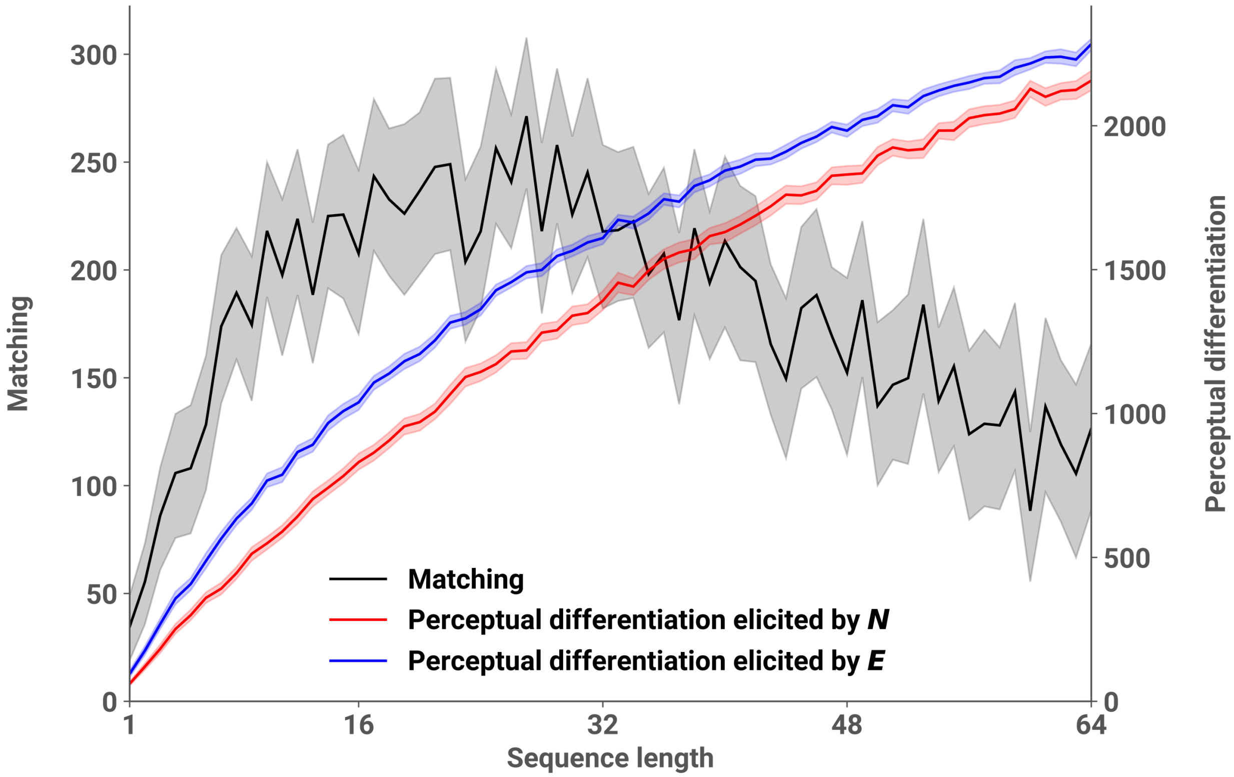


Figure D.1. Matching estimates for different sequence lengths in a 'segment-rich' environment. We computed estimates of matching for stimulus sequences of varying lengths. Sequences were sampled from an environment in which each 'segment' stimulus had probability a and each non-segment stimulus had probability b, where $a = 5b$. For a given sequence length, the matching estimate (black line) was computed as the mean difference between the perceptual differentiation elicited by 100 sequences sampled from the environment (blue line) and that of 100 sequences sampled from independent noise sources (*i.e.*, uniform distribution over stimuli; red line). Shaded regions indicate 95% non-parametric confidence intervals computed from a bootstrapping procedure using 1000 resampling iterations. For this environment, the estimate peaks at sequence length $k = 27$. This is due to several factors: (1) the size of the sensory interface $|\partial S| = 2^8 = 256$; (2) the number of segment stimuli (32) *vs.* non-segment stimuli (224); (3) the relative likelihood of segment *vs.* non-segment stimuli; and (4) the difference in perceptual differentiation for an average segment *vs.* non-segment stimuli. Beyond the peak, matching values decrease as the sequence length becomes large enough that noise samples begin to saturate the system's evoked differentiation capacity (5.18). Note that in practice, this limit (5.24) will be approached extremely slowly, since the uniform probability of a given stimulus decreases exponentially with the size of the sensory interface. In general, properly estimating matching requires a representative and comprehensive sample of the environment; here, for example, the sample sequence must be long enough to include diverse segment stimuli in representative proportions.