Falsification of the Integrated Information Theory of Consciousness

by

Jake R. Hanson

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved March 2021 by the
Graduate Supervisory Committee:

Sara Walker, Chair
Steven Desch
Theodore Pavlic
Christopher Groppi
Sang-Heon Shim

ARIZONA STATE UNIVERSITY

May 2021

ABSTRACT

Astrobiology is premised on the idea that life beyond Earth can exist. Yet, everything known about life is derivative from life on Earth. To understand life beyond Earth, then, requires a definition of life that is abstracted beyond a particular geophysical context. To do this requires a formal understanding of the physical mechanisms by which matter is animated into life. At current, such descriptions are completely lacking for the emergence of life, but do exist for the emergence of consciousness. Namely, contemporary neuroscience offers definitions for universal physical processes that are in one-to-one correspondence with conscious experience. Since consciousness is a sufficient condition for life, these universal definitions of consciousness offer an interesting way forward in terms of the search for life in the cosmos. In this work, I systematically examine Integrated Information Theory (IIT), a well-established theory of consciousness, with the aim of applying it in both biological and astrobiological settings. Surprisingly, I discover major problems with Integrated Information Theory on two fronts: mathematical and epistemological. On the mathematical side, I show how degeneracies buried deep within the theory render it mathematically ill-defined, while on the epistemological side, I prove that the postulates of IIT are scientifically unfalsifiable and inherently metaphysical. Given that IIT is the preeminent theory of consciousness in modern neuroscience, these results have far-reaching implications in this field. In addition, I show that the epistemic issues of falsifiability that hamstring IIT apply quite generally to all contemporary theories of consciousness, which suggests a major reframing of the problem is necessary. The problems that I reveal in regard to defining consciousness offer an important parallel in regard to defining life, as both fields seek to define their topic of study in absence of an existing theoretical framework. To avoid metaphysical problems related to falsifiability, universal theories of both life and consciousness must be framed with respect to independent

empirical observations that can be used to benchmark predictions from the theory. In this regard, I argue that the epistemic debate over scientific theories of consciousness should be used to inform the discussion regarding theoretical definitions of life.

For Nida, next to whom the beauty of science pales in comparison.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my parents for their love and support. It hasn't always been easy, but you have stood by my side and done everything in your power to ensure my well-being. There is no better example of what it means to be loving parents.

Second, I would like to thank my advisor Sara Walker. You have given me the freedom and opportunity to think deeply for five years, which is a gift very few have the pleasure of receiving. Your passion and creativity make our research group a true joy to be a part of. Intellectually, your influence on my ideas cannot be overstated.

Next, I would like to thank my supervisory committee: Chris Groppi, Ted Pavlic, Steve Desch, and Dan Shim. It takes a special group of individuals to supervise a PhD project that starts in astrophysics and ends in cognitive neuroscience but, throughout this process, I have received nothing but unequivocal support and prudent advice from you all. Your support and enthusiasm are a true testament to the interdisciplinary nature of our program and the value of interdisciplinary scientists in general.

To my sister, I am so glad you moved here and I have had the opportunity to build a friendship with you and Stephen. To my in-laws, thank you for the love and generosity you have shown me. You welcomed me with open arms and demonstrate the importance of family every day, and I am honored to be a part of yours. And to my friends - Jack, Dan, Eric, and Dylan - I owe each of you a facet of my personality.

Last, and most importantly, I would like to thank my wife, Nida. Your presence is felt in all that I do, but nowhere is it more readily apparent than my relationship with science. Your love and warmth serve as a crucial counterbalance to the austere realm of theory and abstraction. For this, among countless other things, I am eternally grateful.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF TABLES

Table Page

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Chapter 1

INTRODUCTION

Modern astrophysics provides insight into many existential questions, such as the nature of space and time, the ultimate fate of the universe, and the scale of the cosmos. Yet, the question of whether or not we are alone in the universe remains fundamentally unaddressed. The burgeoning field of exoplanetary science promises insight into this question, as technological advances allow us to infer the atmospheric composition of planets beyond our solar system for the first time ever. From these faint beams of light, we will either detect the presence of life elsewhere in the universe or remain alone indefinitely. The *a priori* balance of outcomes depends entirely on a working understanding of the physical processes by which matter is animated into life.

1.1 Exoplanets

The first confirmed exoplanets were discovered in 1992 by Wolszczan and Frail (Wolszczan and Frail, 1992). At the time, the authors were able to identify periodic variations in the arrival time from a millisecond radio pulsar that were determined to be due to the gravitational influence of at least two terrestrial size planets, with orbital periods of 98 and 67 days respectively. As it turns out, this first method of exoplanet detection is also the rarest, as the solid angle swept out by the beam from a pulsar is relatively small and the supernova process by which they are generated is likely to destroy any primordial planets. Thus, only a handful of exoplanets have been discovered via pulsar timing variations to date (Schneider *et al.*, 2011), but these first two confirmed planets paved the way for future telescope missions devoted entirely

to exoplanet hunting. Of these, the Kepler Space Telescope is undoubtedly the most famous, and responsible in large part for ushering in the modern era of exoplanet science (Borucki *et al.*, 2010, 2011). Launched in 2009, Kepler is devoted entirely to exoplanet hunting via the transit method, which is premised on the idea that as a planet passes in front of its host star the flux from the host star is diminished due to the shadow of the planet on the face of the star which allows the presence of the planet to be inferred. Of course, to confidently make this inference requires that the signal to noise ratio be such that the planetary signal can be unambiguously decoupled from background noise such as starspots on the stellar surface. Thus, a multitude of transits is typically required which biases the transit method towards short-period, large planets (Ananyeva *et al.*, 2020). This bias resulted in the first major surprise of the modern exoplanet era; namely, the ubiquity of "hot Jupiters", which are Jupiter mass planets on extremely close orbits (often on the order of days), for which there is no analog in our solar system. In addition to the discovery of hot Jupiters, "super-Earths" were discovered as a second class of planets with no immediate analog in our solar system (Mayor *et al.*, 2009). Indeed, super-Earths are intermediate-mass planets (2-15 $M_{\oplus}$) that are expected to have either a thick gaseous envelope ("mini-Neptune") or a high abundance of water ("ocean world") (Fu *et al.*, 2009; Adams *et al.*, 2008).

The discovery of water worlds brings into question the difference between a *habitable* planet and an *inhabited* planet. In exoplanet science, the canonical definition of the former is any planet capable of maintaining liquid water on its surface (Lammer *et al.*, 2009). More specifically, the habitable zone refers to the annulus distance surrounding a star within which an Earth analog suffers from neither a runaway greenhouse effect (inner radius) nor frozen water-ice (outer radius) (Kopparapu *et al.*, 2013). Interestingly, the bounds on habitability correspond qualitatively to

Venus and Mars, with the former representing a planet that suffers from a runaway greenhouse effect and the latter representing a planet with frozen water-ice. Thus, the search for habitable worlds is clearly influenced by what it means to be habitable in our solar system, which is a point we will return to in detail later on. Naturally, water worlds satisfy the definition of habitable planets given that their surfaces are primarily liquid water. For example, in the TRAPPIST-1 system (Luger *et al.*, 2017), three of the seven terrestrial exoplanets are in the canonical habitable zone; yet, of these, none are likely to be *inhabited* due to the fact that too much water actually shuts off key geochemical cycles thought to be necessary for life, such as plate tectonics and continental weathering (Unterborn *et al.*, 2018; Kite *et al.*, 2009).

In addition to habitability, *detectability* is another concern at the forefront of modern exoplanet science (Desch *et al.*, 2017; Walker *et al.*, 2018). There are four potential outcomes that can result from an exoplanetary observation: a true positive, a true negative, a false positive, or a false negative. A true positive is the ideal outcome, in which we detect life on an inhabited planet (i.e. life is there and we detect it). However, there is also the potential for an observation to result in a false negative, in which life is present but we fail to detect it. In practice, this situation is no better than a true negative, as failure to detect life on an inhabited planet results in the same practical outcome as failure to detect life on an uninhabited planet - we simply can't tell the difference. Thus, the question of habitability has shifted to one of detectability, as scientists realize the most likely places to detect life are not necessarily the most likely places for life to exist (Unterborn *et al.*, 2018; Glaser *et al.*, 2020). The last of the four outcomes is a false positive, in which a biosignature is detected and falsely attributed to biotic sources when, in reality, it was caused by an abiotic process. This is arguably the deepest issue currently faced by exoplanetary scientists, as an inability to rule out abiotic sources amounts to an

inability to confidently infer the presence of life. To understand this issue further, it helps to investigate the landscape of potential biosignatures, so as to understand the relative confidence we have in each.

1.2 Biosignatures

There are certain biosignatures that, if detected, would result in the unambiguous confirmation of life outside of Earth. Microbial life on Mars, for example, would be an unambiguous confirmation that life on Earth is not alone (so-called "smoking guns"). Yet, if we were to find life on Mars, chances are that either it spread from Earth to Mars or vise versa, in which case we still have only a single example of the emergence of life. Finding life elsewhere in the solar system, such as the moons of Jupiter or Saturn, would likely provide a much stronger constraint on the likelihood of life emerging in the universe, as the chances of a shared origin diminish greatly. This is not to say that finding life on Mars would not be informative but only to emphasize the fact that there is a risk/reward tradeoff when it comes to the search for life outside of Earth. Here, we are solely concerned with the *remote* detection of life outside of our solar system. The reason for this is primarily aesthetic, as the search for life outside our solar system answers a different set of questions than the search for life within our solar system, though the challenges associated with decoding the presence of life from a beam of light are compounded accordingly.

Arguably the only real smoking gun biosignature from an exoplanet would be an unambiguous technosignature - either in the form of decoded communication or the detection of molecules that are the byproduct of a complex technological process (Lin *et al.*, 2014; Stevens *et al.*, 2016; Griffith *et al.*, 2015; Korpela *et al.*, 2015). Yet, the likelihood of this occurring is extremely low due primarily to three different factors. First, intelligent life must overcome many "filters" in order to reach a high enough

level of complexity for interstellar communication (Hanson, 1998). Second, it is unclear whether communication is even possible in the absence of common grounding, as decoding messages requires the use of a mutually agreed-upon code (Brillouin, 2013). And third, technological barriers may limit the ability of advanced civilizations to contact each other (Dyson, 1979). All this to say that the most unambiguous biosignature is also the one that is least likely to occur, which is an inverse correlation that seems to holds quite generally (i.e. the less ambiguous a biosignature is the rarer it is).

In the absence of a clear technosignature, we get into the more plausible realm of biosignatures from unintelligent life, which is what the vast majority of current exoplanet science is focused on. In particular, it is *oxygenic photosynthesis* that the community has established as the most promising biosignature going forward (Schwieterman *et al.*, 2018; Grenfell, 2017; Meadows *et al.*, 2018; Walker *et al.*, 2018). Roadmaps for the future of astrobiology, such as that of Horneck *et al.* (2016), skew heavily towards molecular oxygen (O2) as a biosignature not because it is a smoking gun, but rather, it is the most likely to be detected. Alternatives such as the "all small molecules" program of Seager *et al.* (2016) exist and may very well provide a less ambiguous biosignature than O2, but they are unlikely to be detectable via transmission spectroscopy due to the trace abundances that result from such processes. If one reflects on what biosignatures are technologically feasible with the James Webb Space Telescope (JWST) it is without a doubt oxygenic photosynthesis that shows the most promise. For this reason, there is an abundance of literature surrounding O2 as a biosignature (Meadows, 2017). The emerging consensus is that if one can understand the sources and sinks of abiotic oxygen, then the detection of oxygen in the atmosphere in an abundance that cannot be attributed to abiotic means must be due to the presence of life on a given planet. Thus, there is a considerable amount

of effort currently underway to understand the carbon cycle on earth and beyond, as ultimately this dictates the amount of oxygen gas present in a planetary atmosphere. In particular, coupled tectonic-climate models attempt to understand the interplay between subduction, surface creation, weathering, and climate so that the abiotic rates of gas production can be constrained (Grenfell, 2017; Sleep and Zahnle, 2001; Whipple and Meade, 2004; Roe *et al.*, 2008; Lee *et al.*, 2015). Understanding this problem on Earth is difficult enough, let alone exoplanets; yet, it plays a crucial role in ruling out abiotic sources of oxygen gas.

In the context of the environment, there are a few situations for which molecular oxygen has no known abiotic sources (Meadows *et al.*, 2018). First, the detection of oxygen in combination with methane; and second, the detection of oxygen in absence of carbon monoxide. The reason the former serves as a biosignature is due to the fact that methane and oxygen are in redox disequilibrium (methane is a sink for oxygen), which means the presence of oxygen and methane together implies a high rate of O2 replenishment, likely from a biological source though abiotic sources are not ruled out (Krissansen-Totton *et al.*, 2018). Similarly, the presence of oxygen in absence of carbon monoxide (CO) suggests photochemical generation is not responsible for the abundance of O2, as CO would then be abundant as well. Thus, in the context of other spectral features, abiotic sources of O2 are less likely.

In addition to the detection of molecular oxygen via transmission spectroscopy, another possible biosignature is the so-called "red-edge" (Seager *et al.*, 2005; Pallé *et al.*, 2009; Schwieterman *et al.*, 2015). The red-edge refers to the fact that photosynthetic life on earth uses primarily Chlorophyll *a* as the molecule that harvests sunlight and converts it into energy. This molecule, in turn, has a very particular absorption spectrum - absorbing green wavelengths of light strongly while primarily reflecting the rest. This, in combination with the internal geometry of leaves, results

in a steep increase in reflection at the boundary between visible and infrared radiation (hence the term "red-edge"). In terms of biosignatures, it is possible to imagine detecting this signal in the reflection spectrum of an exoplanet, though this detection is notoriously difficult (Montanes-Rodriguez *et al.*, 2006). Better yet, it is possible to imagine seasonal variations in the reflectance spectrum corresponding to the growth and decay of vegetation (Meadows, 2008), though again a strong biosignature such as this is likely to be incredibly rare and difficult to detect.

1.3 A Bayesian Approach

Given the previous section, it is clear that future biosignatures are likely to be ambiguous rather than unequivocal - begging the question of whether or not an abiotic process is responsible for a false positive. Indeed, there have been numerous "biosignature" detections over the years, with varying degrees of ambiguity. For example Allan Hills meteorite 84001 (ALH 84001), discovered in 1984, was reported to show surface features that were consistent with biological origin (McKay *et al.*, 1996). Naturally, this resulted in an extremely high-profile publication that made world-wide news, prompting former United States president Bill Clinton to make a statement on the discovery (NASA, 1996). However, the scientific community eventually reached the consensus that the surface features in question were not biological in origin (Golden *et al.*, 2001). More recently, another headline-making biosignature was that of the "Alien Megastructure" star KIC 8462852 (also known as Tabby's Star). Discovered by the Kepler Space Telescope, this anomalous star showed occasional dips in brightness on the order of tens of percent (e.g. 20%). At the time, aperiodic dips in brightness of this magnitude had no known astrophysical source, prompting some to argue that the dips could be caused by alien artifacts designed to either harvest starlight for energy or to provide extra living space for inhabitants (Wright

et al., 2015). However, follow-up observations revealed a wavelength dependence to these dips that is indicative of dust, rather than intelligent life (Schaefer *et al.*, 2018). As a third example, there was a recent report of phosphine gas detected in the atmosphere of Venus, for which no known abiotic process could be responsible (Greaves *et al.*, 2020). Unfortunately, this sensational report was short-lived as an error in the data reduction pipeline was found to be responsible for the detection (Greaves *et al.*, 2020; Snellen *et al.*, 2020). In all three of the preceding examples, a biosignature was reported but ultimately concluded to be the result of an abiotic rather than biotic process (i.e. a false positive). The purpose of these examples is to illustrate the logic underlying the search for alien biosignatures; namely, if a biosignature is detected, all known abiotic processes are considered in turn and if none can explain the observation biology is invoked as the solution.

Ideally, the likelihood of a biological process being responsible for a given observation should be weighed against the likelihood of an unknown (or unlikely) abiological source. For this comparison to be done objectively, Bayesian inference must be invoked. Put simply, Bayesian inference is a mathematical formalism that allows one to precisely quantify degrees of belief based on available evidence (Joyce, 2019). In regard to biosignatures, Bayes' Theorem, which is the core of Bayesian inference, can be formulated as follows:

$$P(\text{life}|\text{data}) = \frac{P(\text{data}|\text{life})P(\text{life})}{P(\text{data})} = \frac{P(\text{data}|\text{life})P(\text{life})}{P(\text{data}|\text{no life})P(\text{no life}) + P(\text{data}|\text{life})P(\text{life})}$$

Here, the probability that we have detected life given an observation, $P(\text{life}|\text{data})$, breaks down into a ratio of the likelihood that life generates the data, $P(\text{data}|\text{life})$, and the likelihood of the observation in general, $P(\text{data})$. In practice, the term in the denominator is almost always decomposed into a weighted sum of conditional probabilities which, in this case, represent the likelihood of true and false positives

generating the data: $P(\text{data}|\text{life})$ and $P(\text{data}|\text{no life})$. Thus, the use of Bayes' theorem allows the notion of true and false positives to enter the discussion surrounding biosignatures in a way that is both natural and quantitative. Consequently, all biosignatures can and should be cast in terms of Bayes' theorem in order to formalize their underlying assumptions. For example, the detection of seasonal changes in pigmentation via a reflection spectrum has no known abiotic source, which translates into the mathematical assumption $P(\text{data}|\text{no life}) = 0$. Thus, we have $P(\text{life}|\text{data}) = 1$, quantitatively justifying the claim of a smoking gun. Conversely, in the report of phosphine on Venus, there is a non-negligible probability that the data was due to a source other than life, so $P(\text{data}|\text{no life}) > 0$ which lowers the likelihood of a true positive. Of course, to make this argument precise requires knowledge of all other terms in Bayes' theorem, such as the prior probability of life: $P(\text{life})$.

It is here that the full utility of Bayes' theorem is evident, as it forces us to quantify the assumptions that ultimately result in ambiguity over the interpretation of biosignatures. In particular, the term $P(\text{life})$ represents the prior probability, or initial degree of belief, that there is life elsewhere in the universe. Based on the fact life exists on earth, we know $P(\text{life}) > 0$, but the exact value could be arbitrarily large or astronomically small (Carter, 1983; Spiegel and Turner, 2012; Walker *et al.*, 2018). If the emergence of life is incredibly rare (e.g. $P(\text{life}) = 1\text{e}{-24}$) then it is more likely that an unknown abiotic source is responsible for the observation than life, as the chances of life existing elsewhere in our universe are negligibly small. Conversely, if the emergence of life is relatively common (e.g. $P(\text{life}) = 1\text{e}{-4}$), then the detection of something as simple as O2 will result in a non-negligible probability of a true positive. Unfortunately, there is currently no way of objectively assessing these probabilities due to the fact that the probability of life emerging is entirely unconstrained.

1.4 Defining Life

The inability to constrain the likelihood of life emerging in a given planetary context supervenes on our inability to define life. Earth is the only known example of a planet that has given rise to life, which means we have a single data point (known as the "N=1 problem") from which to draw inferences from (Sterelny and Griffiths, 2012; Smith, 2016). Granted, the diversity of life on earth suggests paths forward in terms of "universal biochemistry" (Mariscal and Fleming, 2018; Kim *et al.*, 2019), but there is no doubt in the scientific community that we lack a mechanistic understanding of the physical conditions that give rise to life. We lack both a theoretical definition of what it means to be alive as well as a formal framework for differentiating between living and non-living systems. Consequently, it can be argued that the current definition of life in the exoplanet community is one of folk-psychology - operating strictly on an informal "know it when you see it" basis (Machery, 2012; Walker *et al.*, 2018).

Unsuccessful attempts to define life have been ongoing for decades. As early as 1943, Nobel laureate Erwin Schrödinger attempted to explicate a working definition of life based on physics, but his conclusion was that life requires "other laws of physics that are hitherto unknown" (Schrödinger, 1992; Walker, 2017). Other notable attempts at defining life over the years included those of Dyson (1999); Hazen (2017); Feinberg and Shapiro (1980); Kamminga (1988); Fleischaker (1990); Koshland (2002). The most popular definition, and that adopted by NASA, is the chemical Darwinism definition of Joyce; namely, that "life is a self-sustaining chemical system capable of Darwinian evolution" (Joyce, 1994). Yet, this definition does little to alleviate the problem faced by the exoplanet community for several reasons. First, it says nothing of the emergence of life and therefore does not constrain $P(\text{life})$; second, it is contingent on life as we know it, thus failing to overcome the N=1 problem; and third, it is

qualitative rather than quantitative, which implies it does not readily translate into the Bayesian formalism. For these reasons, among others, the chemical Darwinism definition of life is not unanimously accepted and is of little operational value (Cleland and Chyba, 2002; Machery, 2012).

The inability to define life despite nearly a century of attempts has lead to the postmodern attempt to understand why life is so difficult to understand (Cleland and Chyba, 2002; Machery, 2012; Trifonov, 2011; Walker *et al.*, 2017b). There is growing support for the idea that the problem with defining life is rooted in the inability to recognize life as a *process* rather than an end result (Walker *et al.*, 2017b). In this paradigm, what we commonly refer to as "life" is recognized as a byproduct of a fundamental dynamical process that coordinates inanimate matter into self-sustaining chemical systems (Dupré, 2017; Smith, 2016). Thus, being alive is not as simple as having a metabolism, for example, as a system without a metabolism such as a screwdriver on Mars cannot be explained without invoking life as a process. The exact physical processes by which matter is concerted into life remain undetermined, but there is growing evidence that what differentiates living from non-living systems is the relationship between *information* and *matter* (Walker and Davies, 2013; Walker *et al.*, 2017a). To quote Walker directly:

> [Treating life as a process] necessitates a re-conceptualization of the origins of life, removing the imposed hard boundary between non-life and life, and recognizing there may exist physical processes that we do not yet understand which are most prominent in living systems but are not necessarily absent elsewhere. One candidate is the physics of information: it is often speculated that information may be a key factor in the origins of life. Just as massive bodies represent ideal example systems to study gravity, life could represent the structures in physical reality where the ef-

> fects of information are most prominent. However, we do not understand how information can structure matter (or precisely what "information" is for that matter), yet it seems apparent this is critical to structuring living systems across the hierarchy of living processes from cells to cities. (Walker *et al.* (2017b))

1.5 The Strong Life-Mind Continuity Thesis

The working hypothesis underlying the current work is that *information processing is what differentiates living and non-living systems*. In particular, both living and non-living systems demonstrate complex collective dynamics but the informational architecture that gives rise to these dynamics is hypothesized to be different (Kim *et al.*, 2021). Oil droplets, for example, are capable of complex dynamics that at least visually imitate the dynamics of biological systems such as ant colonies (Gutierrez *et al.*, 2014). However, ant colonies presumably process more information than oil droplets in the form of long-term spatiotemporal correlations that are critical to the robust implementation of complex computations such as nest site selection (Mallon *et al.*, 2001; Pratt and Sumpter, 2006; Valentini *et al.*, 2020). Similarly, the human brain processes more information than an ant colony as evident by the fact that the space of computations realizable by the human brain is vastly greater than that of an ant colony (Von Neumann and Kurzweil, 2012). Thus, the goal is to quantitatively test the working hypothesis by applying a mathematical measure of information processing to a variety of living and non-living dynamical processes, such as those shown in Figure 1.1.

There are several information-theoretic measures that could potentially be used to quantify the difference between living and non-living dynamical systems. These include active information (Lizier *et al.*, 2012), transfer entropy (Schreiber, 2000;

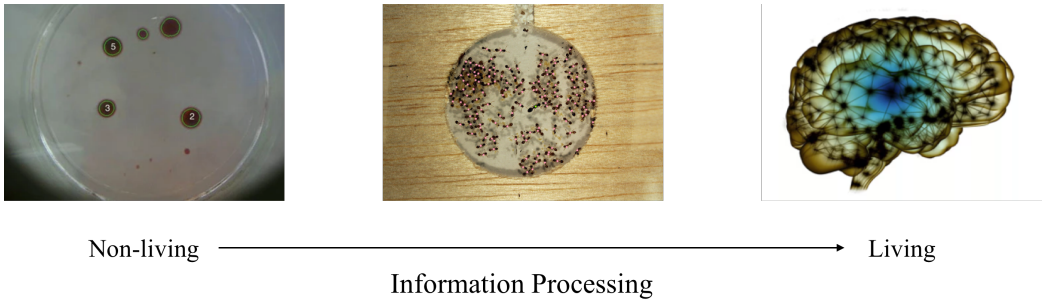


Figure 1.1: The Goal of Our Working Hypothesis Is to Demonstrate That a Quantitative Measure of Information Processing Can Be Used to Differentiate Living and Non-living Systems. Interacting Oil Droplets (Left) Are Hypothesized to Process Less Information than Ant Colonies (Middle) Which, in Turn, Are Hypothesized to Process Less Information than the Human Brain (Right).

Lizier *et al.*, 2008), effective information (Hoel, 2017), excess entropy (Crutchfield and Feldman, 2003), and integrated information (Tononi, 2004; Balduzzi and Tononi, 2008; Oizumi *et al.*, 2014). [1] Of these, integrated information is of particular interest due to the claim that it is both a necessary and sufficient condition for *consciousness*. Of course, a measure of consciousness is not the same as a measure of life, but it is a trivial step to assume consciousness is a sufficient condition for life given the difficulty associated with imagining consciousness in the absence of life. Thus, if integrated information is necessary and sufficient for consciousness, as its proponents claim, it should also be viable as a sufficient condition for life.

In addition to the assumption that consciousness is a sufficient condition for life, there is an alternative reason to focus on a measure of consciousness as opposed to (non-existent) measures of life. Namely, the idea that *cognitive processes are an enrichment of the organizational principles and processes definitive of life*. In its

[1] A full review of these measures in the context of differentiating living and non-living systems can be found in Kim *et al.* (2021)

strongest form, this is known as the "strong life-mind continuity thesis" (Stewart, 1995; Maturana and Varela, 1987; Wheeler, 1997; Kirchhoff and Froese, 2017) and can be outlined as follows:

> The thesis of strong continuity would be true if, for example, the basic concepts needed to understand the organization of life turned out to be self-organization, collective dynamics, circular causal processes, autopoiesis, etc., and if those very same concepts and constructs turned out to be central to a proper scientific understanding of mind. (Clark (2000))

In other words, if strong life-mind continuity holds then a measure of consciousness is one and the same with a measure of life. For our purposes, this suggests there is no need to reinvent the wheel by defining a quantitative measure of life if a quantitative measure of consciousness already exists and is well-established in the cognitive neuroscience community.

1.6 Integrated Information Theory

Integrated Information Theory is the most popular theory of consciousness in contemporary neuroscience (Figure 1.2). In addition to generating hundreds of research papers, it recently received a twenty-five million dollar grant, solidifying its prominence as a top research program for years to come (Reardon, 2019). The theory's popularity is due in large part to three factors. First, it is a "phenomenologically derived" theory of consciousness, meaning that the theory axiomatizes what it is like to be conscious and from these axioms, it derives a mathematical measure. Of course, the word "derives" is used loosely in this context, as the phenomenological axioms must first be translated into physical postulates which are, in turn, translated into a mathematical formalism. Nonetheless, the phenomenology-first approach to

consciousness avoids many epistemic problems that plagued the scientific study of consciousness throughout the twentieth century and is one of IIT's biggest appeals (Negro, 2020). Second, the theory is mathematical in nature, aiming to provide an information-theoretic measure of consciousness - Φ - that can be applied out of the box to any dynamical system (all that is necessary is a transition probability matrix). Thus, for the first time, there is a theory of consciousness that goes beyond ambiguous claims and makes concrete predictions in terms of a real-valued function corresponding to consciousness. This leads to the third factor responsible for IIT's popularity, which is that it is an experimentally falsifiable theory. Predicted Φ values can be compared to experimental results and used to provide evidence for or against the epistemic foundations of the theory. For example, if a conscious human has $\Phi = 0$ the theory is falsified. For these reasons, among others, IIT has seen exponential growth in popularity over the past two decades and recently surpassed global neuronal workspace (Dehaene and Changeux, 2004) as the most popular theory of consciousness to date. [2]

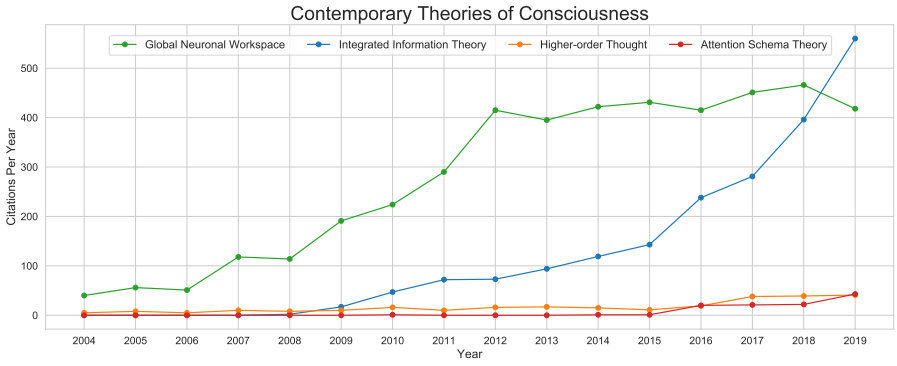


Figure 1.2: Rival Theories of Consciousness in Terms of Citations per Year. IIT Is Currently the Most Popular Theory of Consciousness in Contemporary Neuroscience.

[2]Data is from the SCOPUS database (Burnham, 2006) and includes any citation that uses the name of the theory in the title, abstract, or keywords.

The core idea underlying IIT is, unsurprisingly, integrated information. Measures of integrated information such as Φ quantify the extent to which the whole is more than the sum of its parts (Tegmark, 2016). To do so, one "cuts" the system into parts by preventing information exchange between subcomponents. If the dynamics of the system remain unchanged, then the subcomponents are acting independently and the system is not integrated. Mathematically, this can be formalized by asking whether or not the transition probability matrix M describing the dynamical evolution of a system can be tensor factorized into the product of two (or more) conditionally independent transition probability matrices M_A and M_B. More specifically, if we consider a time-dependent random process $x(t)$ with marginal probability distributions $p(x_0)$ and $p(x_1)$ (corresponding to $x(t_0)$ and $x(t_1)$), we have a Markov process defined by:

$$p(x_1) = Mp(x_0)$$

Similarly, if we tensor factorize the transition probability matrix as $\hat{M} = M_A \otimes M_B$, we have an approximation to this process defined by:

$$\hat{p}(x_1) = \hat{M}p(x_0)$$

Measures of integrated information then quantify the information-theoretic distance D between the marginal probability distribution $p(x_1)$ that results from the natural dynamical evolution and the marginal distribution $\hat{p}(x_1)$ that results under tensor factorization. Minimizing this distance over all possible tensor factorizations yields the Φ value for the system; That is:

$$\Phi = \min D[p(x_1)||\hat{p}(x_1)]$$

There is an intuitive appeal to the notion that consciousness is the extra "stuff" that emerges when a system is considered as a whole rather than its parts. Yet,

this intuition is far from a definition and requires very strong justification if it is to form the axiomatic backbone of a scientific theory. Thus, while the ultimate goal is to apply a measure such as Φ to various living and non-living systems, we must first assess the epistemic foundations of the theory and the arguments justifying its validity.

1.7 Dissertation Context and Layout

Within this context, this dissertation levels a number of serious accusations against the scientific validity of IIT. In Chapter 2, I demonstrate that despite being lauded as a mathematically rigorous theory of consciousness, ambiguity in the definition of Φ renders IIT mathematically ill-defined In Chapter 3, I turn to the epistemic foundations of the theory, demonstrating that the presence or absence of integrated information is ultimately a consequence of arbitrary changes in the binary labels are used to internally represent functional states. This, in turn, proves IIT is either (a) falsified or (b) inherently unfalsifiable. In Chapter 4, I make this argument concrete by building Turing indistinguishable machines with and without $\Phi > 0$, for which only metaphysical justification can be used to explain IIT's predicted difference in subjective experience. In Chapter 5, I prove that the issues with falsification that hamstring IIT apply quite generally to all machine-state functionalist theories of consciousness, a broad class of theories that subsumes IIT. As a whole, this dissertation represents a step towards formally understanding the limits and pitfalls of scientific inquiry as it applies to disciplines lacking a formal definition of their topic of study, such as astrobiology, which is a topic I return to in the conclusion.

Chapter 2

ON THE NON-UNIQUENESS PROBLEM IN INTEGRATED INFORMATION THEORY

2.1 Abstract

Integrated Information Theory is lauded as a mathematically rigorous theory of consciousness due its provision of a scalar mathematical measure of consciousness - Φ - deduced from the phenomenological axioms of the theory. Here, we show that despite its widespread use, Φ is not a well-defined mathematical concept in the sense that the value it specifies is neither unique nor specific. This problem, occasionally referred to as "undetermined qualia", is the result of degeneracies buried deep in the optimization routine used to calculate Φ and completely undermines the validity of the theory. As demonstration, we first apply the mathematical definition of Φ to a simple `AND+OR` logic gate system and show 83 non-unique Φ values result, spanning a substantial portion of the range of possibilities. We then introduce a Python package called `PyPhi-Spectrum` designed to address this issue by calculating the entire spectrum of possible Φ values for any given system and apply it to a host of recently published Φ values. We find that virtually all Φ values in the literature are chosen arbitrarily from a set of non-unique possibilities, often including both conscious and unconscious predictions. Lastly, we review proposed solutions to this problem and find none to be adequate, either because they fail to specify a unique Φ value or yield $\Phi = 0$ for systems that are clearly integrated. We conclude with a philosophy of science discussion, arguing the handling of IIT's core ideas has been unscientific to date.

2.2 Introduction

Integrated Information Theory (IIT) is lauded as a mathematical theory of consciousness due to its provision of a scalar mathematical measure - Φ - that is claimed to predict the overall level of consciousness in virtually any dynamical system. In comparison to other contemporary theories, such as Global Neuronal Workspace Theory (Dehaene and Changeux, 2004) or Predictive Processing (Dolkega and Dewhurst, 2020; Wilkinson *et al.*, 2019; Seth, 2014; Hobson *et al.*, 2014; Hohwy, 2018), the mathematical rigor of IIT is unique and is at least partially responsible for the exponential growth in popularity IIT has experienced over the past decade [1] . It may come as a surprise, then, to learn that a unique Φ value is not guaranteed. This is not to say there is ambiguity in the derivation of Φ from the phenomenological axioms of the theory, which there is (Barrett and Mediano, 2019; Bayne, 2018; McQueen, 2019), but rather, that the accepted mathematical definition of Φ in IIT 3.0 (Oizumi *et al.*, 2014) does not result in a uniquely defined value. Indeed, as we will show, it is possible that IIT simultaneously predicts a system to be both conscious and unconscious which implies the theory is mathematically ill-defined.

As early as 2012, it was clearly known by proponents of IIT that Φ may be "indeterminate" for some systems (Tononi, 2012). Yet, to date, an investigation of the scope of this problem has not been undertaken. How this fundamental flaw came to be overlooked is somewhat of an enigma, as even the simplest of systems can evade a straightforward Φ evaluation. The specifics of this problem result as a consequence of what is occasionally referred to as "underdetermined qualia" or "tied purviews" in the literature (Krohn and Ostwald, 2017; Moon, 2019). Basically, Φ is defined as an information-theoretic distance between two vectors known as cause-effect structures

[1] It is relatively straightforward to track the popularity of rival theories using the SCOPUS database (Burnham, 2006)

(CES) or "constellations". A constellation is comprised of "concepts" which, in turn, are comprised of three things: a "core cause repertoire" (probability distribution), a "core effect repertoire" (probability distribution), and a ϕ^{Max} value (scalar). The core cause and core effect are chosen based on an optimization routine that is designed to select the cause/effect with the highest ϕ^{max} value. However, in practice, equivalent ϕ^{Max} values are ubiquitous - in which case the core cause and core effect are non-unique or "degenerate". Naturally, the constellations are affected by this degeneracy (as they are comprised of the degenerate core cause/effect repertoires) which, in turn, affects the value of Φ. Put simply, cause-effect structures are underspecified and therefore the distance between them (Φ) is non-unique.

Of course, it is possible to select a core cause and core effect repertoire arbitrarily from the set of degenerate values. For example, using an `if less than` statement in the optimization routine will select the first of the degenerate core causes/effects while using an `if less than or equal to` statement will select the last of the degenerate core causes/effects. Thus, it is easy to imagine that numerical routines used to calculate Φ, such as `PyPhi` (Mayner *et al.*, 2018), simply failed to consider this small detail. However, this is not the case. Buried in the configuration files for `PyPhi` is the option to select whether to keep the smallest or largest purview element in the event of a tie (i.e. the same ϕ^{max} value for different causes/effects). Indeed, this is an ad hoc solution that is well known in the small body of literature that covers degenerate core causes and effects (Moon, 2019; Krohn and Ostwald, 2017; Albantakis *et al.*, 2019). Unfortunately, it is equally well known that this is not a valid solution (Moon, 2019), as often the tied purview elements are the same size and the fundamental degeneracy remains unaddressed. [2] In addition, there is nothing

[2] In which case, algorithms such as `PyPhi` default to selecting the first of the degenerate values which depends arbitrarily on the order in which purview elements are considered

phenomenological to suggest why the smallest or largest purview element should be retained as the core cause/effect, which is evident by the fact that different authors reach conflicting conclusions as to whether to select the smallest or largest purview element (Krohn and Ostwald, 2017; Albantakis *et al.*, 2019; Moon, 2019).

Here, we aim to shed light on this issue by attempting to calculate Φ for a very simple model system in the form of an `AND` gate connected to an `OR` gate. Taking the mathematical definition of Φ at face value, we demonstrate that a spectrum of 83 different Φ values results, corresponding to both conscious and unconscious predictions. Next, we provide a modified version of `PyPhi` called `PyPhi-Spectrum` that can be used to calculate the entire spectrum of Φ values for a given dynamical system with a single function call, as opposed to the singular value that is typically reported. We then apply this algorithm to a corpus of ten recently published Φ values, in order to determine the extent to which non-unique Φ values are overlooked in the literature. Last, we investigate whether or not proposed solutions adequately address this problem. We conclude with a philosophy of science discussion related to the scientific handling of ideas.

2.3 Methods

2.3.1 Preliminaries

In IIT 3.0, Φ^{Max} is the overall level of conscious experience that is predicted for a given dynamical system. The calculation of Φ^{Max} is notoriously difficult to perform. In total, five nested optimization steps are required, as shown in Algorithm 1. At the core of this routine is a simple distance measure in the form of an earth mover's distance (Rubner *et al.*, 2000) between two probability distributions. This results in a measure of integration known as ϕ ("little phi"). However, this elementary distance

calculation must be performed for every possible partition of a given "purview" in order to calculate ϕ^{MIP}, then every possible purview for a given "mechanism" in order to find ϕ^{Max} [3] . This results in what is known as a cause-effect structure (CES) or "constellation", which is defined by the set of mechanisms, their ϕ^{Max} values, and two probability distributions per mechanism corresponding to the "core cause" and "core effect". Next, one must generate a constellation for every possible partition of the subsystem under consideration and use a modification of the earth mover's distance to quantify how close this constellation is to that of the unpartitioned subsystem. This results in a second measure of integration known as Φ ("big Phi"), which is designed to quantify the effect of a system-level partition on the underlying ability for a system's components (mechanisms) to integrate information. The system-level partition with the smallest Φ value is the minimum information partition (MIP) and the corresponding Φ value is Φ^{MIP}. Last, this entire process must be repeated for every possible subsystem in a given system in order to find the maximum integrated information Φ^{Max}. In total, this hierarchy of nested optimization routines results in a computational complexity that scales as $\mathcal{O}(13^m)$, where m is the number of elements in the system and is unrealizable in practice for all but the smallest of physical systems (c.f. Appendix A).

2.3.2 Degenerate Core Causes and Effects

To demonstrate the problems inherent in the mathematical definition of Φ^{Max} we will consider a simple system comprised of an `AND` gate and an `OR` gate connected to each another, as shown in Figure 2.1. Since there are only two elements, we need not worry about the outermost optimization as a subsystem must be comprised of at least

[3] Here, we assume the reader is familiar with the basic terms in IIT 3.0. For a more detailed explanation of terms, we refer the reader to the original IIT 3.0 publication (Oizumi *et al.*, 2014).

Algorithm 1: Pseudocode Overview of the Routine to Calculate Phi Max

```
## Calculate Phi_Max
for each subsystem in the powerset of system elements:
    ## Calculate Phi_MIP
    for each unidirectional partition of the subsystem:
        ## Build CES
        for each mechanism in the powerset of the subsystem elements:
            ## Calculate phi_max
            for each element in the past purview:
                ## Calculate phi_cause_max
                for every partition of the purview element:
                    phi = D(partitioned_purview || unpartitioned purview)
                phi_cause_max = max(phi)
            for each element in the future purview:
                ## Calculate phi_effect_max
                for every partition of the purview element:
                    phi = D(partitioned_purview || unpartitioned purview)
                phi_effect_max = max(phi_effect_mip)
            phi_max = min(phi_cause_max,phi_effect_max)
        Phi = D(original_ces || new_ces)
    Phi_MIP = min(Phi)
Phi_Max = max(Phi_MIP)
```

two components in order to generate $\Phi^{MIP} > 0$, so $\Phi^{MIP} = \Phi^{Max}$ in what follows. To calculate Φ^{MIP} we first must initialize the system into a given state. We assume an initial state $s_0 = 00$ in all that follows, though our results are not sensitive to this choice. The next step is to identify the cause-effect structure (CES) or "constellation" C corresponding to the transition probability matrix (TPM) of the unpartitioned system. To do this, one must find the *core cause* and *core effect* of every potential mechanism in the system, where a mechanism is any element in the power set of the subsystem. In our case, the potential mechanisms are in $\mathcal{P}(\{A^cB^c\}) = \{A^c, B^c, AB^c\}$ where the superscript c denotes the mechanism in its current state. For each element in this set, we must identify how well it constrains elements in the past power set $\mathcal{P}(\{A^pB^p\})$, known as the past purview, as well as how well it constrains elements in the future powerset $\mathcal{P}(\{A^fB^f\})$, known as the future purview.

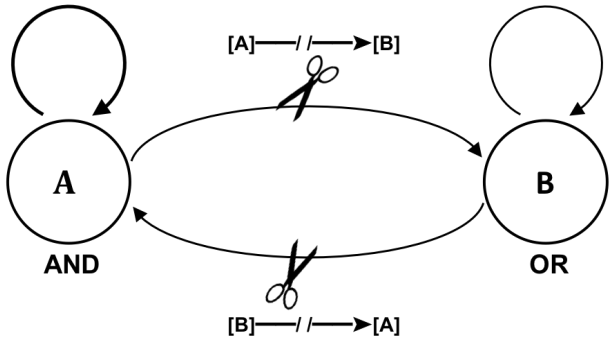


Figure 2.1: A Simple System Comprised of a Fully Connected AND+OR Gate System. Nodes Are Labeled as A and B, Respectively. Partitions Are Found by Cutting the Connection from One Element to the Other in a Unidirectional Fashion.

Next, we measure the earth mover's distance D between the constrained distribution of each purview element and the constrained distribution of each purview element under the minimum information partition (MIP):

$$\phi^{MIP}(m, z) = D[p(z|m = s_0)||p(z|m = s_0/\text{MIP})]$$

where z is the purview element and m is the mechanism. The distribution $p(z|m = s_0)$ tells us the likelihood of z given the current state of m is s_0 which, compared to an unconstrained distribution, tells us how much information m is generating about z. However, we also need to know whether or not that information is "integrated" so we must break m and z up into all possible parts and ask whether the parts acting independently can generate the same amount of information as the whole. For example, to find how much integrated information is generated by the mechanism A^c about the purview element $z = AB^p$ we calculate the probability distribution $p(AB^p|A^c = 0)$ and compare this to the two possible partitions of the purview: $A^c/AB^p \rightarrow (A^c/A^p \times []/B^p)$ and $A^c/AB^p \rightarrow (A^c/B^p \times []/A^p)$. The first partition allows A^c to constrain A^p but leaves B^p unconstrained (denoted by an empty bracket

[]) while the second partition allows A^c to constrain B^p but leaves A^p unconstrained. The distributions generated by these partitions, shown in Figure 2.2, are then compared to the distribution generated by the unpartitioned system, and the partition that minimizes the earth mover's distance to the unpartitioned system is the MIP for this purview/mechanism combination. If multiple partitions yield the same earth mover's distance to the unpartitioned system, as is the case in Figure 2.2, it is irrelevant which one is chosen as all that moves forward in the computation is the scalar value of ϕ^{MIP}.

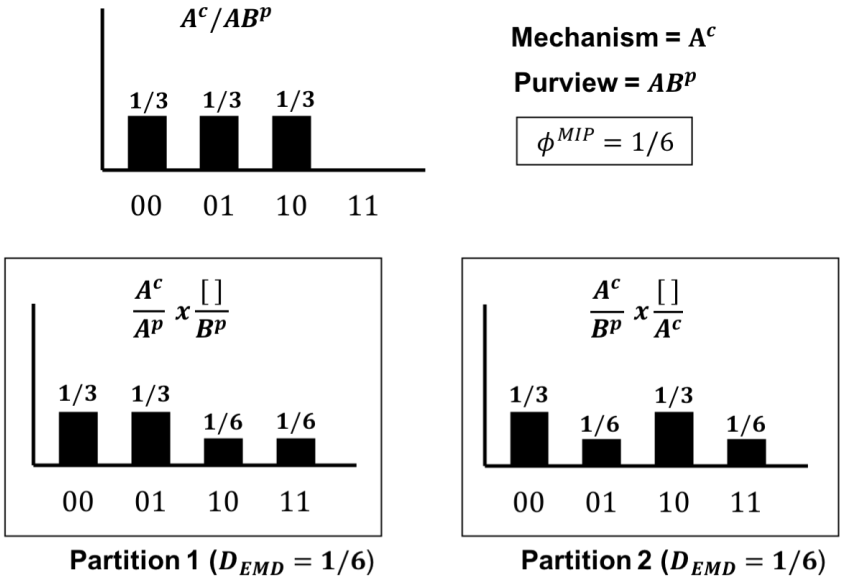


Figure 2.2: All Possible Partitions of a Given Mechanism+Purview Combination and the Resulting ϕ^{MIP} Value.

Once we have identified the MIP (and calculated ϕ^{MIP}) for all purview elements for a given mechanism, we define the core cause and core effect as the past and future purview elements with the greatest ϕ^{MIP}. We denote the integrated information of the former as ϕ^{Max}_{cause} and of the latter as ϕ^{Max}_{effect} and define the total integrated

information ϕ^{Max} of a given mechanism as:

$$\phi^{Max} = \min\left[\phi^{Max}_{cause}, \phi^{Max}_{effect}\right]$$

If $\phi^{Max} > 0$ for a mechanism we say that the mechanism gives rise to a "concept". A concept is fully specified by three things: its ϕ^{max} value, the cause repertoire corresponding to the core cause, and the effect repertoire corresponding to the core effect. We have already provided an example of a cause repertoire in Figure 2.2, namely, it is the distribution over previous state of the purview element given that the current state of the mechanism. In Figure 2.2, the state of A^c constrains the probability of observing AB^p and this constrained distribution is the cause repertoire for the purview element AB^p. Any element not included as part of the purview is left unconstrained and must be independently "noised" (Oizumi *et al.*, 2014). For example, if the purview of mechanism A^c is A^p we generate the constrained distribution $p(A^p|A^c = s_0)$ and combine this with the unconstrained distribution for B^p (denoted $p^{uc}(B^p)$). For the `AND+OR` system, we have $p(A^p|A^c = 0) = [2/3, 1/3]$ and $p^{uc}(B^p) = [1/2, 1/2]$ which yields the cause repertoire: $[1/3, 1/3, 1/6, 1/6]$, where states are ordered in binary with the most significant digit on the left (i.e. $[00, 01, 10, 11]$).

The effect repertoire for a given purview element is generated in the same way as a cause repertoire. For example, the probability of observing A^f given the state of mechanism $A^c = 0$ is $P(A^f|A^c = 0) = [0, 1]$. Combining this with the unconstrained future distribution for B^f yields the effect repertoire $[1/4, 3/4, 0, 0]$. Note, this is an example where the unconstrained future distribution for B^f is not uniform. This is a direct result of the noising procedure as an `OR` gate receiving uniform random input is three times as likely to be in state 1 as it is to be in state 0. Furthermore, when more than one target node is involved, we must send *independent* noise to each target (to avoid correlated input). For example, in a three-node system if we were looking

at the purview element BC^f and noising over the state of A^c, we must imagine that A^c has the ability to send different signals to B^f and C^f at the same time (hence the term "noising").

We can now define the CES or "constellation" C as the set of all concepts for the system in the given state. Recall, each concept corresponds to a single mechanism and is comprised of the mechanism's core cause repertoire, core effect repertoire, and ϕ^{Max} value. In our example, there are at most three concepts, corresponding to the mechanisms $\{A^c, B^c, AB^c\}$. The core cause repertoire for a mechanism is found by optimizing over all past purview elements and identifying the purview with the highest ϕ^{MIP}, where ϕ^{MIP} is found by further optimization over all possible partitions. Figure 2.3 shows ϕ^{MIP} and the corresponding partition for all possible purview elements given the mechanism A^c.

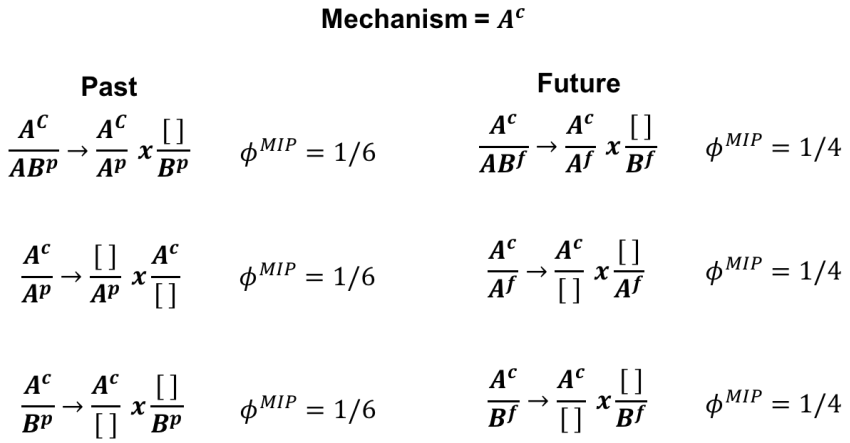


Figure 2.3: All Possible Purview Elements and Their MIPs for a given Mechanism. It Is Here That the Degeneracy Is Introduced, as One Cannot Select a Unique Core Cause or Effect for a given Mechanism If There Are Purview Element with the Same ϕ^{MIP} Values.

At this point, we are faced with a problem. The postulates of IIT (and the exclu-

sion postulate in particular) imply that we must assign a unique core cause to each mechanism, but the purview element that generates ϕ^{Max}_{cause} is not unique. As Figure 2.3 shows, A^p, B^p, and AB^p all generate the same ϕ^{MIP} value for the mechanism in question. Since each purview/mechanism combination is associated with a different cause repertoire, *the core cause repertoire and the resulting constellation C are not well-defined*. If the scalar value of ϕ^{Max}_{cause} was all that mattered to the calculation of Φ, this degeneracy would be inconsequential (as is the case for partitions that generate the same ϕ^{MIP} value for a given purview element). However, system-level integrated information Φ is defined as the cost of transforming the core cause/effect repertoires from one constellation C into another C'. That is:

$$\Phi = D(C||C')$$

where D is an extension of the earth mover's distance that calculates the cost of moving ϕ^{Max} *between repertoires*. If the core cause or effect repertoire changes, the distance between constellations will change accordingly, as the distance metric that goes into the EMD calculation is sensitive to the relative shape of the distributions and not just the scalar ϕ^{Max} values. For example, if we were to choose AB^p as the core cause for mechanism A^c, this generates the concept in C given by the tuple $\{[1/3, 1/3, 1/3, 0], [1/2, 1/2, 0, 0], 1/6\}$ where the first element is the core cause repertoire, the second element is the core effect repertoire, and the third element is the ϕ^{Max} value. However we could just as easily have chosen A^p as our core cause and A^f as our core effect. In which case, the concept generated for A^c would be $\{[1/3, 1/3, 1/6, 1/6], [1/4, 3/4, 0, 0], 1/6\}$. Clearly, these choices have the same Φ^{Max} value but significantly different core cause and effect repertoires.

To illustrate the consequences of this, let C be the constellation consisting only of the concept generated by A^c with core cause AB^p and core effect AB^f and let C' be

the constellation consisting of only the null concept for this system (the unconstrained cause and effect repertoires):

$$C = \{[1/3, 1/3, 1/3, 0], [1/2, 1/2, 0, 0], 1/6\}$$

$$C' = \{[1/4, 1/4, 1/4, 1/4], [3/16, 9/16, 1/16, 3/16], 0\}$$

The extended earth mover's distance is the cost of transforming C into C' by moving $\phi^{Max} = 1/6$ a distance given by the sum of the (regular) earth mover's distance between cause repertoires and effect repertoires. Namely, we have

$$D_{cause} = D_{EMD}([1/3, 1/3, 1/3, 0]||[1/4, 1/4, 1/4, 1/4]) = 1/3$$

$$D_{effect} = D_{EMD}([1/2, 1/2, 0, 0]||[3/16, 9/16, 1/16, 3/16]) = 1/2$$

which results in the integrated conceptual information:

$$\Phi^{MIP} = (D_{cause} + D_{effect})\phi^{Max} = (\frac{1}{3} + \frac{1}{2})\frac{1}{6} = \frac{5}{36}$$

Now, if we instead choose A^p and A^f as our core cause and core effect we have:

$$C = \{[1/3, 1/3, 1/6, 1/6], [1/4, 3/4, 0, 0], 1/6\}$$

$$D_{cause} = D_{EMD}([1/3, 1/3, 1/6, 1/6]||[1/4, 1/4, 1/4, 1/4]) = 1/6$$

$$D_{effect} = D_{EMD}([1/4, 3/4, 0, 0]||[3/16, 9/16, 1/16, 3/16]) = 1/4$$

corresponding to an integrated conceptual information:

$$\Phi^{MIP} = (\frac{1}{6} + \frac{1}{4})\frac{1}{6} = \frac{5}{72}$$

Thus, we get different values of Φ^{MIP} depending on our choice of core cause and core effect.

A Spectrum of Non-unique Φ Values

Each combination of degenerate core cause/effect repertoires results in the potential for a different Φ value. For example, if the unpartitioned system has three degenerate core causes for A^c and two for B^c, then there are $3 \times 2 = 6$ non-unique constellations (CES) for the unpartitioned system. If the partitioned system also has six non-unique CES, then there are a total of 36 different combinations for the distance between constellations (Φ). For the `AND+OR` system, the unpartitioned system has a total of 81 non-unique combinations, while each cut has a total of 9 non-unique constellations. Thus, one must examine the distance between $81 \times 9 \times 2$ different combinations of constellations in order to determine all possible Φ values, which we refer to as the "spectrum" of Φ values for the subsystem. Note, not all Φ values are valid Φ^{MIP} values; it is only those between the upper and lower Φ value of the minimum information partition (MIP) that satisfy the definition of Φ^{MIP}. In total, 83 non-unique Φ^{MIP} values result for the `AND+OR` system, as shown in Figure 2.3.2. Again, we emphasize there is nothing in the axioms or postulates of IIT to suggest which of these 83 values IIT actually predicts, as all are equally valid according to the mathematical definition of Φ. Crucially, both $\Phi^{MIP} = 0$ and $\Phi^{MIP} > 0$ are present in the spectrum, meaning IIT cannot actually predict whether or not this simple system is conscious (or it simultaneously predicts the system is both conscious and unconscious).

`PyPhi-Spectrum`

The Python package `PyPhi` (Mayner *et al.*, 2018) provides all of the basic functionality needed to calculate the spectrum of Φ values for a given system. Namely, it allows the user to calculate purviews, cause/effect repertoires, earth mover's distance,

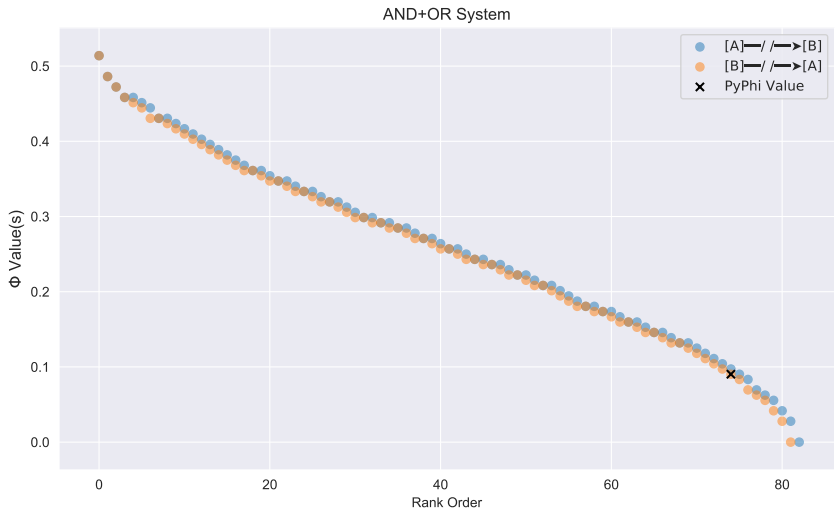


Figure 2.4: The Spectrum of Φ Values That Result for Each Cut of the `AND`+OR Model System. In Total, There Are 83 Different Values, All of Which Are Equally Valid Φ^{MIP} Values According to the Mathematical Definition of Φ^{MIP}. The Single Value Corresponding to the Output From the Python Package `PyPhi` is Shown As a Black 'x'.

CES distance (extended earth mover's distance), and more. In addition, it contains prebuilt classes for data structures such as concepts that are useful in the calculation. Here, we wrap this basic functionality into a modified version of `PyPhi` called `PyPhi-Spectrum` that allows the user to calculate all Φ values for a given subsystem with a single function call. To install this package, one can simply download or clone the entire `Phi-Spectrum` repository (which includes core `PyPhi` functionality) from https://github.com/jakehanson/pyphi-spectrum. An overview of the wrapper, as well as basic usage, can be found in Appendix C.

2.4 Results

We now apply our methodology to a host of recently published Φ values, in order to determine the extent to which degenerate core causes/effects undermine the publication of a unique Φ value. We begin with a pedagogical case study, followed by the broad application of the `PyPhi-Spectrum` package to as large of a corpus as is computationally feasible.

2.4.1 Case Study: Three-node Fission Yeast Cell Cycle

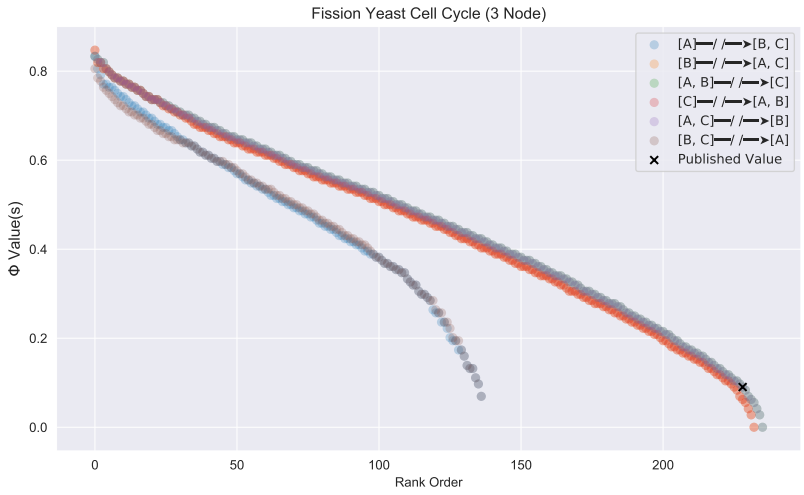
As a case study, we consider the Boolean network model of the fission yeast cell cycle from Marshall *et al.* (2017b). In this study, IIT is used to analyze the causal structure of a minimal biological system, namely, the cell cycle of the fission yeast *S. pombe*. Using Φ, the authors identify three integrated subsystems corresponding to the full system (eight nodes), a six node subsystem, and a three-node subsystem - all potentially of biological importance. Of these three systems, only the smallest may be subject to our analysis, though we expect similar results for the other two systems. Applying the methodology from Section 2.3, we find a spectrum of *244 non-unique* Φ *values*, spanning a range from $0.00 - 0.83$ bits (Figure 2.5). Crucially, only one of these values ($\Phi = 0.09$) is published as the unique Φ^{MIP} value for this subsystem. In reality, all of these values are equally valid according to the mathematical definition of Φ. Furthermore, the inclusion of $\Phi = 0$ in the spectrum of possibilities changes the biological interpretation of the results entirely. If the subsystem under consideration has $\Phi^{MIP} = 0$, rather than $\Phi^{MIP} > 0$, it would not be identified as "integrated" and its biological function would not be deemed of interest. Thus, the conclusion that this subsystem is of biological importance is entirely dependent on the arbitrary selection of a single Φ value from the spectrum of possibilities and, in general, it is impossible

to tell *a priori* whether the narrative being built around a particular Φ value is valid without studying the entire spectrum of possibilities.

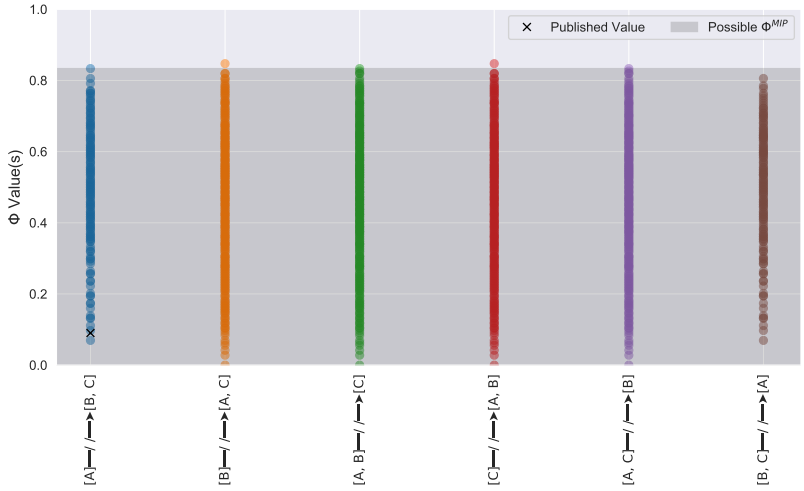
2.4.2 The Non-uniqueness of Published Φ Values

Next, we use `PyPhi-Spectrum` to analyze a corpus of recently published Φ values, with the goal of understanding the extent to which underdetermined qualia render published Φ values ill-defined. The corpus we analyze is intended to be comprehensive, but we are limited by the computational resources required to perform these calculations (see Appendix A). Of the dozen or so Φ values that are published in the literature Haun and Tononi (2019); Albantakis *et al.* (2019); Albantakis and Tononi (2019); Juel *et al.* (2019); Popiel *et al.* (2020); Hanson and Walker (2019, 2020); Sevenius Nilsen *et al.* (2019); Aguilera *et al.* (2018); Marshall *et al.* (2017b); Farnsworth (2021); Niizato *et al.* (2020); Tononi *et al.* (2016); Hoel *et al.* (2016); Chalmers and McQueen (2014), only a handful are small enough to be subjected to our analysis (Albantakis *et al.*, 2019; Hanson and Walker, 2020; Farnsworth, 2021; Marshall *et al.*, 2017b; Oizumi *et al.*, 2014; Tononi *et al.*, 2016; Hoel *et al.*, 2016; Chalmers and McQueen, 2014). These systems, summarized in Table 2.1, are selected primarily for their size, though size alone is not a good indication of computational tractability. For example, certain three-node systems, such as that found in Hanson and Walker (2020), have thousands of degenerate cause-effect structures while others, such as that found in Farnsworth (2021), have just a few. It is not readily apparent what dictates the number of non-unique cause-effect structures that result for a given system, though symmetric inputs almost certainly play a role (see Section 2.5). Consequently, our corpus is limited to small systems (2-4 nodes) which happen to allow relatively fast evaluation via `PyPhi-Spectrum` [4] . While improvements such as parallelization

[4]Less than a day on a Mac Pro; 3.5 GHz 6-Core Intel Xeon E5 Processor



(a)



(b)

Figure 2.5: The Spectrum of Φ^{MIP} Values That Result From Degenerate Core Causes/Effects in the Three-node Fission Yeast System Analyzed by Marshall *et al.* (2017b). The Published Value for this System is Shown As a Black 'x'. Figure (a) Shows the Φ Values for Each Cut in Rank Order, While (b) Shows the Φ Values for Each Cut Relative to the Upper and Lower Bound on Φ^{MIP}. All Φ Values Between the Min and Max Φ Value of the MIP Are Equally Valid Φ^{MIP} Values.

could certainly be made to improve performance and increase the size of our corpus, we do not believe doing so would add much to the interpretation of our results.

Name	Description	Size	Φ Value
`AND+OR`	System from Section 2.3	2	0.0903
Farnsworth 2021 (Full)	Virus-Host Dynamics	5	0.3125
Farnsworth 2021 (Reduced)	Simplified Virus-Host Dynamics	3	0.4375
Gomez et al. 2020	p53-Mdm2 Regulatory Network	4	0.2153
Photodiode	`COPY+COPY`	2	1.0000
Marshall et al. 2017	Fission Yeast Cell Cycle	3	0.0903
Hoel et al. 2016	`AND+AND+AND+AND`	4	0.1139
Tononi et al. 2016	`MAJORITY+OR+AND+AND`	4	0.6597
Oizumi et al. 2014	`OR+AND+XOR`	3	1.9167

Table 2.1: Summary of Corpus in Reverse Chronological Order. Sources Were Selected Based on the Publication of a Unique Φ^{MIP} Value and Computational Tractability. Additional Details Required for Analysis, Such as Transition Probability Matrices and Initial States Are Provided in Appendix D.

Our primary result is shown in Figure 2.6. Namely, it is the calculation of the entire spectrum of Φ^{MIP} values relative to the published value for every text in our corpus. There are several things to note. First, the existence of a unique Φ value is rare, as only the photodiode has a spectrum consisting of a single value (the number of different Φ^{MIP} values is denoted by $|\Phi^{MIP}|$ in Figure 2.6). For the rest of the corpus, the spectra often consist of dozens if not hundreds of non-unique Φ^{MIP} values, of which only one is published (denoted as a black "x" in Figure 2.6). In addition, it is entirely possible for a spectrum to contain both $\Phi = 0$ and $\Phi > 0$ values, which

implies IIT is not a well-defined mathematical theory. This occurs for three out of the ten published Φ values in our corpus: `AND+OR`, Marshall *et al.* (2017b), and Hoel *et al.* (2016). In such cases, IIT does not fail to predict whether or not a system is conscious so much as it simultaneously predicts a system to be both conscious and unconscious, which is a much stronger indictment of the logical foundations of the theory (the exclusion postulate in particular). Last, we would like to point out that the span of Φ^{MIP} values is often comparable to the entire range of possibilities that one would expect for systems of this size. According to Figure 2.6, a typical Φ spectrum spans roughly 1/2 of a bit. In comparison, a (deterministic) two-node Boolean system is bounded from above by $\Phi^{MIP} \leq 1.5$ bits (Appendix B), which implies that the Φ values calculated by IIT are not only non-*unique* but also non-*specific* (i.e. they don't constrain the possible Φ values to a small portion of the range).

2.5 Existing Solutions

The problem of degenerate core causes/effects in IIT is understudied but not entirely unknown (Oizumi *et al.*, 2014; Krohn and Ostwald, 2017; Moon, 2019; Albantakis *et al.*, 2019). To our knowledge, there are four different solutions to this problem, with differing degrees of justification. The first solution is that put forward by Oizumi *et al.* (2014) in the original IIT 3.0 publication. In Figure S1 of their Supporting Information, the authors argue that the degenerate core cause corresponding to the *biggest purview* element should be selected as the core cause, with the justification being that the larger purview "specifies information about more system elements for the same value of irreducibility". By this, what is meant is that the ϕ^{max} values of the degenerate purview elements are the same (same value of irreducibility) but bigger purview elements s constrain more of the system (e.g. AB^p constrains more of the system than A^p). Conveniently, the degenerate core causes in the example they

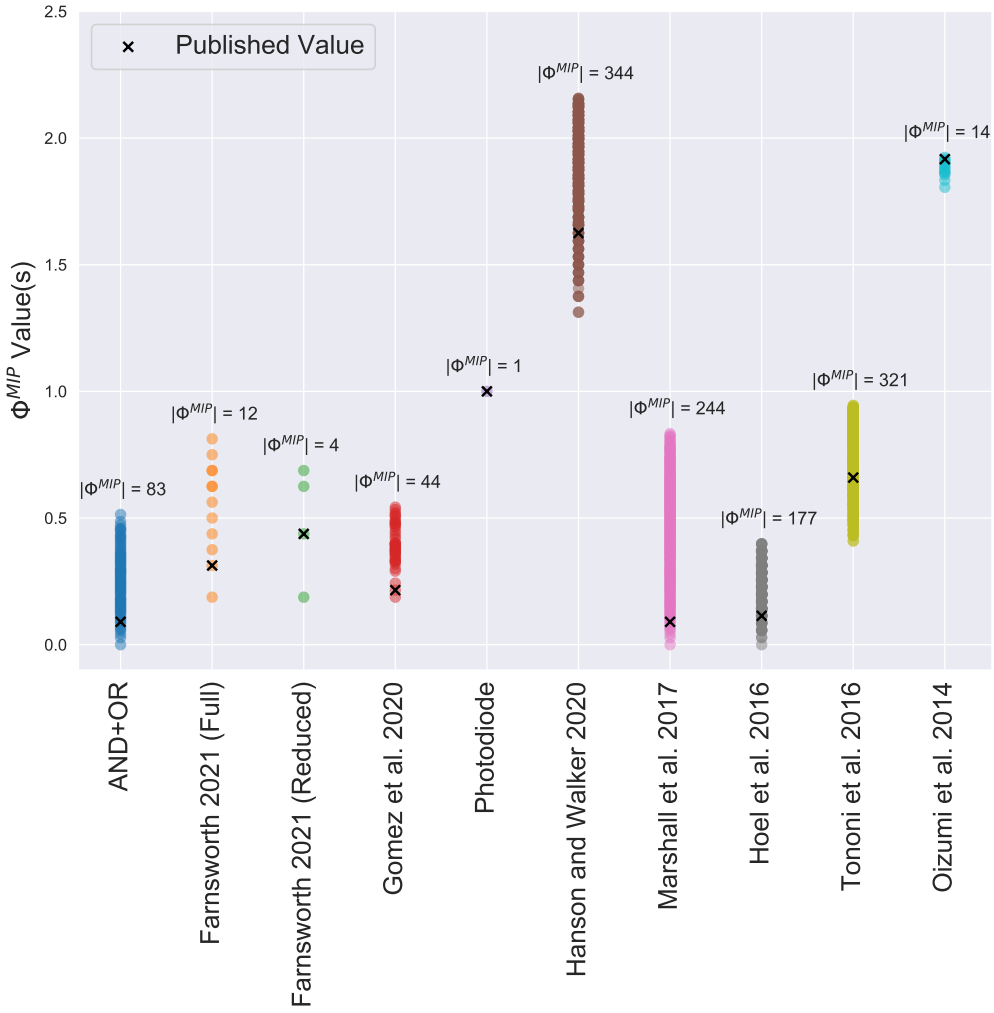


Figure 2.6: Possible Φ^{MIP} Values Relative to the Published Value for the Corpus Shown in Table 2.1. Each Point Represents a Possible Φ^{MIP} Value That Results From the Mathematical Definition of Φ. The Number of Different Φ Values for Each System Is given by the Cardinality of Its Spectrum $|\Phi^{MIP}|$.

consider correspond to purview elements of different sizes, which allows this simple criterion to result in a unique core cause; in general, there is no guarantee this is the case and, therefore, this criterion cannot be used to guarantee a unique Φ value. In addition, it is unclear why the selection of bigger purview elements is more in line with

the axioms and/or postulates of the theory. Constraining more or less of the system does not have a clear interpretation in terms of the phenomenology from which Φ is purportedly derived, which implies additional postulates are required (Moon, 2019). This lack of phenomenological grounding is further evident by the fact that Krohn and Ostwald (hereafter KO17) reach the exact opposite criterion as a proposed solution to the same problem (Krohn and Ostwald, 2017); namely, KO17 argues that it is the *smallest purview* element that should be selected as the core cause/effect in the case of tied purviews, based on the idea that "causes should not be multiplied beyond necessity" (Tononi, 2012). While this verbiage is technically part of IIT's exclusion postulate, it is not clear that it can be applied to the dimensionality of purview elements - is choosing B^p as the core cause of A^c instead of AB^p really multiplying the cause of A^c beyond necessity? The answer is unclear; however, what is clear is that the smallest purview element is still not guaranteed to be unique which means Φ is still not mathematically well-defined. To address this, KO17 presents a completely novel solution in the form of a modified definition of Φ based on the difference in the sum of ϕ^{max} values between constellations, rather than the extended earth mover's distance. The obvious benefit of this definition is that it depends only on the scalar ϕ^{max} values associated with concepts, rather than the non-unique cause/effect distributions. In other words, it does not matter which of the degenerate core causes is selected because they all have the same ϕ^{Max} value and the integrated conceptual information is just the sum of ϕ^{Max} values over concepts. The fourth and final solution is the "differences that make a difference" criterion proposed by Moon (2019). Here, the author argues that if degenerate core causes/effects exist then *none* of the corresponding purview elements should be selected as the core cause/effect (i.e. $\phi^{max} = 0$ for the mechanism). This solution is based on the idea that in IIT "to exist is to cause differences". If tied purview elements exist with the same ϕ^{Max} value, then the ϕ^{Max}

value does not change if one of the tied purview elements is excluded. For example, if A^p and B^p both give rise to a concept with $\phi^{Max} = 1/6$, one can eliminate B^p without changing the ϕ^{Max} value for the mechanism; therefore, the existence of B^p does not make a difference "from the intrinsic perspective of the system". The fact that one can do this individually for each of the degenerate core causes or effects implies that none can give rise to a concept and $\phi^{Max} = 0$ for the mechanism.

Unfortunately, there are major problems with the Φ values that result from all four proposed solutions to the non-uniqueness problem. Namely, selecting the smallest or largest purview element does not guarantee a unique Φ value; and the KO17 and Moon criteria result in $\Phi = 0$ for systems that are clearly integrated. In the case of the KO17 definition of Φ, under a partition, the sum of ϕ^{Max} values can actually *increase*, resulting in the confounding conclusion that the system is integrating more information after a cut than it was before ($\Phi < 0$). These so-called "magic cuts" are discussed by Krohn and Ostwald (2017), as well as in the PyPhi documentation, but they do not bode well for a measure of integration. More immediately, if we apply the KO17 definition of Φ to the `AND+OR` system from Section 2.3, we find $\Phi = 0$ due to the fact that the ϕ^{Max} values for each concept are the same before and after the system level minimum information partition ($\sum \phi^{Max} = 5/12$ in both cases). Consequently, we reject this modified definition of Φ outright based on the idea that an `AND+OR` system *is* integrated. This notion is supported by the general mathematical definition of integrated information, which is an inability to tensor factorize a system without changing the underlying dynamics (Oizumi *et al.*, 2016b; Tegmark, 2016). In the case of the `AND+OR` system, one cannot cut A from B without affecting the state of B, and vise versa, which implies the system is integrated. The same conclusion applies to Moon's criterion, for which the `AND+OR` system again fails to yield $\Phi > 0$ due to the presence of degenerate core causes/effects for all mechanisms. In addition, the

"differences that make a difference" argument relies entirely on the assumption that the ϕ value being measured is an accurate reflection of what it means to make a difference. Put simply, cutting information from an `AND` gate to an `OR` gate *does make a difference* in terms of system-level integration, and the only way to justify that it doesn't is to define differences that make a difference in terms of ϕ values.

As demonstration of the problems inherent with each of the four existing solutions, we reanalyze our corpus enforcing each criterion in turn (Figure 2.7) [5] . As expected, the KO17 and Moon solution yield $\Phi = 0$ for several systems that are clearly integrated and the "smallest" criterion does little to mitigate the degeneracy. At first glance, however, it appears the "biggest" criterion avoids both of these issues and provides a positive Φ value in all cases. Unfortunately, this is nothing more than an idiosyncrasy of our data set, as selecting the biggest purview element suffers from the same problem as selecting the smallest purview element; namely, degenerate core causes/effects are often the same size. The reason that the "biggest" solution appears to yield unique Φ values for the systems under consideration is due entirely to the ubiquitous use of two-input logic gates (`AND`, `OR`,`XOR`, `NOR`, etc.) in our corpus. In such cases, the tied purview elements are almost always A, B and AB for which selecting the largest purview element (AB) results in a unique core cause/effect. However, this does not hold in general, as systems comprised of logic gates with more than two inputs (e.g. neurons in the human brain) have entirely different symmetries. As Figure 2.8 shows, a simple system of majority gates, each with three inputs, is enough to prove that a unique Φ value does not always result from the "biggest" criterion. Thus, the fundamental problem remains unaddressed.

[5]Implementation of the "Smallest","Biggest", and "Moon" solutions are available via keyword arguments in the `PyPhi-Spectrum` package (see Appendix C), while the KO17 solution is available by changing the `USE-SMALL-PHI-FOR-CES` option in the standard `PyPhi` configuration file.

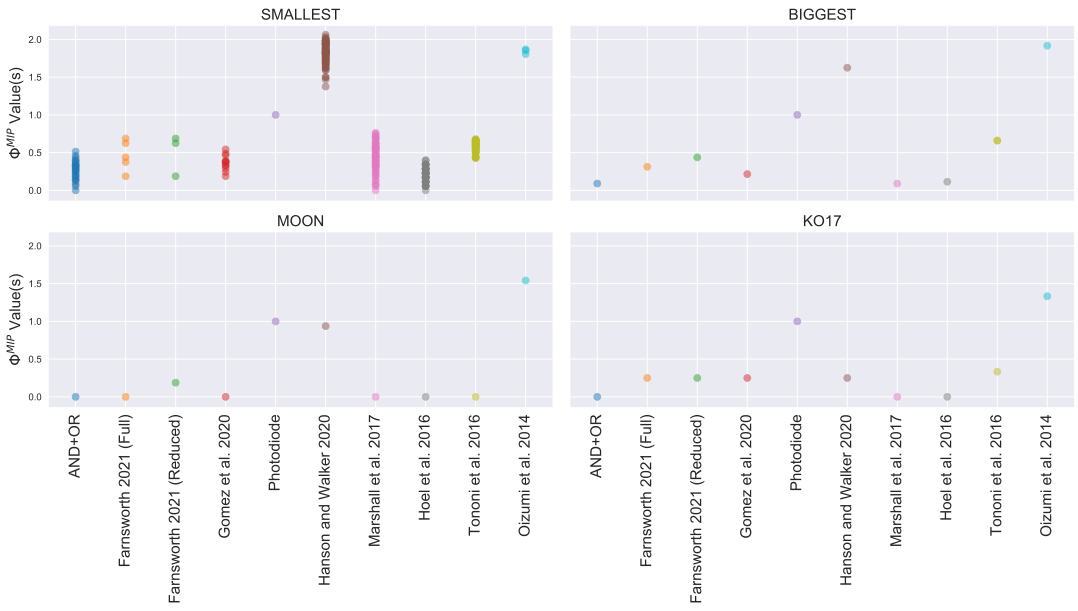
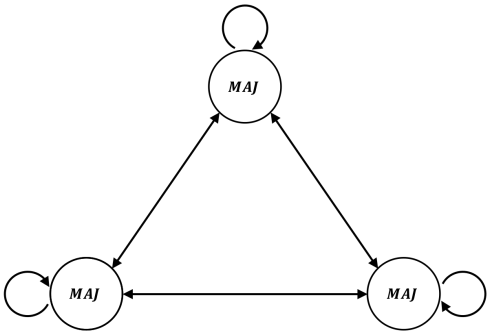


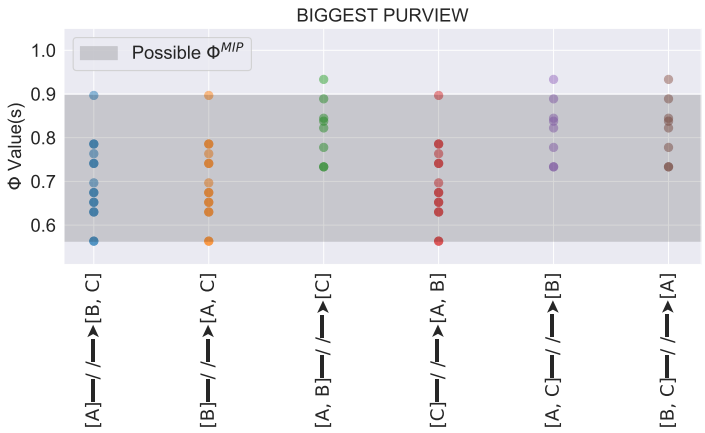
Figure 2.7: The Spectrum of Φ Values That Results From Each of the Four Proposed Solutions. The "Smallest" and "Biggest" Solutions Do Not Guarantee a Unique Φ Value, While the "Moon" and "KO17" Solutions Result in $\Phi = 0$ For Systems That Are Clearly Integrated, Such as the `AND+OR` Gate.

2.6 Discussion

In IIT, the constellation corresponding to the unpartitioned CES is equated with nothing less than subjective experience itself. The geometric shape of this constellation (meaning the shape of the probability distributions that comprise the core cause/effect repertoires) is identified as "what it is like" to be something (Nagel, 1974) while the Φ^{Max} value is identified as the overall "level" of consciousness. Our results prove that this structure is unequivocally underspecified and thus, by IIT's own description, Φ cannot be used to measure the contents or quantity of subjective experience. In the words of Moon, "the qualia underdetermination problem shakes IIT to the ground" (Moon, 2019).



(a)



(b)

Figure 2.8: For a System of Three Fully Connected Majority Gates (a) Selecting the Largest Purview Element Does Little to Mitigate Degeneracy, as Evident by the Range of Possible Φ^{MIP} Values (b).

And yet, IIT is more popular than ever and Φ is being used to bolster arguments in neuroscience and beyond (Reardon, 2019; Farnsworth, 2021; Niizato *et al.*, 2020). From a philosophy of science perspective, this is disconcerting, as IIT demonstrates several hallmark features of an unscientific handling of the core ideas of a theory (Godfrey-Smith, 2009). For one, the use of ad hoc procedures to resolve contradictions that are deduced from the axioms of a theory is never a good sign; that IIT tries to resolve the undetermined qualia problem with the ad hoc procedure of selecting the

smallest or largest purview element is a textbook example of this. In addition, the widespread use of a black box algorithm based on faulty assumptions is a recipe for disaster, especially given IIT's ultimate aims of serving in medical, moral, and legal settings. In fact, we were unable to find a single example of a standard Φ application that does not use `PyPhi` other than those found in the original IIT 3.0 publication (Oizumi *et al.*, 2014). This, in combination with formal proofs that imply IIT is impossible to falsify experimentally (Doerig *et al.*, 2020; Kleiner and Hoel, 2020; Hanson and Walker, 2020), seems to put the theory on the wrong side of the demarcation problem (i.e. what separates science from pseudoscience), despite its widespread use.

Granted, there is hope that IIT may yet provide a blueprint for a successful scientific theory of consciousness (Negro, 2020; Kleiner, 2020). But, to date, it is difficult to overlook the fact that IIT has repeatedly pushed forward before resolving contradictions inherent in its logical foundation (Cerullo, 2015; Doerig *et al.*, 2019; Moon, 2019). On one hand, this could be seen as a positive, as Feyerabend argues that all new theories must proceed counterinductively for a matter of time if they are to succeed - going against well-established principles and holding onto assumptions in the face of overwhelming evidence to the contrary (Feyerabend, 1993). Yet, even Feyerabend is careful to distinguish between scientific and unscientific handling of ideas, stating:

> The distinction between the crank and the respectable thinker lies in the research that is done once a certain point of view is adopted. The crank usually is content with defending the point of view in its original, undeveloped, metaphysical form, and he is not prepared to test its usefulness in all those cases which seem to favor the opponent, or even admit that there exists a problem. It is this further investigation, the details of it,

> the knowledge of the difficulties, of the general state of knowledge, the recognition of objections, which distinguishes the "respectable thinker" from the crank. The original content of his theory does not. (Feyerabend (1981))

Going forward, it is of paramount importance that proponents of IIT seek to address contradictions in the theory head-on, rather than bypass them, as the scientific merit of the theory ultimately depends on it.

2.7 Acknowledgements

The authors would like to thank Cole Mathis, Yanbo Zhang, and Dylan Gagler for early feedback on this manuscript and thoughtful discussions.

Chapter 3

INTEGRATED INFORMATION THEORY AND ISOMORPHIC FEED-FORWARD PHILOSOPHICAL ZOMBIES

3.1 Abstract

Any theory amenable to scientific inquiry must have testable consequences. This minimal criterion is uniquely challenging for the study of consciousness, as we do not know if it is possible to confirm via observation from the outside whether or not a physical system knows what it feels like to have an inside - a challenge referred to as the hard problem of consciousness. To arrive at a theory of consciousness, the hard problem has motivated the development of phenomenological approaches that adopt assumptions of what properties consciousness has based on first-hand experience and, from these, derive the physical processes that give rise to these properties. A leading theory adopting this approach is Integrated Information Theory (IIT), which assumes our subjective experience is a "unified whole", subsequently yielding a requirement for physical feedback as a necessary condition for consciousness. Here, we develop a mathematical framework to assess the validity of this assumption by testing it in the context of *isomorphic* physical systems with and without feedback. The isomorphism allows us to isolate changes in Φ without affecting the size or functionality of the original system. Indeed, the only mathematical difference between a "conscious" system with $\Phi > 0$ and an isomorphic "philosophical zombie" with $\Phi = 0$ is a permutation of the binary labels used to internally represent functional states. This implies Φ is sensitive to functionally arbitrary aspects of a particular labeling scheme, with no clear justification in terms of phenomenological differences. In light of this, we argue

any quantitative theory of consciousness, including IIT, should be invariant under isomorphisms if it is to avoid the existence of isomorphic philosophical zombies and the epistemological problems they pose.

3.2 Introduction

The scientific study of consciousness walks a fine line between physics and metaphysics. On the one hand, there are observable consequences to what we intuitively describe as consciousness. Sleep, for example, is an outward behavior that is uncontroversially associated with a lower overall level of consciousness. Similarly, scientists can decipher what is intrinsically experienced when humans are conscious via verbal reports or other outward signs of awareness. By studying the physiology of the brain during these specific behaviors, scientists can study "neuronal correlates of consciousness" (NCCs), which point to where in the brain conscious experience is generated and what physiological processes correlate with it (Rees *et al.*, 2002). On the other hand, NCCs cannot be used to explain *why* we are conscious or to predict whether or not another system demonstrating similar properties to NCCs is conscious. Indeed, NCCs can only tell us the physiological processes that correlate with what are assumed to be the functional consequences of consciousness and, in principle, may not actually correspond to a measurement of what it is like to have subjective experience (Chalmers, 1995). In other words, we can objectively measure behaviors we assume accurately reflect consciousness but, currently, there exist no scientific tools permitting testing our assumptions. As a result, we struggle to differentiate whether a system is truly conscious or is instead simply going through the motions and giving outward signs of, or even actively reporting, an internal experience that does not exist (Searle, 1980).

This is the "hard problem" of consciousness (Chalmers, 1995) and it is what differ-

entiates the study of consciousness from all other scientific endeavors. Since consciousness is subjective (by definition), there is no objective way to prove whether or not a system experiences it readily accessible to science. Addressing the hard problem, therefore, necessitates an inversion of the approach underlying NCCs: rather than starting with observables and deducing consciousness, one must start with consciousness and deduce observables. This has motivated theorists to develop phenomenological approaches that adopt rigorous assumptions of what properties consciousness must include based on human experience, and, from these, "derive" the physical processes that give rise to these properties. The benefit to this approach is not that the hard-problem is avoided, but rather, that the solution appears self-evident given the phenomenological axioms of the theory. In practice, translating from phenomenology to physics is rarely obvious, but the approach remains promising.

The phenomenological approach to addressing the hard problem of consciousness is exemplified in Integrated Information Theory (IIT) (Tononi, 2008; Oizumi *et al.*, 2014), a leading theory of consciousness. Indeed, IIT is a leading contender in modern neuroscience precisely because it takes a phenomenological approach and offers a well-motivated solution to the hard problem of consciousness (Tononi *et al.*, 2016). Three phenomenological axioms form the backbone of IIT: information, integration, and exclusion. The first, *information*, states that by taking on only one of the many possibilities a conscious experience generates information (in the Shannon sense, *e.g.* via a reduction in uncertainty (Shannon, 1948)). The second, *integration*, states each conscious experience is a single "unified whole". And the third, *exclusion*, states conscious experience is exclusive in that each component in a system can take part in at most one conscious experience at a time (simultaneous overlapping experiences are forbidden). Given these three phenomenological axioms, IIT derives a mathematical measure of integrated information - Φ - that is designed to quantify the extent to

which a system is conscious based on the logical architecture (i.e. the "wiring") underlying its internal dynamics.

In constructing Φ as a phenomenologically-derived measure of consciousness, IIT must assume a connection between its phenomenological axioms and the physical processes that embody those axioms. It is important to emphasize that this assumption is nothing less than a proposed solution to the hard problem of consciousness, as it connects subjective experience (axiomatized as integration, information, and exclusion) and objective (measurable) properties of a physical system. As such, it is possible for one to accept the phenomenological axioms of the theory without accepting Φ as the correct quantification of these axioms and, indeed IIT has undergone several revisions in an attempt to better reflect the phenomenological axioms in the proposed construction of Φ (Tononi, 2004; Balduzzi and Tononi, 2008; Oizumi *et al.*, 2014). Experimental falsification of IIT or any other similarly constructed theory is a matter of sufficiently violating our intuitive understanding of what a measure of consciousness should predict in a given situation which, outside of a few clear cases (e.g. that humans are conscious while awake), varies across individuals and adds a level of subjectivity to assessing the validity of measures derived based on phenomenology. Given that IIT is a phenomenological theory, it is therefore only natural that the bulk of epistemic justification for the theory comes in the form of carefully constructed logical arguments, rather than directly from empirical observation (Godfrey-Smith, 2009). For this reason, it is extremely important to isolate and understand the logical assumptions that underlie any potentially controversial deductions that come from the theory, as this plays an important role in assessing the foundations of the theory.

Here we focus on a particularly controversial aspect of IIT, namely, the fact that philosophical zombies are permitted by the theory. By definition, a philosophical

zombie is an unconscious system capable of perfectly emulating the outward behavior of a conscious system. In addition to being epistemologically problematic (Marcus, 2004; Kirk, 2003), such systems are thought to indicate problems with the logical foundations of any theory that admits them (Harnad, 1995), as it is difficult to imagine how a difference in subjective experience can be scientifically justified without any apparent difference in the outward functionality of the system (Turing, 1950). In IIT, philosophical zombies arise as a direct consequence of IIT's proposed translation of the integration axiom. IIT assumes the subjective experience of a unified whole (the integration axiom) requires feedback in the physical substrate that gives rise to consciousness as a necessary (but not sufficient) condition. This implies any strictly feed-forward logical architecture has $\Phi = 0$ and is unconscious by default, despite the fact that the logical architecture of an "integrated" system with $\Phi > 0$ can always be unfolded (Doerig *et al.*, 2019) or decomposed (Krohn and Rhodes, 1965; Zeiger, 1967a) into a system with $\Phi = 0$ without affecting the outward behavior of the system.

In what follows, we demonstrate the existence of a fundamentally new type of feed-forward philosophical zombie, namely, one that is *isomorphic* to its conscious counterpart in its state-transition diagram. To do so, we implement techniques based on Krohn-Rhodes decomposition from automata theory to isomorphically decompose a system with $\Phi > 0$ onto a feed-forward system with $\Phi = 0$. The result is a feed-forward philosophical zombie capable of perfectly emulating the behavior of its conscious counterpart *without increasing the size of the original system.* Given the strong mathematical equivalence between isomorphic systems, our framework suggests the presence or absence of feedback is not associated with observable differences in function or other properties such as efficiency or autonomy. Our formalism translates into a proposed mathematical criterion that any observationally verifiable measure of consciousness should be invariant under physical isomorphisms. That is, we suggest

conscious systems should form an equivalence class of physical implementations with structurally equivalent state-transition diagrams. Enforcement of this criterion serves as a necessary, but not sufficient, condition for any theory of consciousness to be free from philosophical zombies and the epistemological problems they pose.

3.3 Methods

Our methodology is based on automata theory (Hopcroft *et al.*, 2006; Ginzburg, 2014), where the concept of philosophical zombies has a natural interpretation in terms of "emulation" (Egri-Nagy and Nehaniv, 2015). The goal of our methodology is to demonstrate that it is possible to isomorphically emulate an integrated finite-state automaton ($\Phi > 0$) with a feed-forward finite-state automaton ($\Phi = 0$) using techniques closely related to the Krohn-Rhodes theorem (Krohn and Rhodes, 1965; Zeiger, 1967b).

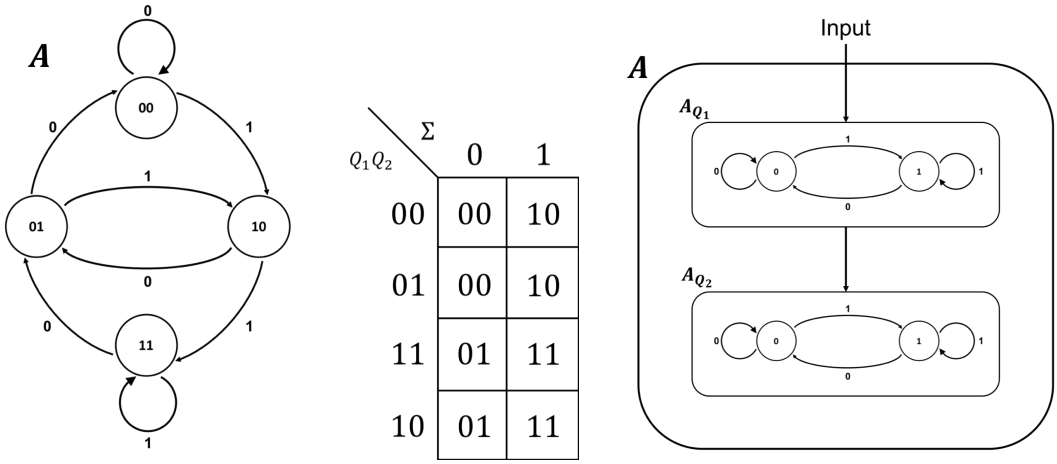
3.3.1 Finite-State Automata

Finite-state automata are abstract computing devices, or "machines", designed to model a discrete system as it transitions between states. Automata theory was invented to address biological and psychological problems (Zeiger, 1968; Shannon and McCarthy, 2016) and it remains an extremely intuitive choice for modeling neuronal systems. This is because one can define an automaton in terms of how specific abstract inputs lead to changes within a system. Namely, if we have a set of potential inputs Σ and a set of internal states Q, we define an automaton A in terms of the tuple $A = (\Sigma, Q, \delta, q_0)$ where $\delta : \Sigma \times Q \to Q$ is a map from the current state and input symbol to the next state, and $q_0 \in Q$ is the starting state of the system. To simplify notation, we write $\delta(s, q) = q'$ to denote the transition from q to q' upon receiving the input symbol $s \in \Sigma$.

For example, consider the "right-shift automaton" A shown in Figure 3.1. This automaton is designed to model a system with a two-bit internal register that processes new elements from the input alphabet $\Sigma = \{0, 1\}$ by shifting the bits in the register to the right and appending the new element on the left (DeDeo, 2011). The global state of the machine is the combined state of the left and right register, so $Q = \{00, 01, 10, 11\}$ and the transition function δ specifies how this global state changes in response to each input, as shown in Figure 3.1b.

In addition to the global state transitions, each individual bit in the register of the right-shift automaton is itself an automaton. In other words, the global functionality of the system is nothing more than the combined output from a system of interconnected automata, each specifying the state of a single component or "coordinate" of the system. Specifically, the right-shift automaton is comprised of an automaton A_{Q1} responsible for the left bit of register and an automaton A_{Q2} responsible for the right bit of the register. By definition, A_{Q_1} copies the input from the environment and A_{Q_2} copies the state of A_{Q1}. Thus, $\Sigma_{Q_1} = \{0, 1\}$ and $\Sigma_{Q_2} = Q_1 = \{0, 1\}$ and the transition functions for the coordinates are $\delta_{Q_1} = \delta_{Q_2} = \{\delta(0, 0) = 0; \delta(0, 1) = 0; \delta(1, 0) = 1; \delta(1, 1) = 1\}$. This fine-grained view of the right-shift automaton specifies its *logical architecture* and is shown in Figure 3.1c. The logical architecture of the system is the "circuitry" that underlies its behavior and, as such, is often specified explicitly in terms of logic gates, with the implicit understanding that each logic gate is a component automaton.

It is important to note that not all automata require multiple input symbols and it is common to find examples of automata with a single-letter input alphabet. In fact, any deterministic state-transition diagram can be represented in this way, with a single input letter signaling the passage of time. In this case, the states of the automaton are the states of the system, the input alphabet is the passage of time,



(a) (b) (c)

Figure 3.1: The Right-shift Automaton A in Terms of Its State-transition Diagram (a), Transition Function δ (b), and Logical Architecture (c).

and the transition function δ is given by the transition probability matrix (TPM) for the system. Because Φ is a mathematical measure that takes a TPM as input, this specialized case provides a concrete link between IIT and automata theory. Non-deterministic TPMs can also be described in terms of finite-state automata (Maler, 1995; DeDeo, 2011) but, for our purposes, this generalization is not necessary.

3.3.2 Cascade Decomposition

The idea of decomposability is central to both IIT and automata theory. As Tegmark (2016) points out, mathematical measures of integrated information, including Φ, quantify the inability to decompose a transition probability matrix M into two independent processes M_A and M_B. Given a distribution over initial states p, if we approximate M by the tensor factorization $\hat{M} \approx M_A \otimes M_B$, then Φ, in general, quantifies an information-theoretic distance D between the regular dynamics Mp and the dynamics under the partitioned approximation $\hat{M}p$ (i.e. $\Phi = D(Mp||\hat{M}p)$). In the latest

version of IIT (Oizumi *et al.*, 2014), only *unidirectional* partitions are implemented (information can flow in one direction across the partition) which mathematically enforces the assumption that feedback is a necessary condition for consciousness.

Decomposition in automata theory, on the other hand, has historically been an engineering problem. The goal is to decompose an automaton A into an automaton A' which is made of simpler physical components than A and maps *homomorphically* onto A. Here, we define a homomorphism h as a map from the states, stimuli, and transitions of A' onto the states, stimuli, and transitions of A such that for every state and stimulus in A' the results obtained by the following two methods are equivalent (Zeiger, 1968):

1. Use the stimulus of A' to update the state of A' then map the resulting state onto A.

2. Map the stimulus of $A'$and the state of A' to the corresponding stimulus/state in A then update the state of A using the stimulus of A.

In other words, the map h is a homomorphism if it *commutes* with the dynamics of the system. The two operations (listed above) that must commute are shown schematically in Figure 3.2. If the homomorphism h is bijective then it is also an *isomorphism* and the two automata necessarily have the same number of states.

From an engineering perspective, homomorphic/isomorphic logical architectures are useful because they allow flexibility when choosing a logical architecture to implement a given computation (i.e. the homomorphic system can perfectly emulate the original). Mathematically, the difference between homomorphic automata is the internal labeling scheme used to encode the states/stimuli of the global finite-state machine, which specifies the behavior of the system. Thus, the homomorphism h is a dictionary that translates between different representations of the same computation.

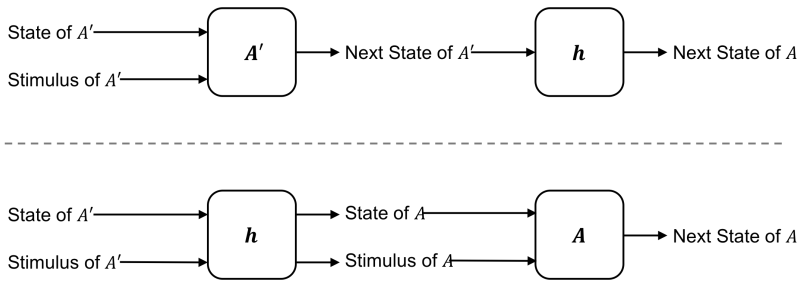


Figure 3.2: For the Map h to be a Homomorphism From A' Onto A, Updating the Dynamics Then Applying h (Top) Must Yield the Same State of A As Applying h Then Updating the Dynamics (Bottom).

Just as the same sentence can be spoken in different languages, the same computation can be instantiated using different encodings. Under this view, what gives a computational state meaning is not its binary representation (label) but rather its causal relationship with other global states/stimuli, which is what the homomorphism preserves.

Because we are interested in isolating the role of feedback, the specific type of decomposition we seek is a feed-forward or *cascade decomposition* of the logical architecture of a given system. Cascade decomposition takes the automaton A and decomposes it into a homomorphic automaton A' comprised of several elementary automata "cascaded together". By this, what is meant is that the output from one component serves as the input to another such that the flow of information in the system is strictly unidirectional (Figure 3.3). The resulting logical architecture is said to be in "cascade" or "hierarchical" form and is functionally identical to the original system (i.e. it realizes the same global finite-state machine).

At this point, the connection between IIT and cascade decomposition is readily apparent: if an automaton with feedback allows a homomorphic cascade decomposition, then the behavior of the resulting system can emulate the original but utilizes

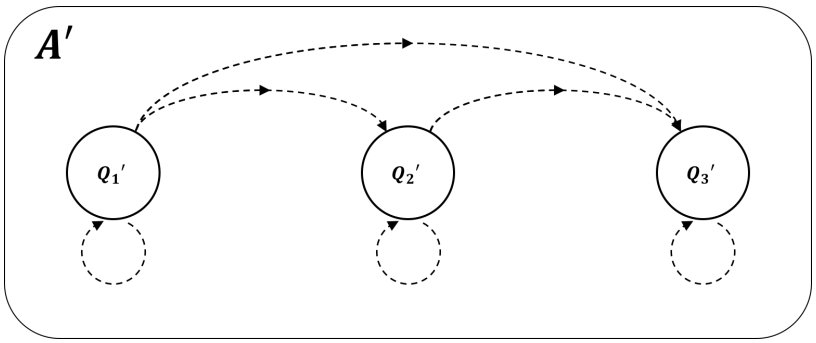


Figure 3.3: An Example of a Fully Connected Three Component System in Cascade Form. Any Subset of the Connections Drawn Above Meets the Criteria for Cascade Form Because All Information Flows Unidirectionally.

only feed-forward connections. Therefore, there exists a unidirectional partition of the system that leaves the dynamics of the new system (i.e. the transition probability matrix) unchanged such that $\Phi = 0$ for all states.

In the language of Oizumi *et al.* (2014), we can prove this by letting $C_{\rightarrow}$ be the constellation that is generated as a result of any unidirectional partition and C be the original constellation. Because $C_{\rightarrow}$ has no effect on the TPM, we are guaranteed that $C_{\rightarrow} = C$ and $\Phi^{MIP} = D(C|C_{\rightarrow}) = 0$. We can repeat this process for every possible subsystem within a given system and, since the flow of information is always unidirectional, $\Phi^{MIP} = 0$ for all subsets so $\Phi^{Max} = 0$. Thus, $\Phi = 0$ for all states and subsystems of a cascade automaton.

Pertinently, the Krohn-Rhodes theorem proves that every automaton can be decomposed into cascade form (Krohn and Rhodes, 1965; Zeiger, 1967a), which implies *every system for which we can measure non-zero* Φ *allows a feed-forward decomposition with* $\Phi = 0$. These feed-forward systems are "philosophical zombies" in the sense that they lack subjective experience according to IIT (i.e. $\Phi = 0$), but they nonetheless perfectly emulate the behavior of conscious systems. Yet, the Krohn-

Rhodes theorem does not tell us *how* to construct such systems. Furthermore, the map between systems is only guaranteed to be homomorphic (many-to-one) which allows for the possibility that Φ is picking up on other properties (e.g. such as the efficiency and/or autonomy of the computation) in addition to the presence or absence of feedback (Oizumi *et al.*, 2014).

To isolate what Φ is measuring, we must go one step further and insist that the decomposition is isomorphic (one-to-one) such that the original and zombie systems can be considered to perform the same computation (Egri-Nagy and Nehaniv, 2015) (same global state-transition topology) under the same resource constraints. In this case, the feed-forward system has the *exact same number of states* as its counterpart with feedback. Provided the latter has $\Phi > 0$, this implies Φ is not a measure of the efficiency of a given computation, as both systems require the same amount of memory. This is not to say that feedback and Φ do not *correlate* with efficiency because, in general, they do (Albantakis *et al.*, 2014). For certain computations, however, the presence of feedback is not associated with increased efficiency but only increased interdependence among elements.

It is these specific corner cases that are most beneficial if one wants to assess the validity of the theory, as they allow one to understand whether or not feedback is important in absence of the benefits typically associated with its presence. In other words, IIT's translation of the integration axiom is that *feedback* is a minimal criterion for the subjective experience of a unified whole; yet, Φ is described as quantifying "the amount of information generated by a complex of elements, above and beyond the information generated by its parts" (Tononi, 2008), which seems to imply feedback enables something "extra" feed-forward systems cannot reproduce. An isomorphic feed-forward decomposition allows us to carefully track the mathematical changes that destroy this additional information, in a way that lets us preserve the efficiency

and functionality of the original system. This, in turn, provides the clearest possible case to assess whether or not this additional information is likely to correspond to a phenomenological difference between systems.

3.3.3 Feed-forward Isomorphisms via Preserved Partitions

The special type of computation that allows an isomorphic feed-forward decomposition is one in which the global state-transition diagram is amenable to decomposition via a nested sequence of preserved partitions. A preserved partition is a way of partitioning the state space of a system into blocks of states (macrostates) that transition together. Namely, a partition P is preserved if it breaks the state space S into a set of blocks $\{B_1, B_2, ..., B_N\}$ such that every state within each block transitions to a state within the same block (Hartmanis, 1966; Zeiger, 1968). If we denote the state-transition function $f : S \to S$, then a block B_i is preserved when:

$$\exists j \in \{1, 2, ..., N\} \text{ such that } f(x) \in B_j \forall x \in B_i$$

In other words, for B_i to be preserved, $\forall x$ in B_i x must transition to some state in a single block B_j ($i = j$ is allowed). Conversely, B_i is *not* preserved if there exist two or more states in B_i that transition to different blocks (i.e. $\exists\ x_1, x_2 \in B_i$ such that $f(x_1) = B_j$ and $f(x_2) = B_k$ with $j \neq k$). In order for the entire partition P_i to be preserved, each block within the partition must be preserved.

For an isomorphic cascade decomposition to exist, we must be able to iteratively construct a hierarchy or "nested sequence" of preserved partitions such that each partition P_i evenly splits the partition P_{i-1} above it in half, leading to a more and more refined description of the system. For a system with 2^n states where n is the number of binary components in the original system, this implies that we need to find exactly n nested preserved partitions, each of which then maps onto a unique

component of the cascade automaton, as demonstrated in Section 3.3.3.

If one cannot find a preserved partition made of disjoint blocks or the blocks of a given partition do not evenly split the blocks of the partition above it in half, then the system in question does not allow an isomorphic feed-forward decomposition. It will, however, still allow a *homomorphic* feed-forward decomposition based on a nested sequence of preserved *covers*, which forms the basis of standard Krohn-Rhodes decomposition techniques (Zeiger, 1968; Egri-Nagy and Nehaniv, 2008, 2015). Unfortunately, there does not appear to be a way to tell *a priori* whether or not a given computation will ultimately allow an isomorphic feed-forward decomposition, although a high degree of symmetry in the global state-transition diagram is certainly a requirement.

Example: `AND/OR` ≅ `COPY/OR`

As an example, we will isomorphically decompose the feedback system X, comprised of an `AND` gate and an `OR` gate, shown in Figure 3.4a. As it stands, X is not in cascade form because information flows bidirectionally between the components Q_1 and Q_2. While this feedback alone is insufficient to guarantee $\Phi > 0$, one can readily check that X does indeed have $\Phi > 0$ for all possible states (Mayner *et al.*, 2018). The global state-transition diagram for the system X is shown in Figure 3.4c. Note, we have purposefully left off the binary labels that X uses to instantiate these computational states, as the goal is to relabel them in a way that results in a different (feed-forward) instantiation of the same underlying computation. In general, one typically starts from the computation and derives a single logical architecture but, here, we must start and end with fixed (isomorphic) logical architectures - passing through the underlying computation in between. The general form of the feed-forward logical architecture X' that we seek is shown in Figure 3.4b.

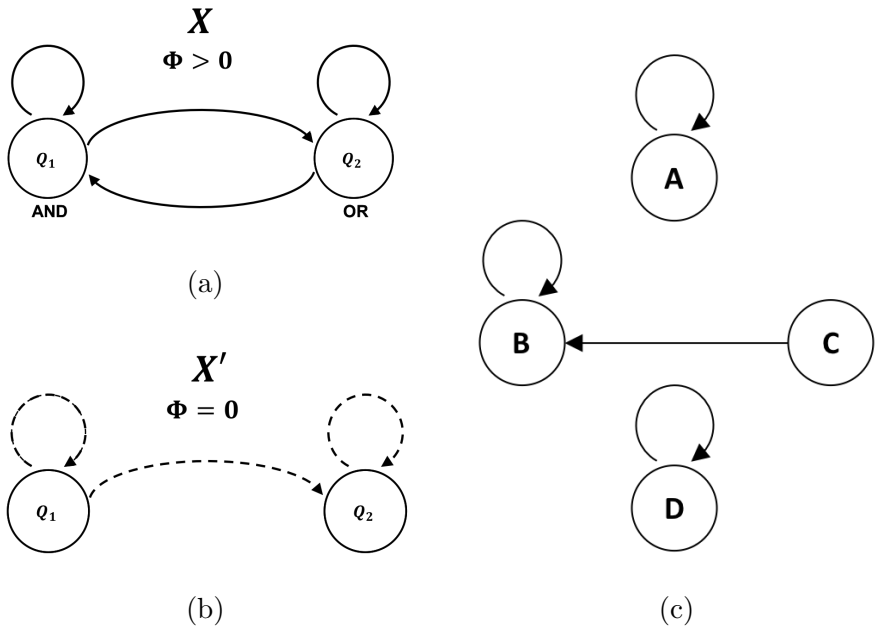
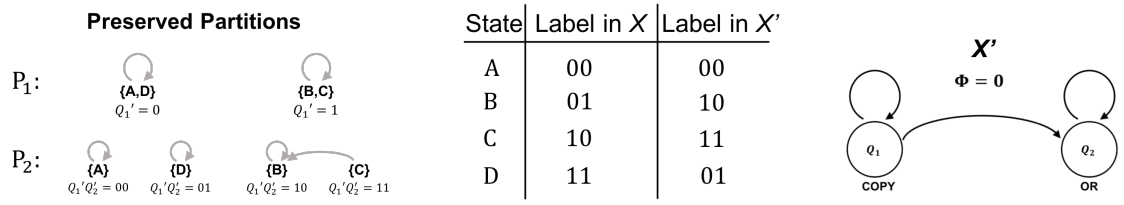


Figure 3.4: The Goal of an Isomorphic Cascade Decomposition is to Decompose the Integrated Logical Architecture of the System X (a) so That It Is in Cascade Form X' (b) Without Affecting the State-transition Topology of the Original System (c).

Given the global state-transition diagram shown in Figure 3.4c, we let our first preserved partition be $P_1 = \{B_0, B_1\}$ with $B_0 = \{A, D\}$ and $B_1 = \{B, C\}$. It is easy to check that this partition is preserved, as one can verify that every element in B_0 transitions to an element in B_0 and every element in B_1 transitions to an element in B_1 (shown topologically in Figure 3.5a). We then assign all the states in B_0 a first coordinate value of 0 and all the states in B_1 a first coordinate value of 1, which guarantees the state of the first coordinate is independent of later coordinates. If the value of the first coordinate is 0 it will remain 0 and if the value of the first coordinate is 1 it will remain 1, because states within a given block transition together. Because 0 goes to 0 and 1 goes to 1, the logic element (component automaton) representing the first coordinate Q'_1 is a `COPY` gate receiving its previous state as input.

The second preserved partition P_2 must evenly split each block within P_1 in half. Letting $P_2 = \{\{B_{00}, B_{01}\}, \{B_{10}, B_{11}\}\}$ we have $B_{00} = \{A\}$, $B_{01} = \{B\}$, $B_{10} = \{C\}$, and $B_{11} = \{D\}$. At this stage, it is trivial to verify that the partition is preserved because each block is comprised of only one state which is guaranteed to transition to a single block. As with Q'_1, the logic gate for the second coordinate (Q'_2) is specified by the way the labeled blocks of P_2 transition. Namely, we have $B_{00} \to B_{00}$, $B_{01} \to B_{01}$, $B_{10} \to B_{01}$, and $B_{11} \to B_{11}$. Note, the transition function δ_{Q2} is completely deterministic given input from the first two coordinates (as required) and is given by $\delta_{Q2} = \{00 \to 0; 01 \to 1; 10 \to 1; 11 \to 1\}$. This implies Q'_2 is an `OR` gate receiving input from both Q'_1 and Q'_2.



State	Label in X	Label in X'
A	00	00
B	01	10
C	10	11
D	11	01

(a) (b) (c)

Figure 3.5: The Nested Sequence of Preserved Partitions in (a) Yields the Isomorphism (b) Between X and X' Which Can Be Translated Into the Strictly Feed-forward Logical Architecture With $\Phi = 0$ Shown in (c).

At this point, the isomorphic cascade decomposition is complete. We have constructed an automaton for Q'_1 that takes input from only itself and an automaton for Q'_2 that takes input only from itself and earlier coordinates (i.e. Q'_1 and Q'_2). The mapping between the states of X and the states of X', shown in Figure 3.5b, is specified by identifying the binary labels (internal representations) each system uses to instantiate the abstract computational states A, B, C, D of the global state-transition diagram. Because X and X' operate on the same support (the same four

binary states) the fact that they are isomorphic implies the difference between representations is nothing more than a permutation of the labels used to instantiate the computation. By choosing a specific labeling scheme based on isomorphic cascade decomposition, we can induce a logical architecture that is guaranteed to be feedback-free and has $\Phi = 0$. In this way, we have "unfolded" the feedback present in X without affecting the size/efficiency of the system.

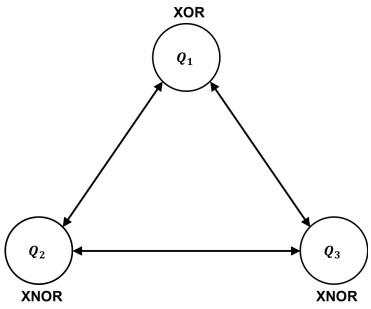
3.4 Results/Discussion

We are now prepared to demonstrate the existence of isomorphic feed-forward philosophical zombies in systems similar to those found in Oizumi *et al.* (2014). To do so, we will decompose the integrated system Y shown in Figure 3.6 into an isomorphic feed-forward philosophical zombie Y' of the form shown in Figure 3.3. The system Y, comprised of two `XNOR` gates and one `XOR` gate, clearly contains feedback between components and has $\Phi > 0$ for all states for which Φ can be calculated (Figure 3.6c). As in Section 3.3.3, the goal of the decomposition is an isomorphic relabeling of the finite-state machine representing the global behavior of the system, such that the induced logical architecture is strictly feed-forward.

We first evenly partition the state space of Y into two blocks $B_0 = \{A, C, G, H\}$ and $B_1 = \{B, D, E, F\}$. Under this partition, B_0 transitions to B_1 and B_1 transitions to B_0, which implies the automaton representing the first coordinate in the new labeling scheme is a `NOT` gate. Note, this choice is not unique, as we could just as easily have chosen a different preserved partition such as $B_0 = \{A, D, E, H\}$ and $B_1 = \{B, C, F, G\}$, in which case the first coordinate would be a `COPY` gate; as long as the partition is preserved, the choice here is arbitrary and amounts to selecting one of several different feed-forward logical architectures - all in cascade form. For the second preserved partition, we let $P_2 = \{\{B_{00}, B_{01}\}, \{B_{10}, B_{11}\}\}$ with $B_{00} = \{C, G\}$,

t_0 \ t_1	000	001	010	011	100	101	110	111
000	0	0	0	1	0	0	0	0
001	0	0	0	0	0	1	0	0
010	0	0	0	0	0	0	1	0
011	1	0	0	0	0	0	0	0
100	1	0	0	0	0	0	0	0
101	0	0	0	0	0	0	1	0
110	0	0	0	0	0	1	0	0
111	0	0	0	1	0	0	0	0

(a)



(b)

State	Φ Value
000	1.875
001	n/a
010	n/a
011	1.875
100	n/a
101	1.875
110	1.875
111	n/a

(c)

Figure 3.6: The Transition Probability Matrix (a), Logical Architecture (b), and All Available Φ Values (c) For the Example System Y. Note, "N/A" Implies Φ Is Not Defined for a given State Because It Is Unreachable.

$B_{01} = \{A, H\}$, $B_{10} = \{B, F\}$. and $B_{11} = \{D, E\}$. The transition function for the automaton representing the second coordinate, given by the movement of these blocks, is: $\delta_{Q2'} = \{00 \to 0; 01 \to 1; 10 \to 0; 11 \to 1\}$, which is again a `COPY` gate receiving input from itself. The third and final partition P_3 assigns each state to its own unique block. As is always the case, this last partition is trivially preserved because individual states are guaranteed to transition to a single block. The transition function for this coordinate, read off the bottom row of Figure 3.7, is given by:

$$\delta_{Q3'} = \{000 \to 0; 001 \to 0; 010 \to 1; 011 \to 1; 100 \to 0; 101 \to 0; 110 \to 1; 111 \to 1\}$$

Using Karnuagh maps (Karnaugh, 1953), one can identify δ_{Q3} as a `COPY` gate receiving input from Q_2'. With the specification of the logic for the third coordinate, the cascade decomposition is complete and the new labeling scheme is shown in Figure 3.7. A side-by-side comparison of the original system Y and the feed-forward system Y' is shown in Figure 3.8. As required, the feed-forward system has $\Phi = 0$ but executes the same sequence of state transitions as the original system, modulo a permutation

of the labels used to instantiate the states of the global state-transition diagram.

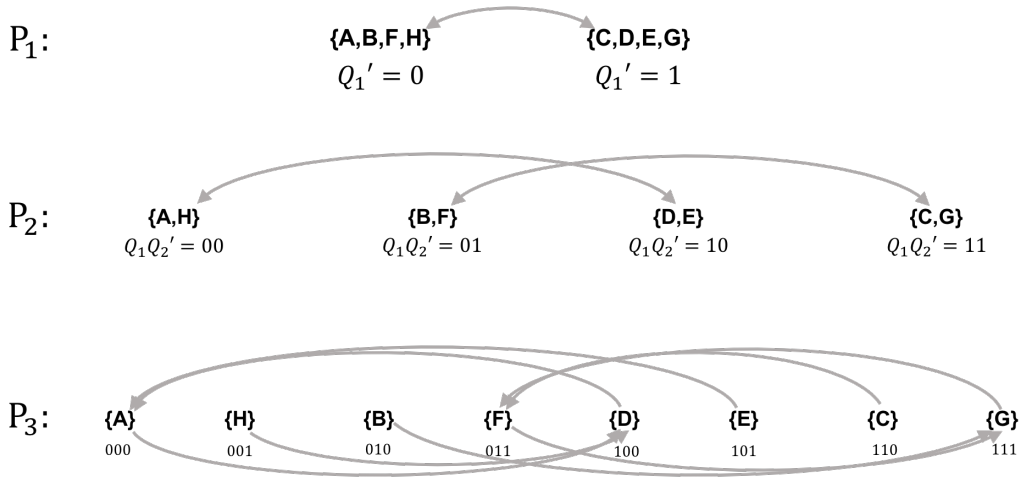


Figure 3.7: Nested Sequence of Preserved Partitions Used to Decompose Y into Cascade Form.

3.4.1 Discussion

Behavior is most frequently described in terms of *abstract* states/stimuli, which are not tied to a specific representation (binary or otherwise). Examples include descriptors of mental states, such as being asleep or awake, etc.: these are representations of system states that must be defined either by an external observer or internally in the system performing the computation by its own logical implementation, but are not necessarily an intrinsic attribute of the computational states themselves (e.g. these states could be labeled with any binary assignment consistent with the state transition diagram of the computation). The analysis presented here is based on this premise, such that behavior is defined by the topology of the state-transition diagram, independent of a particular labeling scheme. And, indeed, it is this premise that enables Krohn-Rhodes decomposition to be useful from an engineering perspective, as one can swap between logical architectures without affecting the operation of

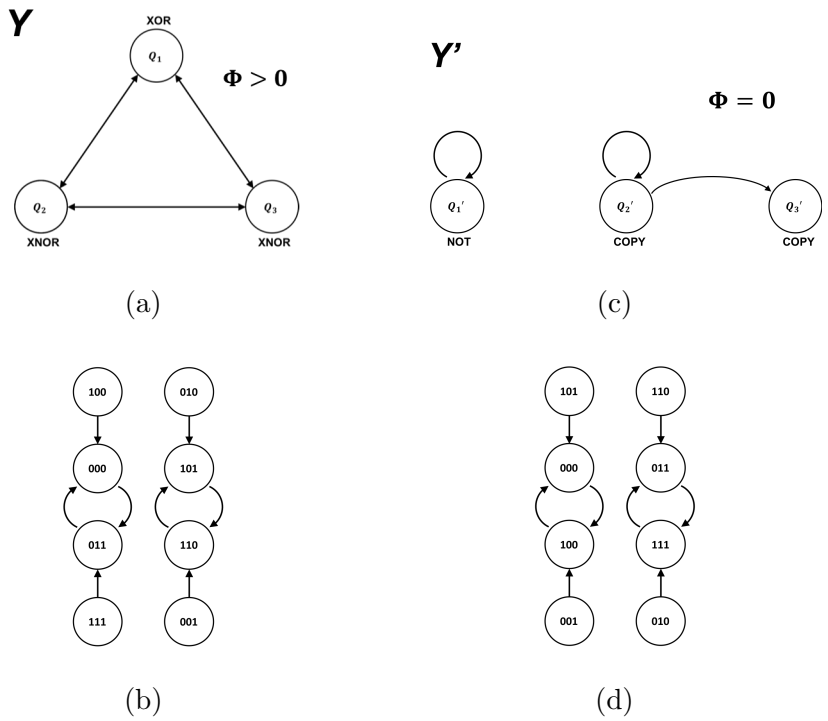


Figure 3.8: Side-by-side Comparison of the Feedback System Y With $\Phi > 0$ (a) and Its Isomorphic Feed-forward Counterpart Y' With $\Phi = 0$ (c). The Respective Global State-transition Diagrams (b) and (d) Differ Only by a Permutation of Labels.

a system in any way.

Phenomenologically, consciousness is often associated with the concept of "top-down causation", where "higher level" mental states exert causal control over lower-level implementation (Ellis, 2016). Under this view, the "additional information" provided by consciousness above and beyond non-conscious systems is considered to be functionally relevant by affecting how states transition to other states. Typically this is associated with a macroscale intervening on a microscale, which historically has been problematic due to the issue of supervenience whereby a system can be causally overdetermined if causation operates across multiple scales (Kim, 2017). Ellis and others have described functional equivalence classes (Ellis, 2016; Auletta *et al.*, 2008)

as one means of implementing top-down causation without causal overdetermination because what does the actual causal work is a functional equivalence class of microstates, which as a class have the same causal consequences. Building on the idea of functional equivalence classes, our formalism introduces a different kind of top-down causation that also avoids issues of supervenience. In our formalism, consciousness is most appropriately thought of as a computation having to do with the topology (causal architecture) of global state transitions, rather than the labels of the states or a specific logical architecture. Thus, the computation/function describes a functional equivalence class of logical architectures that all implement the same causal relations among states, *i.e.* the functional equivalence class is the computational abstraction (macrostate) which can be implemented in any of a set of isomorphic physical architectures (microstates/physical implementations). There is no additional "room at the bottom" for a particular logical architecture to exert more causal influence when it instantiates a particular abstraction than another architecture instantiating the same abstraction, because the causal structure of the abstraction remains unchanged. Any measure of consciousness that changes under the isomorphism we introduce here, such as Φ, cannot, therefore, account for "additional information" related to executing a particular function, because of the existence of zombie systems within the same functional equivalence class.

It is important to recognize there exists an alternative perspective where one defines differences relevant to consciousness not in terms of abstract computation, but in terms of specific logical implementations, as IIT adopts. However, this does not address, in our view, whether isomorphic systems ultimately experience a phenomenological difference as there is no way to test that assumption other than accepting it axiomatically. In particular, for the examples we consider here, there is only one input signal, meaning there are not multiple ways to encode input from the environment.

Therefore there is no physical mechanism by which the environment can dictate a privileged internal representation. Instead, the choice of internal representation is arbitrary with respect to the environment and depends only on the physical constraints of the architecture of the system performing the computation. For a system as complex as the human brain, there are presumably many possible logical architectures that define an equivalence class capable of performing the same computation given the same input, differing only in how the states are internally represented (*e.g.* by how neurons are wired together). This could, for example, explain why human brains all have the potential to be conscious despite differences in the particular wiring of their neurons. Why the internal representations that have evolved were selected for in the first place is likely important for understanding why consciousness emerged in the universe.

The foregoing provides examples where Φ is independent of both functionality and efficiency. Historically within IIT, the presence of feedback is associated with *efficiency*, such that unconscious feed-forward systems like those presented in Oizumi *et al.* (2014) and Doerig *et al.* (2019) operate under drastically different resource constraints than their conscious counterparts with feedback. This motivates arguments for an evolutionary advantage toward efficient representation and, by proxy, Φ/consciousness (Albantakis *et al.*, 2014). Under isomorphic decomposition, however, systems with feedback can be assumed to have equivalent efficiency to their counterparts without feedback, because the size of the system and its state transitions are equivalent - demonstrating that Φ is fundamentally distinct from efficiency. This, in turn, implies there is no inherent evolutionary benefit to the presence or absence of Φ because it is not selectable as being distinctive to a particular computation an organism must perform for survival, but only how that computation is internally represented.

Given that isomorphic systems exhibit the same behavior, meaning that they take the exact same trajectory through state space, modulo a permutation of the labels used to represent the states, they can also be considered to be equally autonomous. This is because the internal states of isomorphic systems are in one-to-one correspondence (Figure 3.5b), meaning the presence/absence of autonomy does not affect the transitions in the global state diagram of the system (e.g. Figure 3.8b/3.8d). In our examples, the future state of the system as a whole is completely determined its current state for both the zombie system and its conscious counterpart. Similarly, the future state of each individual component within a given system is completely determined by its inputs (i.e. it is a deterministic logic gate). This presents a challenge for understanding in what sense a system with $\Phi > 0$ can be said to be dictating its own future from within, while the system with $\Phi = 0$ is not, as IIT suggests. The notion of autonomy that IIT adopts to address this is one of interdependence - autonomous systems rely on bi-directional information exchange between components while non-autonomous systems do not. Yet, given that each component can store only one bit of information, components cannot store *where* the information came from (e.g. whether or not they are part of an integrated architecture). Since the information stored by the system as a whole is nothing more than the combined information stored by individual components, it is unclear to us why feedback between elements should result in autonomy while feed-forward connections between elements should not, given isomorphic state-transition diagrams.

This leads us to the central question of this manuscript, which is what is experienced as the isomorphic system with $\Phi > 0$ cycles through its internal states that is not experienced by the isomorphic system with $\Phi = 0$? Since in our examples the environment is not dictating the representation of the input, and all state transitions are isomorphic, the representation and therefore the logic is arbitrary so long

as a logical architecture is selected with the proper input-output map under all circumstances. In light of this, our formalism suggests any mathematical measure of consciousness, phenomenologically motivated or otherwise, must be invariant with respect to isomorphic state-transition diagrams. This minimal criterion implies measurable differences in consciousness are always associated with measurable differences in the computation being performed by the system (though the inverse need not be true), which is nothing more than precise mathematical enforcement of the precedent set by Turing (1950). From this perspective, measures of consciousness should operate on the topology of the state-transition diagram, rather than the logic of a particular physical implementation. That is, they should probe the computational capacity of the system without being biased by a particular logical architecture - allowing identifying equivalence classes of physical systems that could have the same or similar conscious experience.

Our motivation in this work is to provide new roads to address the hard problem of consciousness by raising new questions. Our framework focuses attention on the fact that we currently lack a sufficiently formal understanding of the relationship between physical implementation and computation to truly address the hard problem. The logical architectures in Figures 3.8a and 3.8c are radically different, and yet, they perform the same computation. The fact that this computation allows a feed-forward decomposition is a consequence of redundancies that allow a compressed description in terms of a feed-forward logical architecture. There are symmetries present in the computation that allow one to take advantage of shortcuts and reduce the computational load. This, in turn, shows up as flexibility in the logical architecture that can generate the computation. In other words, the computation in question does not appear to require the maximum computational power of a three-bit logical architecture. For sufficiently complex eight-state computations, however, the full capacity

of a three-bit architecture is required, as there is no redundancy to compress. Such systems cannot be generated without feedback, as the presence of feedback is accompanied by indispensable functional consequences. In this case, the *computation* is special because it cannot be efficiently represented without feedback - a relationship that can, in principle, be understood but is only tangentially accounted for in current formalisms. It is up to the community to decide if the causal mechanisms of consciousness are at the level of particular logical architectures or the computations they instantiate, our goal in this work is simply to point out where the distinction between the two sets of ideas is very apparent and clear-cut mathematically so that additional progress can be made.

3.5 Acknowledgements

The authors would like to thank Doug Moore for his assistance with the Krohn-Rhodes theorem and semigroup theory, as well as Dylan Gagler and the rest of the emergence@asu lab for thoughtful feed-back and discussions. SIW also acknowledges funding support from the Foundational Questions in Science Institute.

Chapter 4

FORMALIZING FALSIFICATION OF CAUSAL STRUCTURE THEORIES OF CONSCIOUSNESS ACROSS COMPUTATIONAL HIERARCHIES

4.1 Abstract

There is currently a global, multimillion-dollar effort to experimentally confirm or falsify neuroscience's preeminent theory of consciousness: Integrated Information Theory (IIT). Yet, recent theoretical work suggests major epistemic concerns regarding the validity of IIT and all so-called causal structure theories. In particular, causal structure theories are based on the assumption that consciousness supervenes on a particular causal structure, despite the fact that different causal structures can lead to the same input-output behavior and global functionality. This, in turn, leads to epistemic problems when it comes to the ability to falsify such a theory - if two systems are functionally identical, what remains to justify a difference in subjective experience? Here, we ground these abstract epistemic problems in a concrete example of functionally indistinguishable systems with different causal architectures. Our example comes in the form of an isomorphic feed-forward decomposition (unfolding) of a simple electronic tollbooth, which we use to demonstrate a clear falsification of causal structure theories such as IIT. We conclude with a brief discussion regarding the level of formal description at which a candidate measure of consciousness must operate if it is to be considered scientific.

4.2 Introduction

If, and if so how, theories for consciousness can be brought within the purview of science is a subject of intense debate and equally intense importance. Resolution of this debate is necessary for validating theory against experiments in human subjects. It is also critical to recognizing and/or engineering consciousness in non-human systems such as machines. Currently, there is a global, multi-million dollar effort devoted to scientifically validating or refuting the most promising candidate theories (Reardon, 2019), specifically Integrated Information Theory and Global Neuronal Workspace. At the same time, it is becoming increasingly unclear whether these theories meet the required scientific criteria for validating them.

Since the early 1990s, scientific studies of consciousness have primarily focused on identifying spatiotemporal patterns in the brain that correlate with what we intuitively consider to be conscious experience. This is due in large part to advances in medical imaging such as electroencephalograms (EEG) and functional magnetic resonance imaging (fMRI) that assess brain activity during different activities (e.g., sleeping, verbal reports, etc.). The empirical data that results from such tests provide evidence for links between spatiotemporal patterns and inferred conscious states. These links, known as Neural Correlates of Consciousness (NCCs), are well-established and form the basis for an entire subfield of contemporary neuroscience (Rees *et al.*, 2002; Metzinger, 2000). Despite the success of NCCs, however, there is an underlying epistemic issue with the scientific study of consciousness because conscious states are never directly observed in the NCC framework. Instead, they are inferred based on our own phenomenological experience. For example, when a person is asleep we infer they are less conscious than when awake because we have a first-hand subjective experience of being awake but not of being in deep sleep.

While this epistemic issue is widely known, Kleiner and Hoel (2020) (abbreviated herein as KH) have recently formalized the scientific issues arising when consciousness is inferred based on correlates, rather than directly measuring it, revealing a pervasive problem with current theoretical frameworks attempting to formalize consciousness. Their analysis leads them to the startling conclusion that all contemporary theories of consciousness are either already falsified or unfalsifiable. In KH, falsification is formally defined as a mismatch between what a theory predicts based on observations, $pred(O)$, and what is inferred from observations, $inf(O)$. Consequently, a theory is falsified if one can substitute a physical system for another in a way that changes $pred(O)$ but preserves $inf(O)$. The authors prove such a substitution exists for all contemporary theories of consciousness that treat inferences and observations independently including Integrated Information Theory (IIT) (Tononi, 2008; Oizumi *et al.*, 2014), Global Neuronal Workspace (Sergent and Dehaene, 2004), Recurrent Processing Theory (Lamme, 2006), and Higher-Order Thought Theory (Rosenthal, 2002). Conversely, if a theory of consciousness treats inferences and predictions as strictly dependent, then the theory is necessarily unfalsifiable, as no experiment could possibly find a mismatch between what is predicted and what is inferred. Contemporary theories of consciousness that suffer from this issue include Global Workspace Theory (Baars, 2005), Attention Schema Theory (Graziano and Webb, 2015), and any behaviorist theory in general (Graham, 2019).

The argument made by KH is actually a generalization of a previous argument made by Doerig *et al.* (2019), wherein the authors focused on a particular theory (IIT) and a particular type of substitution known as "unfolding". According to IIT, feedback plays an essential role in generating conscious experience. The motivation for this assumption is that, phenomenologically, we experience consciousness as an "undivided whole", meaning, for example, that our left and right visual fields are

integrated into a single conscious experience. IIT offers a mathematical measure of integration Φ that equates to the overall level of consciousness. Integration, as a phenomenological axiom of the theory, therefore must have a direct translation in terms of mathematical machinery. The way this is accomplished in IIT is by assuming integrated experience is mirrored by integration of the physical substrate that gives rise to consciousness, where the latter use of the term integration has a precise mathematical definition in terms of the presence of feedback between the physical components in a system (e.g., neurons). Consequently, any system that is strictly feed-forward is unconscious, by definition in IIT, due to an assumed inability for such physical structures to generate a unified subjective experience. What Doerig *et al.* (2019) showed was that the input-output behavior of any conscious system with feedback and $\Phi > 0$ can be perfectly emulated by a strictly feed-forward system with $\Phi = 0$. To do so, one simply needs to "unfold" the feedback present in the causal structure of the conscious system in a way that preserves the underlying functionality of the system (i.e. the input-output behavior) - a feat that can be accomplished in the forward or backward direction using feed-forward and recurrent neural networks, respectively (Doerig *et al.*, 2019). In the formalism of KH, this unfolding argument proved that within IIT one can always find a substitution of causal structures that preserves $inf(O)$ but changes $pred(O)$, therefore falsifying the theory.

Interestingly, unfolding substitutions were known in IIT prior to the work of Doerig *et al.* (2019) but were not necessarily considered detrimental. In fact, Oizumi *et al.* (2014) explicitly considered feed-forward substitutions in the development of IIT 3.0 but subsequently dismissed them as inconsequential. The justification for this was primarily the assumption that feedback is a necessary condition for an "integrated experience" but, again, we stress that this assumption, known as the integration axiom, has two distinct interpretations: the phenomenological axiom and the mathematical

translation of the phenomenological axiom. While few would argue against integration as a phenomenological axiom, the way that it is translated into mathematical machinery is the subject of the epistemic concerns raised by the unfolding argument. In particular, how does one justify, scientifically, that feedback is indeed what embodies the subjective experience of an integrated whole in absence of any particular functional consequences?

In answer to this question, many authors have put forth the idea that meaningful differences can and do exist between functionally identical systems at a formal level of description below the finite-state automaton (FSA) description of the system. In particular, causal structure theories posit that it is the way a computation is *instantiated* rather than the computation in the abstract that is relevant in determining consciousness. In a previous work, we showed that a particular instantiation of a computation is a direct consequence of the labels that are assigned to represent the abstract functional states of a computation, meaning that different causal structures result from different *encodings* of the same computation (Hanson and Walker, 2019). Because causal structure supervenes on a particular encoding, the so-called "combinatorial-state automaton" (CSA) description of a system is nothing more than a labeled version of the FSA description (see Figure 4.1). This implies causal structure theories such as IIT are assuming that the way a system encodes a computation is relevant to whether or not it is conscious of the computation which, in computer science terms, is analogous to the claim that it is binary (compiled) code that determines whether or not artificially intelligent machines are conscious rather than the abstract (functional) code being executed. Similarly, one can go one step beyond traditional causal structure theories (which act at the CSA level) and posit that it is the specific material properties or the choice of logical basis that is relevant for determining consciousness. In light of this hierarchy (Harnad, 2006), the main claim of the unfolding

argument and its subsequent generalization is that we must infer consciousness at the level of the FSA description of a system, as only this level has a phenomenological grounding in terms of first-hand experience. Consequently, any measure of consciousness that is not invariant with respect to changes that preserve the FSA description of a system is either falsified or unfalsifiable, depending on whether one assumes the inference procedure or the prediction from the theory is correct.

In this work, we seek to ground the abstract, epistemic problems associated with the unfolding argument – and indeed, more general arguments of falsifiability - in a concrete, easily visualizable system that can readily be realized using widely available tabletop electronics. In particular, we construct isomorphic causal structures (digital circuits) designed to operate a simple electronic tollbooth. The utility of this approach is that it provides a clear falsification of causal structure theories such as IIT in terms of the scale at which they operate. In other words, formalizing this epistemic hierarchy and its degeneracies in the context of inference and prediction allows investigating not only how theories of consciousness might be falsified or are unfalsifiable following on the work of Doerig *et al.* (2019) and Kleiner and Hoel (2020), but also at *what level of the computational hierarchy* (FSA or CSA) a theory of consciousness is falsified or unfalsifiable.

4.3 Results

The different levels of abstraction at which it is possible to specify a computation can be assessed explicitly for theories of consciousness, such as IIT, by constructing automata and circuits representing different levels in the hierarchy in Figure 4.1. We do so, using the formalism we developed in Hanson and Walker (2019), by constructing functionally identical machines operating under the same resource constraints using different causal architectures (circuits). We consider a very simple case of the

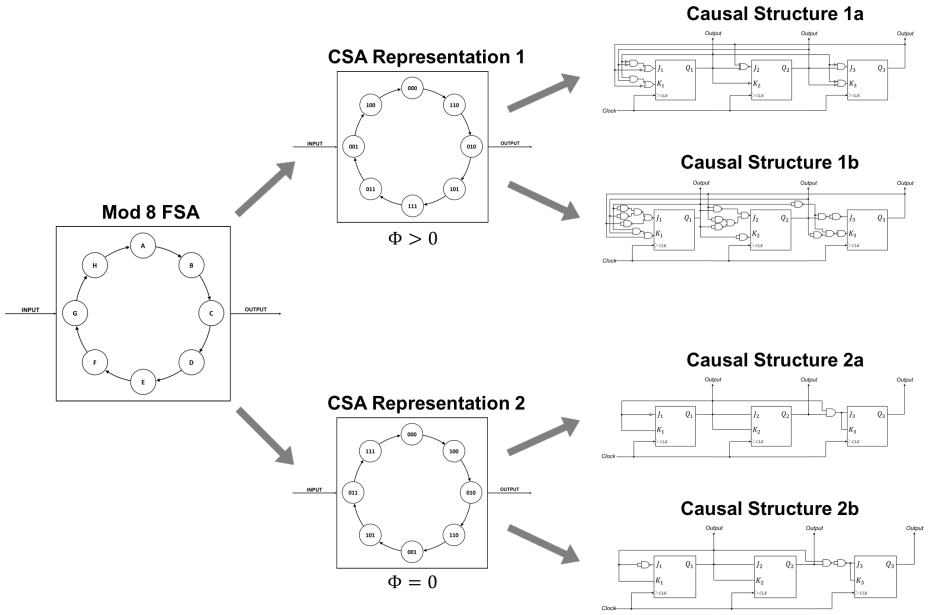
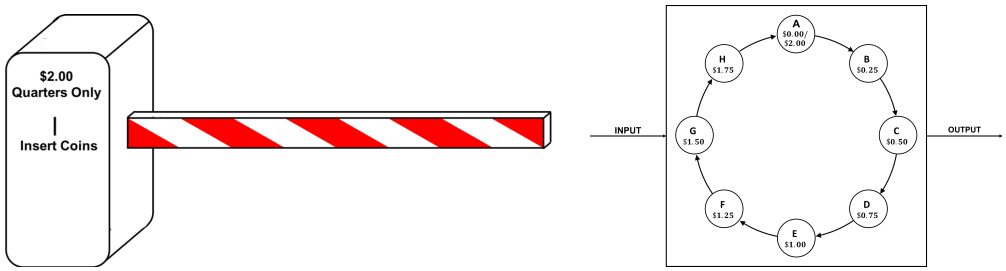


Figure 4.1: Different Levels of Abstraction Where a Computational Theory of Consciousness Can Apply. At the Level of the Computation in the Abstract, the Topology of a Finite-state Automaton (FSA) Is Specified – in This Case, a Mod-eight Counter (Left). Encoding These Abstract Functional States with a Specific Binary Representation Results in a Combinatorial-state Automaton (CSA), Which Constrains Local Dependencies Between Subcomponents Within a System and Is the Level at Which Φ Is Calculated (Center). To Fully Specify the Causal Structure, However, One Must Still Choose a Set of Elementary Logic Gates to Realize a given CSA (Right). In This Case, We Have Shown Two Different Choices for an Elementary Logical Basis: `AND`/`OR`/`NOT` Gates (1a, 2a) and Universal `NAND` Gates (1b, 2b).

design of a simplified electronic tollbooth using a causal architecture with and without feedback. Focusing on feedback, as opposed to some other difference in causal architecture, allows us to ground our results in the specifics of Integrated Information Theory (IIT), where feedback is assumed to be a necessary condition for the presence of consciousness (i.e., $\Phi > 0$). We focus on IIT as it is the most mathematically rigor-

ous theory of consciousness developed to date. However, we expect the quantitative approach to addressing epistemic issues of falsification to also be possible for other theories of consciousness.

The qualitative description of the tollbooth's behavior is to lift a boom barrier upon receipt of exactly eight quarters, as shown in Figure 4.2a. To do this, the circuit governing the behavior of the tollbooth must transition through eight internal memory states, corresponding to the eight functional states in the FSA description of the machine shown in Figure 4.2b). To control for system size, we insist that both circuits are constructed on a three-bit logical architecture, which serves to enforce a strict one-to-one correspondence (isomorphism) between the internal states of the two systems.



(a) (b)

Figure 4.2: Schematic Illustration of a Simplified Electronic Tollbooth (a) and Its FSA Description (b). The General Behavior of the Tollbooth Is to Lift a Boom Barrier upon Receipt of Eight Quarters ($2.00). To Do This Requires the Ability to Cycle Through Eight Internal Memory States $\{A, B, ..., H\}$, Sending Each Internal State as Output to the Boom Barrier. Note, for the Tollbooth to Function Correctly, the Boom Barrier Must Be Programmed to Recognize Internal State A as Functionally Important, as This Is the Output That Causes the Boom Barrier to Lift and Reset.

We will first demonstrate a "conscious" circuit with feedback (and $\Phi > 0$), fol-

lowed by a functionally identical but "unconscious" circuit with strictly feed-forward connections (and $\Phi = 0$). The general construction of both circuits is the same: first, we assign binary labels to the functional states of the system; then, we map these binary state transitions onto JK flip-flops, which are the "bits" in our digital electronics; and last, we use Karnaugh Maps to simplify the logic tables of the JK flip-flops in a way that results in simple elementary logic gate operations (e.g. `AND`, `OR`, `XOR`). As we show, the presence or absence of feedback ultimately stems from the initial choice of the binary labels used to *represent* or *encode* the functional states of the system. For the system with feedback, we randomly assign these labels in a way that happens to result in $\Phi > 0$ for all states. For the feed-forward system, however, we carefully decompose the underlying dynamics in a way that exploits hierarchical relations such that information flow between components in the system is strictly uni-directional. The result is a "hierarchical coordinate scheme" wherein each JK flip-flop is responsible for keeping track of a particular symmetry in the state transition diagram of the original system.

4.3.1 Constructing a Conscious Tollbooth

The construction of any particular causal architecture requires specification of the way in which functional relationships are physically instantiated. In other words, there are two distinct "levels" at which the functional topology can be specified: in terms abstract states and their mathematical relations (the FSA level of description) (Chalmers, 1993), or in terms of specific causal relations between subcomponents (the CSA level of description) (Chalmers, 1993). The difference between these two levels of abstraction is equivalent in the automata formulation we present here to whether or not binary labels have been assigned to represent functional states such as those shown in Figure 4.2b. The presence of binary labels restricts the causal relationships

between subcomponents within the system, thereby constraining the causal structure.

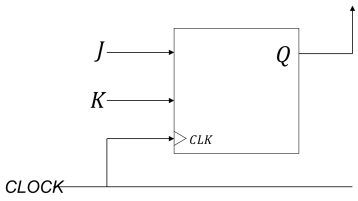
To construct the conscious tollbooth, we randomly assign the following binary labels to represent the eight functional states of the tollbooth:

$$A = 000, B = 110, C = 010, D = 101, E = 111, F = 011, G = 001, H = 100$$

This assignment of labels fully specifies the Boolean logic of the system, as each binary component (bit) now must transition in accordance with the global state of the system. For example, the transition from state A to state B requires that the first component of the system transitions from state 0 to state 1 when the system is in the global state 000. Similarly, the transition from state B to state C specifies that the first component of the system must transition from 1 to 0 when the system is in global state 110. Taken together, the constraints on each individual component in the system at each moment in time provide sufficient criteria for constructing a digital circuit that governs this system. As a sidenote, we remark that for the boom barrier to function correctly it now must be programmed to recognize the binary state $A = 000$ as the "$2.00 state". This means the sensorimotor hardware of the system must be wired in such a way that it "knows" to lift the boom barrier (and reset) when there is a lack of voltage on the three output lines coming from the circuit, which can be accomplished via an encoder/decoder device that translates signals from the internal circuitry to the external hardware or by hardwiring the machinery of the boom barrier directly.

To finish the construction of the causal architecture, we must specify the elementary building blocks of our system. In a human brain, these building blocks would be neurons but in a digital circuit, these building blocks are "JK flip-flops", which are binary memory storage devices (bits) widely used in the construction simple digital circuits (Moore, 1958; Cavanagh, 2018). The behavior of a JK flip-flop is quite simple:

there are two stable internal states (0 and 1), two input channels (the J input and the K input), and a "clock" that serves to synchronize multiple flip-flops within a system. Upon receipt of voltage on a line from the clock, the flip-flop does one of four things depending on the input from the J and K channels: if the JK input is 00 the internal state remains constant ("latch"), if the JK input is 01 the internal state resets to 0 ("reset"), if the JK input is 10 the internal state is set to 1 ("set"), and if the JK input is 11 the internal state is swapped ("toggle"). Thus, for any given internal state transition - $Q_i(t_0) \rightarrow Q_i(t_1)$ - there are two different possibilities for JK inputs that will correctly realize this transition, as shown in Figure 4.3. This degeneracy provides much-needed flexibility when it comes to the design of the elementary logic gate operations required to actually realize the underlying Boolean logic.



(a)

$Q(t_0) \rightarrow Q(t_1)$	Behavior	JK Value
$0 \rightarrow 0$	Latch/Reset	00/01
$1 \rightarrow 0$	Toggle/Reset	11/01
$0 \rightarrow 1$	Toggle/Set	11/10
$1 \rightarrow 1$	Latch/Set	00/10

(b)

Figure 4.3: A JK Flip-flop Is a Commonly Used Binary Storage Device (Bit) in Digital Electronics (a). The Internal State of the Flip-flop Takes One of Two Values ($Q \in \{0, 1\}$) and Is Continuously Sent as Output. Upon Receipt of a Voltage from a Clocked Input, the Voltages on the Two Input Channels J and K Dictate the State Transitions of Q (See Main). For Any given State Transition $Q(t_0) \rightarrow Q(t_1)$, There Are Two Combinations of JK Inputs That Will Correctly Realize the Transition (b), Which Provides Much-needed Flexibility When It Comes to Elementary Logic Gate Descriptions.

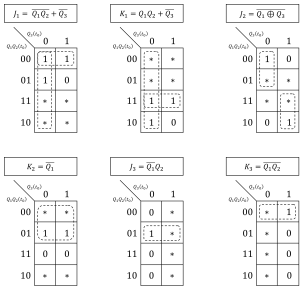
With the specification of the binary labels and the choice of electronic components,

we can now go about actually building the digital circuit using elementary logic gates. To do so, we first convert the state transitions of each individual component into its associated JK value. As mentioned, there is degeneracy in the choice of JK input which means we only have to specify one of the input channels (either J or K) to get the correct transition. For each component in the circuit, there is a column in Figure 4.4a corresponding to the JK value that is required; note, inputs that do not need to be specified are denoted with an asterisk. Next, we must determine the elementary logic gates required to get the correct JK values given the current state of the system. For instance, when the system is in global state 110, the value of K_1 (the K-input to the first component) must be 1, but when the system is in global state 111 the value of K_1 must be 0. Taken together, the eight states of the system comprise a truth table of JK input as a function of the global state of the system, as shown in Figure 4.4b. Ordering these truth tables in gray code yields "Karnaugh maps", which allow straightforward identification of the elementary logic gates required to operate the circuit (Karnaugh, 1953). The elementary logic expression for each of the six input channels, in terms of `AND`,`OR`, `XOR`, and `NOT` gates, is shown above the corresponding Karnaugh map in Figure 4.4b.

The elementary logic expressions for the behavior of each JK input complete the construction of our circuit, which is shown in Figure 4.5a. Clearly, this circuit contains meaningful feedback between components, as the state of the first component depends on the state of the second and third and vis versa (e.g. $J_1 = \overline{Q_1 Q_2}$ and $K_1 = Q_2 \oplus Q_3$). The last thing to check is whether or not this feedback is associated with the presence of consciousness according to IIT, as feedback is a necessary (but not sufficient) condition for $\Phi > 0$. Using the python package PyPhi (Mayner *et al.*, 2018), we find $\Phi > 0$ for all states (Figure 4.5b), meaning this system is indeed considered conscious according to IIT.

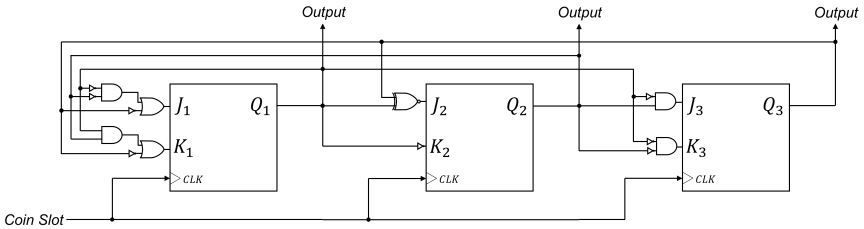
$Q_1Q_2Q_3(t_0)$	$Q_1Q_2Q_3(t_1)$	J_1K_1	J_2K_2	J_3K_3
000	110	$1*$	$1*$	$0*$
110	010	$*1$	$*0$	$0*$
010	101	$1*$	$*1$	$1*$
101	111	$*0$	$1*$	$*0$
111	011	$*1$	$*0$	$*0$
011	001	$0*$	$*1$	$*0$
001	100	$1*$	$0*$	$*1$
100	000	$*1$	$0*$	$0*$

(a)



(b)

Figure 4.4: To Construct the Digital Circuitry for a given Labeling Scheme, We Must Convert the Global State Transitions into Their Associated JK Values (a). Then, We Use Karnaugh Maps to Determine the Elementary Logic Required to Correctly Update Each Component (b). The Presence of Feedback in the Resultant Digital Circuit Is Evident by the Dependence of Earlier Components on Later Components (e.g. $J_1 = \overline{Q_1Q_2} + \overline{Q_3}$) and Vice Versa (e.g. $K_3 = \overline{Q_1Q_2}$).



s_0	$\Phi(s_0)$
000	0.9775
110	1.4687
010	0.9775
101	1.7187
111	1.7187
011	1.4688
001	1.4688
100	1.4688

(a)

(b)

Figure 4.5: A Three-bit Causal Architecture Comprised of JK Flip-flops Capable of Perfectly Operating the Electronic Tollbooth Shown in Figure 4.2. Clearly, This System Contains Feedback in the Form of Bidirectional Dependence Between Elements (a). In Addition, It Has $\Phi > 0$ for All States (b) Which Implies It Is Conscious According to IIT.

4.3.2 Constructing an Unconscious Tollbooth

In the previous section, we demonstrated the construction of a causal structure with feedback that is designed to operate the electronic tollbooth shown in Figure

4.2a. We did so by randomly assigning 3-bit binary labels to represent the function states ($\{A, B, ..., H\}$) of the system and constructing the logic of the digital circuit in a way that correctly realizes these labeled state transitions. The result was a circuit that utilized feedback connections (i.e. there was bi-directional information exchange between components) and had $\Phi > 0$ for all states (Figure 4.5). In this section, we demonstrate that it is possible to assign binary labels in a different way, such that the causal architecture that results instantiates the same functional topology (Figure 4.2b) without the use of feedback. In other words, we "unfold" the underlying dynamics of the system in a way that guarantees a causal architecture with $\Phi = 0$ for all states in the system.

The process of unfolding a finite-state description of a system is based on techniques closely related to the Krohn-Rhodes theorem from automata theory, which states: any abstract deterministic finite-state automata (FSA) can be realized using a strictly feed-forward causal architecture comprised solely of simple elementary components (Krohn and Rhodes, 1965; Zeiger, 1967b). To do so isomorphically, one must find a "nested sequence of preserved partitions", which creates a hierarchical labeling scheme wherein earlier components (flip-flops) transition independently of later components (Zeiger, 1968; Hanson and Walker, 2019). Due to this hierarchical independence, information is guaranteed to flow unidirectionally from earlier components to later components, thereby ensuring a strictly feed-forward logical architecture and $\Phi = 0$ for all states. While a full discussion of Krohn-Rhodes decomposition is well beyond the scope of this paper (Egri-Nagy and Nehaniv, 2015), we briefly describe the relevant methodology for constructing a nested sequence of preserved partitions in the Methods section. The result, applied to the finite-state description of the toll-booth shown in Figure 4.2b, is the following set of binary labels used to represent the

functional states of our system:

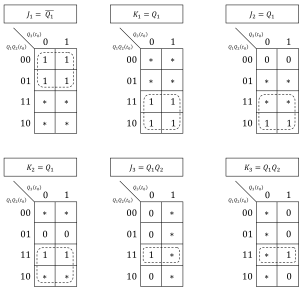
$$A = 000, B = 100, C = 010, D = 110, E = 001, F = 101, G = 011, H = 111$$

Notice, in this labeling scheme, the value of the first component (also called a "coordinate") partitions the underlying state space of the system into two macrostates: $\{A, C, E, G\}$ and $\{B, D, F, H\}$. These macrostates are relevant due to the fact they transition deterministically back and forth between one another. Thus, knowing the future state of the first component depends solely on knowing the current state of the first component. Similarly, the future state of the second component is completely deterministic given the current state of the first and second components and is agnostic to the third. In this way, each additional component offers a refined estimate as to where in a given macrostate the current microstate is located (DeDeo, 2011), which justifies the claim that the labeling scheme is "hierarchical".

With hierarchical labels assigned, the circuit construction now proceeds identically to the previous section. Namely, we convert the binary state transitions into their associated JK values, shown in Figure 4.6a. Then, we construct truth tables for the state of each J and K input given the global state of the system; and last, we order these truth tables in gray code (Karnaugh Maps) and assign elementary logic gates to each input channel (Figure 4.6b). The resulting logical architecture is shown in Figure 4.7a). As required, the circuit is indeed strictly feed-forward, as evident by the fact that each component depends solely on itself or earlier components. This, in turn, guarantees $\Phi = 0$ for all states of the system (Figure 4.7b) as the presence of feedback connections is assumed to be a necessary condition for consciousness in IIT.

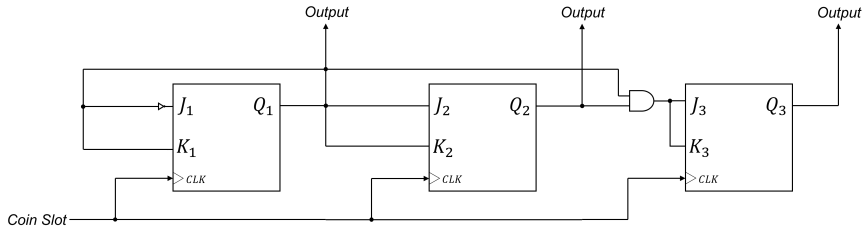
$Q_1'Q_2'Q_3'(t_0)$	$Q_1'Q_2'Q_3'(t_1)$	J_1K_1	J_2K_2	J_3K_3
000	100	$1*$	$0*$	$0*$
100	010	$*1$	$1*$	$0*$
010	110	$1*$	$*0$	$0*$
110	001	$*1$	$*1$	$1*$
001	101	$1*$	$0*$	$*0$
101	011	$*1$	$1*$	$*0$
011	111	$1*$	$*0$	$*0$
111	000	$*1$	$*1$	$*1$

(a)



(b)

Figure 4.6: The State Transitions and JK Values (a) Corresponding to the Hierarchical Labeling Scheme Described in the Main Text. Figure (b) Shows the Karnaugh Maps Used to Determine the Elementary Logic Gates Used in the Construction of the Feed-forward Logical Architecture. Note, the Logical Dependence Between Components Is Strictly Unidirectional (e.g. J_2 and K_2 Depend Only On the State of Q_1).



(a)

s_0	$\Phi(s_0)$
000	0.0000
100	0.0000
010	0.0000
110	0.0000
001	0.0000
101	0.0000
011	0.0000
111	0.0000

(b)

Figure 4.7: The Unfolded Causal Structure That Results from the Hierarchical Labeling Scheme Described in the Main Text (a). This Strictly Feed-forward Causal Structure Operates under the Same Resource Constraints as the Feedback System (Three-bit Logical Architecture) but Has $\Phi = 0$ for All States (b).

4.4 Discussion

While we have presented a specific example of a feed-forward isomorphic transformation as a way to probe causal structure theories such as IIT, the arguments

and their applicability are general. We do not need to realize additional feed-forward automata emulators in a lab to know that they exist, as the Krohn-Rhodes theorem guarantees such a decomposition is always possible. This has implications for the design of machines with "human" intelligence. Importantly, our example highlights how the assessment of falsification depends on what level of the computational hierarchy one assumes consciousness arises within a given theory. In our example, we can consider the implications at either the FSA or CSA level. At the level of FSAs we provided an explicit case of equivalent computations implemented with different causal structures implemented at the CSA level, and each CSA can further be instantiated in different circuits depending on the choice of logic gates being used. This forms a hierarchy of supervenient computations FSA $\rightarrow$ CSA $\rightarrow$ circuit (Figure 4.1). Recall, according to KH, a theory of consciousness is falsified if there is a mismatch between what the theory predicts based on observation $pred(O)$ and what the inferred conscious experience of the system is $inf(O)$ (Kleiner and Hoel, 2020). Conversely, a theory is unfalsifiable if $pred(O)$ is dependent on $inf(O)$. The setup of the tollbooth example highlights how these two criteria play out in a concrete example. The tollbooth setup carefully controls for all confounding factors, creating a situation in which everything in the automata that could be used to *infer* a difference in the conscious states of the two tollbooths (e.g. behavior, functional topology, and efficiency) is fixed - the only difference is that which emerges at the CSA level, as a consequence of the difference in labeling. Causal structure theories are falsified at the FSA level because there is a mismatch between the inference and prediction: equivalent systems at the FSA level can be conscious or not depending on lower-level implementations. Inferences about any differences in conscious experience must therefore occur at the CSA level if causal structure theories. However, this is also where predictions are made about what systems are conscious or not in causal structure theories. While

this coupling of prediction and inference may save a theory from falsification, it has the unintended consequence of rendering the theory unfalsifiable at the CSA level, as a difference in prediction and inference (a requirement for falsification) is no longer possible (see e.g., Kleiner and Hoel (2020)). Thus, causal structure theories, such as IIT, are show to be falsified at the FSA level and unfalsifiable at the CSA level.

By formalizing the epsitemic issues surrounding consciousness in the concerete mathematical formalization of automata theory, the approach we present here enables applying the formal arguments of Kleiner and Hoel to theories of consciousness at specific levels of abstraction in order to test their validity. Not the least of which is that it directly connects validation of theories for consciousness to foundations of computer science. The formalization we present is, in fact, is a concrete mathematical analysis of the Turing test: at the FSA level our machines are Turing indistinguishable. It is only in the assignment of representation (labeling) that they become distinguishable. Thus, our unconscious tollbooth could be said to pass the Turing test at the FSA level. Furthermore, our focus on isomorphisms supports prior proposals that the measurement of "isomorphic experiences", which differ in content but not overall quality of experience, may indeed be possible - even if the subjectivity of the experience itself is not. To approach analyzing such isomorphic physical systems (humans and/or machines) one must find a level of abstraction that is invariant between two physical implementations. There is a direct correspondence between a FSA - conventionally called a deterministic finite-state automaton (DFA) - description of a behavior and the algebraic theory of semigroups. Namely, every DFA maps directly onto a transformation semi-group, which implies the rich mathematics of semigroup theory may be relevant in ascertaining the mathematical structure of phenomenological experience. Semigroup measures such as the group complexity act at the level of the abstract transformation semigroup (corresponding to an abstract DFA) and are

therefore invariant with respect to changes in the causal architectures and/or material properties that instantiate a given computation (Rhodes *et al.*, 2010). In addition, such measures would capture much of the intuitive notions associating consciousness with complexity (such is at the heart of IIT (Tononi and Edelman, 1998)). However, it is not clear if such a framework could avoid ultimately issues of falsification. For that, new ideas about theories of consciousness may be needed which must be directed not at measuring experience itself, but instead whether or not the physical act of having a conscious experience leads to meaningful causal differences in the implementation of computations across different levels of abstraction and implementation.

4.5 Methods

4.5.1 Isomorphic Unfolding via Preserved Partitions

The Krohn-Rhodes theorem guarantees that any finite-state transition diagram can be "unfolded" such that the resultant causal architecture is feedback-free and has $\Phi = 0$. Typically, however, this unfolding process results in a causal architecture that is much larger than the minimum number of bits to instantiate the functional topology of the system using feedback. In other words, Krohn-Rhodes decomposition, and other unfolding methodologies (Oizumi *et al.*, 2014; Doerig *et al.*, 2019), inevitably result in a clear difference in efficiency between feed-forward and recurrent representations of the same underlying computation. To control for this, we must find a system that allows an *isomorphic* feed-forward representation, which can be done using a nested sequence of preserved partitions.

A preserved partition is a way of grouping microscopic states into macroscopic equivalence classes (blocks) based on symmetries present in dynamics. In particular, a partition P is preserved if it breaks the microscopic state space S into a set of blocks

$P = \{B_1, B_2, ..., B_N\}$ such that every microstate within a given block transitions to the same macrostate (i.e. the same block) (Hartmanis, 1966; Zeiger, 1968). If we denote the underlying microscopic dynamics as a function $f : S \to S$, then a block B_i is preserved when:

$$\exists j \in \{1, 2, ..., N\} \text{ such that } f(x) \in B_j \forall x \in B_i$$

In other words, for B_i to be preserved, $\forall x$ in B_i x must transition to some state in a single block B_j ($i = j$ is allowed). Conversely, B_i is *not* preserved if there exist two or more states in B_i that transition to different blocks (i.e. $\exists\ x_1, x_2 \in B_i$ such that $f(x_1) = B_j$ and $f(x_2) = B_k$ with $j \neq k$). In order for the entire partition P_i to be preserved, each block within the partition must be preserved.

For an isomorphic cascade decomposition to exist, we must be able to heirarchically construct preserved partitions in a maximally efficient way. Namely, each partition in the nested sequence of preserved partitions ($\{P_1, P_2, ..., P_N\}$) must consist of blocks that evenly split the blocks in the partition above it in half. If this is the case, then a single bit of information can be used to specify where in the preceding block the current state is located. This, in turn, allows a straightforward mapping from the blocks of the preserved partition P_i onto the first i binary coordinates used to represent these blocks. Thus, a system with 2^n microstates requires only n binary components, meaning the representation is maximally compact. If one cannot find a preserved partition made of disjoint blocks or the blocks of a given partition do not evenly split the blocks of the partition above it in half, then the system in question does not allow an isomorphic feed-forward decomposition and traditional Krohn-Rhodes decomposition techniques (Zeiger, 1968; Egri-Nagy and Nehaniv, 2008, 2015) must be employed.

To isomorphically decompose the finite-state automaton shown in Figure 4.2b,

we let our first preserved partition be $P_1 = \{B_0, B_1\}$ with $B_0 = \{A, C, E, G\}$ and $B_1 = \{B, D, F, H\}$. It is easy to check that this partition is preserved, as one can verify that every element in B_0 transitions to an element in B_1 and every element in B_1 transitions to an element in B_0 (shown topologically in Figure 4.8). To keep track of the blocks, we assign all the states in B_0 a binary coordinate value of $Q_1' = 0$ and all the states in B_1 a binary coordinate value of $Q_1' = 1$, which serves as the first of the three binary components $(Q_1'Q_2'Q_3')$ assigned to represent the global state of the system. The logic of the first coordinate is given by the corresponding state transitions of the blocks in P_1. Since block 0 goes to 1 and vis versa, the first component is essentially a `NOT` gate taking input from itself, or a JK flip-flop receiving a "toggle" signal.

The second preserved partition P_2 must evenly split each block within P_1, such that every block in P_2 is half the size of the blocks in P_1. Denoting $P_2 = \{\{B_{00}, B_{01}\}, \{B_{10}, B_{11}\}\}$, we let $B_{00} = \{A, E\}$, $B_{01} = \{C, G\}$, $B_{10} = \{B, F\}$, and $B_{11} = \{D, H\}$. One can quickly check that these blocks are indeed preserved, and that the component logic for Q_2' (based on the state of $Q_1'Q_2'$) is given by: $\{00 \to 0; 01 \to 1; 10 \to 1; 11 \to 0\}$. In a single-channel input scheme, this corresponds to Q_2' as an `XOR` gate (i.e. $Q_2' = Q_1' \oplus Q_2'$) but, again, the two-channel logic corresponding to a JK flip-flop will differ slightly.

The third and final partition P_3 must also split the blocks of P_2 in half, which implies each of the eight states corresponds to its own block in P_3. Naturally, this partition is preserved since there is only a single state in each block (making it impossible for two states within a given block to transition to separate blocks). Since P_3 is at the bottom of the hierarchy, the state of Q_3' can depend on the global state of the system $(Q_1'Q_2'Q_3')$. Unlike the previous two coordinates, this truth table is too large to be captured with a single elementary logic gate (e.g. `NOT`,`XOR`,etc.).

Instead, we must rely on a combination of elementary logic gates, which is drastically simplified by the use of JK flip-flops. Indeed, it is this third coordinate (and the potential for more complicated logical descriptions in general) that motivated our use of two-channel flip-flops rather than single-channel devices (e.g. D flip-flops). Reading the block transitions off of the bottom of Figure 4.8, we have $\{000 \to 0; 001 \to 0; 010 \to 1; 011 \to 1; 100 \to 0; 101 \to 0; 110 \to 1; 111 \to 1\}$. Clearly, there is no single binary logic gate that implements this truth table, and we must instead refer to the Karnaugh maps shown in Figure 4.4b.

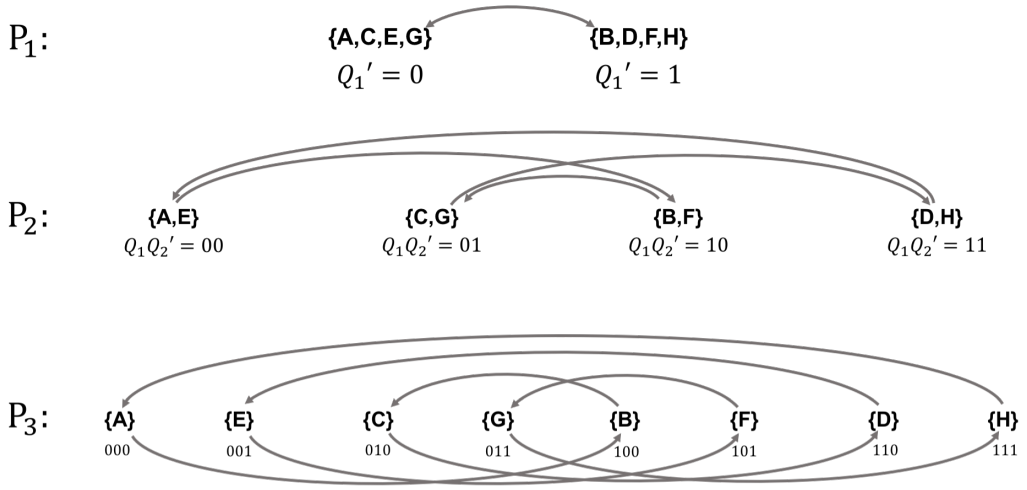


Figure 4.8: A Nested Sequence of Preserved Partitions $\{P_1, P_2, P_3\}$ Used to Isomorphically Decompose (Unfold) the Dynamics Underlying the Finite-state Description of the Tollbooth Shown in Figure 4.2. Blocks Within Any given Partition Transition Deterministically, Which Implies the Logic for Individual Components Can Be Constructed Hierarchically. The Binary Labels Assigned to the Blocks of P_3 Correspond to a Labeling Scheme That Is Isomorphic to the Original and Strictly Feed-forward (See Main).

At this point, the isomorphic cascade decomposition is complete. The values assigned to the blocks of Q_3 correspond to our new binary labeling scheme, namely:

$$A = 000, B = 100, C = 010, D = 110, E = 001, F = 101, G = 011, H = 111$$

As demonstrated in the main text, these labels result in a causal architecture that is strictly feed-forward and has $\Phi = 0$ for all states, as desired. This can easily be seen by the fact that the transitions of blocks in any given level of the nested sequence of preserved partitions are fully deterministic without the need to specify lower levels (Figure 4.8). Thus, downstream information from later coordinates is inconsequential to the action of earlier coordinates, which enforces the "hierarchical" relationship between components. Note, this result is by no means unique; there are other nested sequences of preserved partitions for this system that are equally valid. Choosing a different nested sequence of preserved partitions simply amounts to changing the labels assigned to each block which, in turn, changes the Boolean logic governing the system. As long as the partitions are preserved, however, the causal architecture that results is guaranteed to be strictly feed-forward and isomorphic to the logical architecture we present.

4.6 Acknowledgements

We thank the Emergence@ASU team for helpful feedback on this work as well as two anonymous reviewers whose comments significantly improved the manuscript.

Chapter 5

FALSIFICATION OF MACHINE STATE FUNCTIONALISM VIA AN UNFOLDING ARGUMENT

5.1 Abstract

Recently postulated mathematical measures of consciousness have generated an intense debate regarding the epistemic underpinnings required for a theory of consciousness to be considered scientific. In particular, the "unfolding argument" proved that causal structure theories of consciousness such as Integrated Information Theory (IIT) must either be false or outside the realm of science, as it is possible to vary predictions within the theory without changing the input-output behavior of a given system which is used to infer conscious experience. Here, we apply this same line of reasoning to machine state functionalist theories of consciousness, a much broader class of theories that subsumes causal structure theories. We prove that machine state functionalism is subject to its own version of the unfolding argument, in which machine states are allowed to vary under fixed input-output conditions. If consciousness must be inferred at the level of input-output behavior, our results imply machine state functionalism as a whole is either falsified or outside the realm of science in a way completely analogous to causal structure theories.

5.2 Introduction

The turn of the twenty-first century saw a shift in the scientific study of consciousness from *what* a system does to *how* it does it. This shift was brought about in large part due to advances in medical imaging such as electroencephalographs (EEG)

and functional magnetic resonance imaging (fMRI), which allow one to see the inner workings of the brain in a way previously unaccessible. This, in turn, brought about the rise of so called "causal structure theories" of consciousness, which posit that certain logical architectures are responsible for the subjective feeling of consciousness, in absence of any particular (outward) functional consequences.

The most famous causal structure theory is Integrated Information Theory (IIT), which claims to derive the requisite causal structures responsible for consciousness directly from first-hand phenomenological axioms (Tononi, 2004, 2008; Oizumi *et al.*, 2014). This "derivation" involves the translation from phenomenological axioms to physical postulates thought to embody these axioms. For example, the phenomenological experience of consciousness as a unified whole (the integration axiom) is born out by the assumption that physical feedback between elementary components in a system is a necessary condition for instantiating such an experience (the integration postulate). In this way, a correspondence between phenomenological experience and the physical substrate of consciousness is built - therein solving the hard problem of consciousness (Chalmers, 1995).

Crucially, the translation from phenomenological axioms to physical postulates is not inherently objective. Indeed, it is well known that the mathematical measure of consciousness - Φ - specified by IIT is just one of many possible measures consistent with the axioms of the theory (Barrett and Mediano, 2019). Given this, *justification* for a given set of necessary and sufficient physical conditions for consciousness is of the utmost importance, as it is possible to agree on the phenomenological underpinnings of a given theory of consciousness without agreeing on the qualitative or quantitative predictions from the theory.

Such epistemological issues have recently come to a head in a debate over the "unfolding argument" - a claim that all causal structure theories (most notably IIT) are

either (a) falsified or (b) unfalsifiable (Doerig *et al.*, 2019; Hanson and Walker, 2020). The basis of the unfolding argument is that it is always possible to fix the input-output behavior of a given system while allowing the internal causal structure to vary. Since causal structure theories posit that it is *how* a system computes rather than *what* a system computes, the ability to vary the former while fixing the latter successfully decouples predictions of the theory from the outward behavior of the system. Logically, this decoupling is problematic as one can no longer rely on behavioral states such as being asleep to benchmark predictions from the theory. In other words, one can no longer test the predictions from the theory against our common-sense notion that certain behaviors such as being asleep correspond to certain conscious states, as the same behavior is consistent with the opposite prediction from the theory. Therefore, one must choose between believing that behavior is an accurate reflection of subjective experience, in which case causal structure theories such as IIT are falsified, or believing the predictions from the theory, in which case predictions from the theory can never be falsified by third-party information.

The unfolding argument sheds light on a hitherto unformalized issue in the scientific study of consciousness. Namely, the importance of comparing predictions from a theory of consciousness to something *objective* and *independent* from the theory itself. In scientific fields other than consciousness, the objective benchmark to which to compare predictions from the theory is rarely the subject of debate, as the ontological status of the empirical data is usually unambiguous. For consciousness, however, one can present the same empirical observations to two different parties and receive contradictory answers as to whether or not the system is conscious. In other words, a theory of consciousness is required to *interpret* whether or not a system is conscious - rendering it impossible to compare the predictions from a theory to something objective. In short, the scientific study of consciousness is theory-laden, making it difficult

to ground any theory in the traditional scientific notion of falsifiability.

A general formalization of the epistemic issues surrounding falsification and consciousness was recently presented by Kleiner and Hoel (2020). In this formalism, falsification is defined as a mismatch between a prediction from a theory and the results from an (independent) inference procedure. Ideally, the inference procedure is objective, in the sense that falsification results when there is a mismatch between what a theory predicts and what is objectively true. But, as mentioned, the inference procedure itself requires some theory of consciousness, which often couples inference and prediction into a single procedure - rendering the theory inherently unfalsifiable. In light of this, the falsification side of the unfolding argument can be recast in terms of a mismatch between the predictions of causal structure theories and an inference procedure based on the input-output behavior of the system: if one assumes behavior is reflective of subjective experience (e.g. sleep is indicative of a lower subjective experience than being awake) then the ability to vary the causal structure without affecting the input-output behavior of the system implies experimental falsification of the theory. Conversely, if one uses the inner workings of the system to *infer* whether or not it is conscious, rather than the input-output behavior, then the inference procedure is one and the same with the prediction procedure, resulting in an inability to falsify the theory. Thus, even with the epistemic issue of inference being theory-laden, the unfolding argument serves as a way to prove when a theory of consciousness is no longer scientifically viable by considering all possible inference procedures in turn Hanson and Walker (2020)

Given the utility of the unfolding argument, it is important to understand the scope of its validity. In this work, we examine whether the unfolding argument can be applied to a broader class of consciousness theories beyond causal structure theories. In particular, we focus on machine state functionalist theories of consciousness, which

subsume causal structure theories of consciousness. We find that, like causal theories, machine state functionalist theories are subject to their own version of the unfolding argument, wherein one can logically prove that machine state functionalism is either (a) falsified or (b) unfalsifiable, depending on the inference procedure being used.

5.3 Preliminaries

5.3.1 Defining Machine State Functionalism

Machine state (MS) functionalist theories are based on the idea that conscious states are in one-to-one correspondence with machine states Putnam (1960, 1992). In what follows, we consider machine state functionalist theories based on a deterministic finite-state automaton (DFA) descriptions of a system, with the understanding that all MS functionalist formalisms suffer from the issues that we prove using DFAs (see Discussion). DFA descriptions are natural MS models of both brains and circuits, as they formalize the relationship between inputs, outputs, *and internal states* - a feature central to the idea of functionalism. Mathematically, a DFA A is defined by the tuple $\{S, \Sigma, \delta\}$, where S is a set of internal states, Σ is a set of inputs (known as the alphabet), $\delta : S \times \Sigma \rightarrow S$ is a function that specifies how the system transition through internal states based on the input $\alpha \in \Sigma$ and current state $s \in S$.

There are several different brands of MS functionalism, depending on where the boundary is drawn between *abstract* machine states and *physical* inputs and outputs (Figure 5.1). In the most conservative case, MS functionalist theories predict that conscious states supervene on the abstract automaton description of a system in isolation. We will refer to this class of theories as **Type 1 MS Functionalism** and define it as all theories whose prediction function *pred* can be written strictly as a function of $A = \{S, \Sigma, \delta\}$ - i.e. $pred(A)$. Note, causal structure theories such as IIT are subsumed

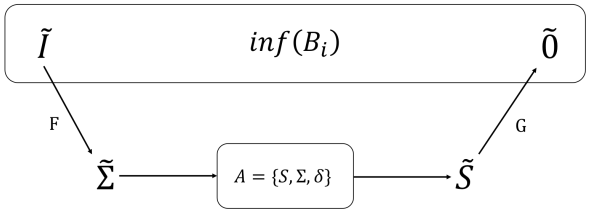
by this brand of MS functionalism, as a transition probability matrix (the mathematical object which is used to make predictions in IIT) is nothing more than a DFA with a single input representing the passage of time. **Type 2 MS Functionalism** takes into account the *temporal sequence* of abstract inputs $\tilde{\Sigma} = (\Sigma(t_0), \Sigma(t_1), ...)$ that is fed into the stationary automaton description A and the corresponding temporal sequence of machine states $\tilde{S}$ - i.e. $pred(\tilde{\Sigma}, A, \tilde{S})$. The difference between Type 1 and Type 2 theories is entirely whether or not *specific trajectories* through internal states are relevant in predicting conscious states, meaning the former is based off of a static description of the automaton while the latter accounts for the temporal sequence of internal state transitions. **Type 3 MS Functionalism** adds an additional layer of physicality by accounting for the relationship between *physical* inputs/outputs and *abstract* machine states. If we denote the sequence of physical inputs as $\tilde{I}$ and the sequence of physical outputs as $\tilde{O}$, then Type 3 functionalist theories are those that explicitly include the maps $F : I \to \Sigma$ and $G : S \to O$ in the static description of the machine - i.e. $pred(F, A, G)$. While it is possible to go beyond Type 3 theories and include the temporal sequence of physical inputs and outputs themselves (i.e. $\tilde{I}$ and $\tilde{O}$), such a theory is no longer under the umbrella of MS functionalism, as the prediction function can depend entirely on physical inputs and behavioral outputs without explicitly accounting for internal machine states.

5.3.2 Defining Falsification

A formalization of *falsification* arguments for theories of consciousness has recently been provided by Kleiner and Hoel (2020). In their formalism, falsification is defined based on a set of observations, O, as a mismatch between what a theory predicts, $pred(O)$, and the state of consciousness that can be inferred $inf(O)$ from the observations. It is always assumed that output behavior (e.g. self-report) is the

Classification	Definition
Type 1 M/S Functionalism	$pred(A)$
Type 2 M/S Functionalism	$pred(\tilde{\Sigma}, A, \tilde{S})$
Type 3 M/S Functionalism	$pred(F, A, G)$

(a)



(b)

Figure 5.1: MS Functionalist Theories Can Be Broken down into Three Different Types, According to Where the Boundary Is Drawn Between Machine States and Physical Inputs/Outputs (a). In All Cases, the Inference Procedure Is Dictated by a Specific Sequence of Physical Input-output Behavior (b).

basis for inference, as this is the only empirical data that can be uncontroversially grounded in phenomenology (Doerig *et al.*, 2020). Conversely, *internal data* such as functional magnetic resonance imaging (fMRI) is often what is used to make predictions within the theory. Thus, the goal for a successful theory of consciousness is to make predictions based on internal data that are in agreement with the conscious states we infer based on output behavior. Likewise, if one can demonstrate a case in which $pred(O) \neq inf(O)$ then the theory is considered falsified. Of course, one can always insist that the predictions from the theory are correct and the results from the inference procedures are incorrect but, in doing so, the theory becomes inherently unfalsifiable and strictly metaphysical (Doerig *et al.*, 2019; Kleiner and Hoel, 2020).

In what follows, we assume inference is based on output behavior, as is the case for the standard unfolding argument (Doerig *et al.*, 2019). To formalize this notion, we imagine a set of behaviors $\{B\} = \{B_0, B_1, ..., B_N\}$ that map to a scalar value of consciousness via an inference procedure: $inf : \{B\} \to \mathbb{R}$. For clarity, we treat $\{B\}$ as a totally ordered set, meaning that $i < j$ implies $inf(B_i) < inf(B_j)$. Qualitatively, this corresponds to the notion that we can infer when a behavior B_j is "more conscious"

than a behavior B_i (e.g. being awake corresponds to a higher level of consciousness than being asleep. Furthermore, we assume that any system capable of performing behavior B_j is also capable of performing behavior B_i. In terms of abstract computation, this assumption is natural, as higher computational complexity can easily emulate lower computational complexity via a homomorphism (many-to-one map) between computational states. When sensori-motor hardware is considered, however, this assumption makes less sense as the ability to perform a given behavior depends strongly on the sensorimotor capabilities of the system in question. Fortunately, MS functionalism is concerned primarily with the computational capabilities of a system, rather than its sensorimotor functionality, and therefore the assumption of a total ordering is justified. For clarity, one can imagine that the sensorimotor hardware of the systems in question is fixed, such that behavioral capability is dictated strictly by computational abilities.

5.4 Results

With falsification and the various types of machine state functionalism formally defined, we can now state our main results.

Theorem 1. *Type 1 MS Functionalist theories are falsified (or unfalsifiable).*

Proof. Consider an automaton A capable of generating behavior B_i and an automaton A' capable of generating behavior up to B_j with $j > i$. When A' emulates A, it must be prescribed the same prediction, as inference is fixed based on the input-output behavior of the system. So $pred(A) = pred(A')$ if the theory is to avoid falsification. Conversely, when A' goes beyond A, we have $inf(A') > inf(A)$ since $B_j > B_i$, in which case $pred(A') > pred(A)$ if the theory is to avoid falsification. This implies a contradiction wherein $pred(A') = pred(A)$ when realizing B_i but $pred(A') \neq pred(A)$

when realizing B_j. Since the DFA description of A and A' is unchanged, a measure of consciousness cannot make predictions based solely on the DFA description of a system, hence Type 1 theories are falsified. □

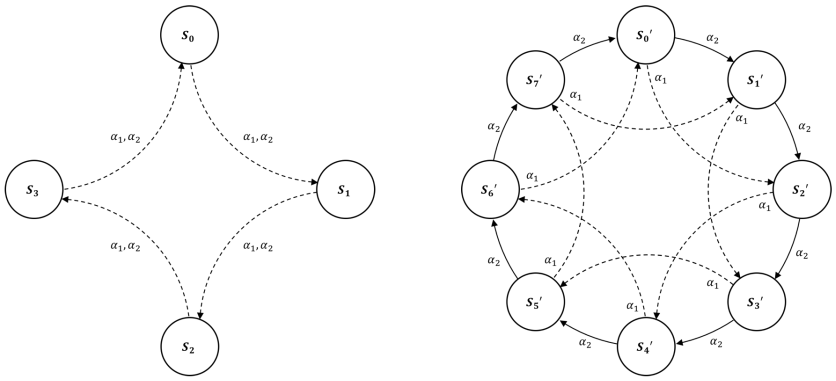


Figure 5.2: Falsification of Type 1 MS Functionalism via Emulation. Here, the Automaton on the Right Is Capable of Emulating the Behavior of the Automaton on the Left, Depending on the Input Sequence Fed into the Machine. Certain Input Sequences Lead to $inf(A) = inf(A')$, While Others Lead to $inf(A) \neq inf(A')$. Consequently, Type 1 MS Functionalist Theories Are Falsified, as the Prediction Function Is Fixed but the Results from the Inference Procedure Are Allowed to Vary.

As concrete demonstration of this proof, consider the two automata shown in Figure 5.2. On the left, the automaton A is designed to count mod-four while, on the right, the automaton A' counts mod-four or mod-eight, depending on the input sequence $\tilde{\Sigma}$. If we assume behavior B_i corresponds to counting mod-four while behavior B_j corresponds to counting mod-eight then, under input sequence $\tilde{\Sigma}_i = \{\alpha_1, \alpha_1, \alpha_1, ..., \}$, we have $inf(A) = inf(A')$ so $pred(A) = pred(A')$. But, under input sequence $\tilde{\Sigma}_j = \{\alpha_2, \alpha_2, \alpha_2, ..., \}$, we have $inf(A) \neq inf(A')$, as A is counting mod-four while A' is counting mod-eight. Thus, under $\tilde{\Sigma}_j$ we must have $pred(A) \neq pred(A')$. However, the input sequence is not explicitly considered by Type 1 MS functionalist

theories, as only the static automaton ("machine table") description $A = \{S, \Sigma, \delta\}$ is relevant. Since this description is fixed for both B_i and B_j, Type 1 theories are falsified. Granted, counting mod-eight is not a behavior we typically associate with a particular subjective experience, but the same principle applies regardless of the complexity of the behavior in question.

What changes in going from B_i to B_j is not the machine-state description, but the sequence of inputs that is fed into it. Since the input sequence is not part of the machine table description of the DFA, a measure of consciousness that acts solely on the DFA description is invariant with respect to such changes and, consequently, falsified. This suggests that it is the *trajectory* that a system takes through internal states that is relevant when predicting consciousness, which is Type 2 MS functionalism. However, Type 2 MS functionalism is also falsified, as the following theorem proves.

Theorem 2. *Type 2 MS Functionalist theories are falsified (or unfalsifiable).*

Proof. Consider a single automaton A executing behavior B_i under input sequence $\tilde{\Sigma}$. Now, consider the same automaton under an intervention † that disconnects the automaton from its motor output, such that $G_\dagger : S \to \emptyset$. Under this intervention, $A_\dagger$ generates the trivial behavior B_0, as the sequence of abstract machine states $\tilde{S}$ is mapped onto the empty set of behavior. Yet, the prediction function $pred(\tilde{\Sigma}, A, \tilde{S})$ is invariant with respect to this intervention, as all inputs to the prediction function take place prior to the movement of G (Figure 5.3). Thus, the intervention $G \to G_\dagger$ changes the results from the inference procedure while leaving the results from the prediction function intact, which implies falsification of Type 2 MS functionalism. □

Theorem 2 suggests that MS functionalist theories must include not only the sequence of abstract inputs and outputs but also the maps that encode/decode physical

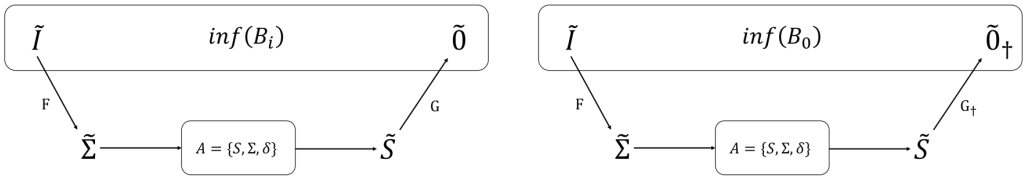


Figure 5.3: Falsification of Type 2 MS Functionalism via an Intervention $G \rightarrow G_\dagger$ That Disconnects Machines from Their Motor Hardware. Under This Intervention, the Results from the Inference Procedure Change While the Prediction Function Remains Fixed, Implying Falsification of Type 2 MS Functionalism.

stimuli and responses - i.e. Type 3 MS functionalism. Again, however, this generalization is not sufficient to save MS functionalism from falsification, as Theorem 3 proves.

Theorem 3. *Type 3 MS Functionalist theories are falsified (or unfalsifiable).*

Proof. Consider an automaton A capable of generating behavior B_j under the sequence of physical inputs $\tilde{I}_j$. Since $\{B\}$ is a total ordering, any such system can also generate behavior B_k $(k < j)$ under the sequence of physical inputs $\tilde{I}_k$. Thus, under $\tilde{I}_j$ we must have $pred(A) = inf(B_j)$ to avoid falsification while under $\tilde{I}_k$ we must have $pred(A) = inf(B_k)$. By definition, Type 3 MS functionalism is invariant with respect to the physical input sequence $\tilde{I}$, thus, $pred(A)$ is fixed for both $\tilde{I}_k$ and $\tilde{I}_j$. However, the results from the inference procedure have changed as a consequence of the physical inputs, since $inf(B_j) \neq inf(B_k)$. Thus, the results from the inference procedure have changed while the prediction function remained fixed, implying falsification of Type 3 theories (Figure 5.4). □

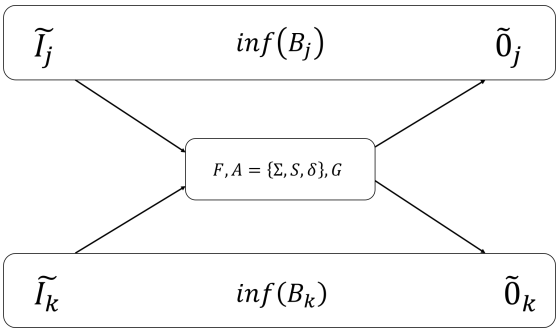


Figure 5.4: Falsification of Type 3 MS Functionalism Occurs by Fixing the Internal Circuitry (A) and Sensorimotor Hardware (F, G) of a System and Allowing the Physical Input Sequence to Vary. Under $\tilde{I}_j$, the System Generates Output Behavior $\tilde{O}_j$ with Inference Contents $inf(B_j)$ While under $\tilde{I}_k$, the System Generates Output Behavior $\tilde{O}_k$ with Inference Contents $inf(B_k) \neq inf(B_j)$. In Both Cases, the Prediction Function ($pred(F, A, G)$) Is Fixed, Implying Falsification of Type 3 MS Functionalism.

5.5 Discussion

The second half of the twentieth century saw both the rise and fall of machine state functionalism (Shagrir, 2005). Indeed, Hilary Putnam, one of the early proponents of functionalism, was responsible in large part for its demise. In 1988, he put forth an argument against functionalism based on the idea that many different functional topologies can realize the same input-output behavior (Putnam, 1988). However, this argument was not universally accepted due to the fact that it was unclear where boundaries were being drawn between the abstract logical properties of a system (software) and their concrete, physical instantiation (hardware) (Chalmers, 1996).

Our results can be seen as a return to this early work, with the contemporary knowledge that clearly defined inference and prediction procedures are of the utmost importance. In addition to defining falsification, we explicitly treat the encoding process by which physical inputs are transformed into machine state inputs (and

likewise for motor outputs). Therefore, we have a well-defined mathematical notion of the interface between "internal" and "external" properties of the system in terms of the maps F and G. Using this, we can define non-functionalist theories as those that make predictions based on physical input-output behavior (i.e. $pred(\tilde{I}, \tilde{O})$), while functionalist theories are those that explicitly depend on the internal machine-state description (i.e. $pred(\tilde{I}, A, \tilde{O})$). Our main conclusion is that any theory of consciousness must be invariant with respect to changes in A that leave $(\tilde{I}, \tilde{O})$ fixed if it is to avoid *a priori* falsification via the unfolding argument. In other words, *for a functionalist theory to be considered scientific it must group predictions into equivalence classes that align with the results from inference procedures based on input-output behavior*. Crucially, this is not necessarily the same as a prediction function based on input-output behavior (i.e. $pred(\tilde{I}, \tilde{O})$), but rather, the prediction function utilizes internal data in a way that independently *predicts* the results of the inference procedure. Thus, input-output behavior is not used for prediction, as is the case for behaviorist theories, but rather, as the independent benchmark used for inference.

There is a strong analogy between the formal argument we present and the unfolding argument as it applies to causal structure theories (Doerig *et al.*, 2019). In terms of causal structure theories, one fixes the functional topology of a system and allows the causal structure that realizes it to vary in a way that varies the prediction from the theory. For example, one can realize the same finite-state automaton A using a circuit with or without feedback connections (Hanson and Walker, 2020). By fixing the automaton description, one necessarily fixes the input-output behavior (though the inverse is not necessarily true) and therefore the results from the inference procedure. Any theory that makes predictions based on causal structure, therefore, is falsified. Similarly, in terms of MS functionalist theories one fixes the input-output behavior of the system and allows the MS description of the system to vary in a way

that affects the prediction from the theory. Thus, our results can be seen as an unfolding argument applied at the level of machine states rather than causal structure. In both cases, the ability to vary predictions below the level of input-output behavior is the cause for falsification, as inference is assumed to supervene on the input-output behavior of the system.

Going forward, the problem theories of consciousness must contend with is that only inference procedures based on input-output behavior are considered theory-independent. It seems paradigm cases, and all other empirical data to be explained by a theory of consciousness, are at the level of input-output behavior (Doerig *et al.*, 2020). In light of this, it is no surprise that a theory must be fixed below the level of input-output behavior if it is to avoid falsification. In fact, this precise conclusion was reached by Turing (1950), who argued that behaviorally indistinguishable systems must be prescribed the same subjective experience on the basis that there is no empirical differences below the level of input-output behavior that can be used to *justify* a difference in subjective experience. This means any theory that resolves differences in subjective experience based on something other than input-output behavior is ultimately doing so without proper justification and, consequently, is metaphysical (Turing, 1950). That Turing reached this conclusion seven decades before the current formalization is perhaps a testament to the depth at which he understood this issue. The notion of Turing-indistinguishability is not a simplification of the problem of consciousness, but rather, a scientific requirement.

As has historically been the case when it comes to the study of consciousness, it is much easier to prove what consciousness isn't rather than what it is. Here, we have shown that the unfolding arguments that undermine causal structure theories apply equally well to machine state functionalist theories. We hope that in building this analogy it is easier to understand the epistemic issues that surround the scientific

study of consciousness and, in particular, the crucial role that falsification plays. In addition, we stress that the inability to falsify theories based on anything other than input-output behavior has serious consequences in terms of the resolution at which one can make theoretical predictions. While we can continue to build increasingly complex theories of consciousness, they will always be limited scientifically by what we can ground experimentally using independent inference procedures.

Chapter 6

CONCLUSION

6.1 Summary of Results

This dissertation was motivated by the search for life elsewhere in the universe. We began with a brief discussion of the history of exoplanets, with an emphasis on the capability for telescopes in the near future to remotely detect biosignatures. We then briefly reviewed the most promising biosignatures, of which oxygen gas (O2) is the preeminent candidate. However, there is a substantial body of literature devoted to understanding the abiotic processes by which biosignatures such as O2 can be produced. In light of this, it is extremely difficult to tell a priori what the likelihood of false positives relative to true positives will be. To address this problem quantitatively requires Bayesian inference, in which the probability of a true biosignature is mathematically compared to that of a false positive. Crucial to this quantitative analysis is the marginal distribution $p(\text{life})$ corresponding to the prior probability that life emerges in a given planetary context. At this point, $p(\text{life})$ is completely unconstrained which implies the detection of a biosignature will be statistically inconclusive in all but the most straightforward of cases.

To remedy this problem requires taking a step back and examining the state of knowledge surrounding life's origin. In this regard, the connection between information and life is pertinent, as information seems to play a vital role in differentiating living and non-living processes. To investigate this hypothesis further, we sought a quantitative measure of information processing that could be applied to dynamical systems in a way that ideally recovered our intuitive understanding of living and non-

living systems. There were several candidate measures that could be used to test this hypothesis, but Integrated Information Theory (IIT) was by far the most promising. The reason for this was because IIT provides an out-of-the-box mathematical measure - Φ - that is purportedly one and the same with conscious experience. Since consciousness is a sufficient condition for life, Φ serves as a quantitative measure of life in addition to consciousness. Also, the strong mind-body continuity thesis suggests that the organizational principles of the mind are nothing more than a distilled and enhanced version of the organizational principles of life. Under this hypothesis, a measure of consciousness *is* a measure of life.

In Chapter 2, we studied the mathematical formalism used to calculate Φ in practice. Surprisingly, we found that despite its widespread usage in contemporary neuroscience, IIT is not a well-defined mathematical theory. The reason for this was due to the fact that deep within the optimization routine used to calculate Φ there are unresolved degeneracies. These degeneracies are the result of a procedure that requires the user to select the minimum of local ϕ ("little phi") values which, in practice, is not guaranteed to be unique. Depending on which ϕ value is chosen, the "core cause" and "core effect" corresponding to the ϕ value changes. This, in turn, causes the global integrated information Φ ("big Phi") to change. Since Φ is the predicted value of consciousness for a given system, failure to resolve these degeneracies can amount to non-unique and non-specific predictions. To demonstrate this, we attempted to calculate Φ for a simple `AND+OR` logic gate system and found 83 different Φ values resulted - spanning virtually the entire range of possibilities and including both conscious and unconscious predictions. To further investigate the scope of this problem, we created a Python package called `PyPhi-Spectrum` that can be used to calculate the entire spectrum of Φ values for a given system with a single function call. We applied `PyPhi-Spectrum` to a corpus comprised of ten recently

published Φ values and found that, of the ten, only one of the published Φ values was actually unique. The rest were selected arbitrarily for a host of alternatives, which heavily impacts the validity of their results.

In Chapter 3, we left the mathematical problems associated with Φ aside and turned instead to the epistemic foundations of the theory. IIT is presented as a "phenomenology first" theory, meaning that it starts with phenomenological axioms of what it is like to be conscious and, from these, it "derives" a mathematical measure of consciousness. In particular, the integration *axiom* states that consciousness is experienced as a single "unified whole", which IIT then translates into the physical *postulate* that feedback between elements (e.g. neurons) is a necessary condition for consciousness. However, the Krohn-Rhodes theorem from automata theory states that any system with feedback can be perfectly emulated by a strictly feed-forward system. In the context of IIT, we showed that this implies every conscious system with $\Phi > 0$ has a feed-forward emulator with $\Phi = 0$ that is functionally indistinguishable from the original (i.e. a "philosophical zombie"). We then proved that the *homomorphism* guaranteed by the Krohn-Rhodes Theorem can also be an *isomorphism*, meaning the only difference between conscious and unconscious systems in IIT is a permutation of the binary labels used to represent functionally states. Since labels are arbitrary, we argued IIT fails to justify the assumption that feedback is a necessary condition for consciousness, instead simply assuming it to be true.

In Chapter 4, we sought a concrete demonstration of the epistemic problem presented in Chapter 3. For this, we turned to a simple digital counter designed to operate an electronic tollbooth. The first step in the design process was to choose the labels that are assigned to implement the functional states of the system. For this, we first chose labels arbitrarily and showed that a circuit with feedback and $\Phi > 0$ resulted (i.e. a conscious tollbooth). Then, we performed an isomorphic cas-

cade decomposition using the methodology we developed in Chapter 3 and showed that the result is a functionally indistinguishable circuit with $\Phi = 0$ (i.e. an unconscious tollbooth). Since both systems operated under the same resource constraints and instantiated the exact same computation, the only way to justify a difference in subjective experience was by assuming the integration postulate is true. Yet, in absence of functional consequences *this assumption cannot be tested* which implies IIT is either falsified by our example or outside of the realm of science.

In Chapter 5, we generalized our arguments beyond the limited scope of IIT. In particular, we focused on machine-state (MS) functionalism which is a broad class of theories that subsumes causal structure theories of consciousness such as IIT. We first created a taxonomy of different MS functionalist theories before proceeding to prove each type falsified/unfalsifiable in turn. The basis for these proofs was the same as the argument we previously applied to IIT, namely, the idea that inference procedures used to independently test theories of consciousness take place at the level of input-output behavior but MS functionalism makes predictions that can vary under fixed input-output conditions. Put differently, MS functionalist theories allow philosophical zombies in a way completely analogous to causal structure theories and are therefore subject to the same epistemological concerns. These issues with MS functionalism were well known in the latter half of the twentieth century but were based in large part on thought experiments rather than formal mathematical proofs such as those we provide. Our arguments suggest that in absence of a mechanistic understanding of the processes that generate the behaviors we associate with consciousness, the Turing test is the only epistemologically sound metaphysical position.

6.2 Outlook: What Theories of Consciousness Teach Us about Defining Life

Both consciousness and life are "pre-theoretical" sciences, meaning they study their topic of interest in absence of a theoretical definition. Integrated Information Theory was the first convincing attempt to define consciousness in a way that connects it to empirical science. For this connection to be made, the properties we intuitively associate with consciousness must be given a causal account in terms of the processes by which they are generated. In this regard, IIT was arguably successful. The theory assumed a causal connection between the phenomenological properties we associate with consciousness, such as unified experience, and the physical mechanisms that give rise to these properties, such as feedback connections.

Unfortunately, the causal account provided by IIT failed to generate testable hypotheses. From its inception, IIT embraced "philosophical zombies" in a way that all but guaranteed the theory was metaphysical. There is a long history of arguments explicitly against theories of consciousness that admit philosophical zombies and IIT is not even superficially immune to these arguments (Harnad, 1995; Cohen and Dennett, 2011; Doerig *et al.*, 2019; Hanson and Walker, 2020). Instead, it relies on the unfalsifiable belief that their assumed causal explanation is correct. Consequently, current experiments designed to falsify or confirm IIT in a laboratory setting are ill-founded, as they are unwittingly based on metaphysical assumptions.

It is easy to forget that IIT is *the most popular theory of consciousness* in contemporary neuroscience. That a theory can reach such status without proper scientific justification is an admonishing lesson for origins of life research. Like consciousness, theories of life are trying to make the transition from metaphysical definitions to testable hypotheses. In doing so, they must explicitly make the connection between the properties we associate with life and its causal mechanisms, while simultaneously

taking care not to conflate predictions and assumptions. A canonical example of a successful transition from a pre-theoretical to a theoretical science is the definition of water (Murphy and Medin, 1985; Cleland and Chyba, 2002). As a pre-theoretical concept, water was defined by its properties, namely, that it is a colorless, odorless, liquid that boils at 100 °C. After the advent of molecular theory, however, water was identified as H2O, which not only redefined the concept but also explained why water has the properties that we associate with it, such as its boiling point. In doing so, the molecular theory of water not only connected our intuitive understanding of water with a causal account but also provided additional explanatory power in the form of predictions that could be independently tested. These predictions serve as a crucial demarcation between physical and metaphysical theories.

In contrast, current definitions of life fail to provide additional explanatory power or make concrete predictions. For example, NASA's adopted definition of life as "a self-sustaining chemical system capable of Darwinian evolution" (Joyce, 1994), fails to provide a causal mechanism that explains *why* life has the salient features we attribute to it, such as homeostasis, chemical replication, evolution, metabolism, etc. In other words, a scientific definition is more than the aggregate of its properties - it requires a causal explanation. It is virtually impossible to come up with a scientific theory based on a collection of salient features, as it is a well-established fact that categorization does not lend itself to definition (Murphy and Medin, 1985; Machery, 2012). This is not because categorization is entirely vacuous, but rather, categorization provides insufficient criteria for generating a definition with any additional explanatory power.

The next step for life as a pre-theoretical science is to connect the properties we attribute to life, such as replication and metabolism, with a causal understanding of the physical mechanisms that generate these properties. It is here that the analogy with consciousness is most productive, as many of the salient features we attribute to

life are one and the same with those we attribute to consciousness (e.g. complexity, integration, and emergence). Crucially, theories of consciousness have already tried (and failed) to turn these salient attributes into scientific theories in a way that foreshadows problems faced by theories of life. The current discussion regarding the use of complexity as a biosignature (Marshall *et al.*, 2017a), for example, mirrors early discussion regarding the use of complexity as a measure of consciousness (Tononi and Edelman, 1998). The reason complexity should not be used as a measure for life is the same reason it cannot be used as a measure of consciousness. Namely, we lack experimental evidence to ground the assumption that it is a necessary and/or sufficient condition.

To be a necessary condition, it must not be possible to have life without complexity. While this may make sense intuitively, it is impossible to test experimentally. If we assume that complexity *is* life, then whatever the measured complexity for a system is dictates our interpretation of whether or not it is alive. Conversely, if our intuition is used as an independent inference procedure to which predictions from the theory are compared then, at best, the theory recovers our intuition. In the former case, the theory is unfalsifiable, while in the latter it fails to provide additional explanatory power. With regard to sufficiency, the problem is the same. It is intuitive to assume that the presence of a salient feature such as complexity is a sufficient condition for life. However, to test this assumption empirically requires the ability to independently isolate what we call "life" so we can compare it to the predictions that complexity is sufficient for life. Thus, we still need to know *a priori* whether or not a system is alive in order to test the sufficiency condition. In addition, a single salient feature such as complexity fails to provide an explanation for any of the other properties we associate with life, such as replication or metabolism. So long as a definition fails to provide a mechanistic understanding of *why* life has the properties

it has, its use as a measure of life is scientifically unfounded. The same problem holds for any salient feature we would like to turn into a causal account.

In summary, definitions of life should abide by the following three conditions. First, a definition must contain a causal account of the properties we associate with a given concept if it is to be considered scientific. Definitions lacking a mechanistic component are scientifically devoid of meaning due to a lack of additional explanatory power. Second, the mechanistic account provided must generate testable consequences, and these predictions must be compared to observations whose interpretation is independent of the theory in question. As we demonstrated with IIT, interpreting predictions through the lens of the theory in question leads to an inherently unfalsifiable (metaphysical) theory. And last, definitions cannot be deduced from salient attributes. Using intuitive features of life such as complexity or replication to try and define the causal mechanisms at play will at best result in a theory of complexity or replication, rather than a general theory of life.

While studies of consciousness are informative in guiding studies of life, they are not the same scientific question. Consciousness has its own set of epistemological concerns surrounding the fact that consciousness is subjective, by definition, which makes it difficult to study as external observers. Life, on the other hand, is like any other empirical phenomenon we have yet to explain, such as water in the absence of molecular theory. As much as we would like to know *a priori* what steps we can take to discover the analogous theory for life, history makes it clear there is no one path forward. We may discover evidence for life on Mars or an unambiguous biosignature from a distant world. Or, perhaps we will remain alone indefinitely. At this point, what is most important is maintaining a diversity of ideas, as progress often comes from the most unlikely of sources. In this regard, we have attempted a wholly original solution to the quantification of life using a preeminent theory of consciousness from

contemporary neuroscience. Unfortunately, our trust in this theory was misguided as it proved to be based on a set of assumptions that render it metaphysical at best. However, our results highlight the fact that science as a whole is not an immutable body of knowledge, but rather, a tangled web of human design. Indeed, it is precisely the deviations from perfection that lead to progress, as ill-founded assumptions must be winnowed out in favor of those that stand the test of time. With theories of consciousness, as with theories of life, it is to be expected that there will be many false starts before a stable paradigm is established. On the surface, this may look like chaos but, in reality, it is an indispensable aspect of the scientific process.

Adams, E., S. Seager and L. Elkins-Tanton, "Ocean planet or thick atmosphere: on the mass-radius relationship for solid exoplanets with massive atmospheres", The Astrophysical Journal **673**, 2, 1160 (2008).

Aguilera, M., C. Alquézar and M. G. Bedia, "Agency and integrated information in a minimal sensorimotor model", in "Artificial Life Conference Proceedings", pp. 396–403 (MIT Press, 2018).

Albantakis, L., A. Hintze, C. Koch, C. Adami and G. Tononi, "Evolution of integrated causal structures in animats exposed to environments of increasing complexity", PLOS Computational Biology **10**, 12, 1–19, URL `https://doi.org/10.1371/journal.pcbi.1003966` (2014).

Albantakis, L., W. Marshall, E. Hoel and G. Tononi, "What caused what? a quantitative account of actual causation using dynamical causal networks", Entropy **21**, 5, 459 (2019).

Albantakis, L. and G. Tononi, "Causal composition: Structural differences among dynamically equivalent systems", Entropy **21**, 10, URL `https://www.mdpi.com/1099-4300/21/10/989` (2019).

Ananyeva, V. I., A. E. Ivanova, A. A. Venkstern, I. A. Shashkova, A. V. Yudaev, A. V. Tavrov, O. I. Korablev and J.-L. Bertaux, "Mass distribution of exoplanets considering some observation selection effects in the transit detection technique", Icarus **346**, 113773, URL `https://www.sciencedirect.com/science/article/pii/S0019103520301603` (2020).

Arsiwalla, X. D. and P. F. Verschure, "The global dynamical complexity of the human brain network", Applied network science **1**, 1, 16 (2016).

Auletta, G., G. F. Ellis and L. Jaeger, "Top-down causation by information control: from a philosophical problem to a scientific research programme", Journal of the Royal Society Interface **5**, 27, 1159–1172 (2008).

Baars, B. J., "Global workspace theory of consciousness: toward a cognitive neuroscience of human experience", Progress in brain research **150**, 45–53 (2005).

Balduzzi, D. and G. Tononi, "Integrated information in discrete dynamical systems: motivation and theoretical framework", PLoS computational biology **4**, 6, e1000091 (2008).

Barrett, A. B. and P. A. Mediano, "The phi measure of integrated information is not well-defined for general physical systems", Journal of Consciousness Studies **26**, 1-2, 11–20 (2019).

Bayne, T., "On the axiomatic foundations of the integrated information theory of consciousness", Neuroscience of consciousness **2018**, 1, niy007 (2018).

Borucki, W. J., D. Koch, G. Basri, N. Batalha, T. Brown, D. Caldwell, J. Caldwell, J. Christensen-Dalsgaard, W. D. Cochran, E. DeVore *et al.*, "Kepler planet-detection mission: introduction and first results", Science **327**, 5968, 977–980 (2010).

Borucki, W. J., D. G. Koch, G. Basri, N. Batalha, T. M. Brown, S. T. Bryson, D. Caldwell, J. Christensen-Dalsgaard, W. D. Cochran, E. DeVore, E. W. Dunham, T. N. Gautier, J. C. Geary, R. Gilliland, A. Gould, S. B. Howell, J. M. Jenkins, D. W. Latham, J. J. Lissauer, G. W. Marcy, J. Rowe, D. Sasselov, A. Boss, D. Charbonneau, D. Ciardi, L. Doyle, A. K. Dupree, E. B. Ford, J. Fortney, M. J. Holman, S. Seager, J. H. Steffen, J. Tarter, W. F. Welsh, C. Allen, L. A. Buchhave, J. L. Christiansen, B. D. Clarke, S. Das, J.-M. Désert, M. Endl, D. Fabrycky, F. Fressin, M. Haas, E. Horch, A. Howard, H. Isaacson, H. Kjeldsen, J. Kolodziejczak, C. Kulesa, J. Li, P. W. Lucas, P. Machalek, D. McCarthy, P. MacQueen, S. Meibom, T. Miquel, A. Prsa, S. N. Quinn, E. V. Quintana, D. Ragozzine, W. Sherry, A. Shporer, P. Tenenbaum, G. Torres, J. D. Twicken, J. V. Cleve, L. Walkowicz, F. C. Witteborn and M. Still, "Characteristics of planetary candidates observed by kepler. ii analysis of the first four months of data.", The Astrophysical Journal **736**, 1, 19, URL `https://doi.org/10.1088/0004-637x/736/1/19` (2011).

Brillouin, L., *Science and information theory* (Courier Corporation, 2013).

Burnham, J. F., "Scopus database: a review", Biomedical digital libraries **3**, 1, 1–8 (2006).

Carter, B., "The anthropic principle and its implications for biological evolution", Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences **310**, 1512, 347–363 (1983).

Cavanagh, J., *Sequential logic: analysis and synthesis* (CRC Press, 2018).

Cerullo, M. A., "The problem with phi: a critique of integrated information theory", PLoS Comput Biol **11**, 9, e1004286 (2015).

Chalmers, D. and K. McQueen, "Consciousness and the collapse of the wave function", Quantum Mechanics and Consciousness. New York: Oxford University Press, Forthcoming (2014).

Chalmers, D. J., "A computational foundation for the study of cognition", URL `http://cogprints.org/319/`, unpublished (1993).

Chalmers, D. J., "Facing up to the problem of consciousness", Journal of Consciousness Studies **2**, 3, 200–19 (1995).

Chalmers, D. J., "Does a rock implement every finite-state automaton?", Synthese **108**, 3, 309–333 (1996).

Clark, A., *Mindware: An introduction to the philosophy of cognitive science.* (Oxford University Press, 2000).

Cleland, C. E. and C. F. Chyba, "Defining 'life"', Origins of Life and Evolution of the Biosphere **32**, 4, 387–393 (2002).

Cohen, M. A. and D. C. Dennett, "Consciousness cannot be separated from function", Trends in cognitive sciences **15**, 8, 358–364 (2011).

Crutchfield, J. P. and D. P. Feldman, "Regularities unseen, randomness observed: Levels of entropy convergence", Chaos: An Interdisciplinary Journal of Nonlinear Science **13**, 1, 25–54 (2003).

DeDeo, S., "Effective theories for circuits and automata", Chaos: An Interdisciplinary Journal of Nonlinear Science **21**, 3, 037106 (2011).

Dehaene, S. and J.-P. Changeux, "Neural mechanisms for access to consciousness", The cognitive neurosciences **3**, 1145–58 (2004).

Desch, S., H. Hartnett, S. Kane and S. Walker, "Detectability, not habitability", in "Habitable Worlds 2017: A System Science Workshop", vol. 2042, p. 4070 (2017).

Doerig, A., A. Schurger and M. H. Herzog, "Hard criteria for empirical theories of consciousness", Cognitive Neuroscience pp. 1–22 (2020).

Doerig, A., A. Schurger, K. Hess and M. H. Herzog, "The unfolding argument: Why iit and other causal structure theories cannot explain consciousness", Consciousness and cognition **72**, 49–59 (2019).

Dolkega, K. and J. E. Dewhurst, "Fame in the predictive brain: A deflationary approach to explaining consciousness in the prediction error minimization framework", Synthese pp. 1–26 (2020).

Dupré, J., "The metaphysics of evolution", Interface focus **7**, 5, 20160148 (2017).

Dyson, F., "Disturbing the universe", New York: Basic Books Inc (1979).

Dyson, F., *Origins of life* (Cambridge University Press, 1999).

Egri-Nagy, A. and C. L. Nehaniv, "Hierarchical coordinate systems for understanding complexity and its evolution, with applications to genetic regulatory networks", Artificial Life **14**, 3, 299–312 (2008).

Egri-Nagy, A. and C. L. Nehaniv, "Computational holonomy decomposition of transformation semigroups", arXiv preprint arXiv:1508.06345 (2015).

Ellis, G., "How can physics underlie the mind", Springer2016 (2016).

Farnsworth, K. D., "An organisational systems-biology view of viruses explains why they are not alive", Biosystems **200**, 104324 (2021).

Feinberg, G. and R. Shapiro, "Life beyond earth: the intelligent earthling's guide to life in the universe.", Life beyond earth: the intelligent earthling's guide to life in the universe (1980).

Feyerabend, P., *Against method* (Verso, 1993).

Feyerabend, P. K., *Realism and instrumentalism*, vol. 1, p. 176–202 (Cambridge University Press, 1981).

Fleischaker, G. R., "Origins of life: an operational definition", Origins of Life and Evolution of the Biosphere **20**, 2, 127–137 (1990).

Fu, R., R. J. O'Connell and D. D. Sasselov, "The interior dynamics of water planets", The Astrophysical Journal **708**, 2, 1326 (2009).

Ginzburg, A., *Algebraic theory of automata* (Academic Press, 2014).

Glaser, D. M., H. E. Hartnett, S. J. Desch, C. T. Unterborn, A. Anbar, S. Buessecker, T. Fisher, S. Glaser, S. R. Kane, C. M. Lisse *et al.*, "Detectability of life using oxygen on pelagic planets and water worlds", The Astrophysical Journal **893**, 2, 163 (2020).

Godfrey-Smith, P., *Theory and reality: An introduction to the philosophy of science* (University of Chicago Press, 2009).

Golden, D., D. W. Ming, C. S. Schwandt, J. Lauer, Howard V., R. A. Socki, R. V. Morris, G. E. Lofgren and G. A. McKay, "A simple inorganic process for formation of carbonates, magnetite, and sulfides in Martian meteorite ALH84001", American Mineralogist **86**, 3, 370–375, URL `https://doi.org/10.2138/am-2001-2-321` (2001).

Gomez, J. D., W. G. Mayner, M. Beheler-Amass, G. Tononi and L. Albantakis, "Computing integrated information (ϕ) in discrete dynamical systems with multivalued elements", Entropy **23**, 1, 6 (2021).

Graham, G., "Behaviorism", in "The Stanford Encyclopedia of Philosophy", edited by E. N. Zalta (Metaphysics Research Lab, Stanford University, 2019), spring 2019 edn.

Graziano, M. S. and T. W. Webb, "The attention schema theory: a mechanistic account of subjective awareness", Frontiers in psychology **6**, 500 (2015).

Greaves, J. S., A. M. Richards, W. Bains, P. B. Rimmer, H. Sagawa, D. L. Clements, S. Seager, J. J. Petkowski, C. Sousa-Silva, S. Ranjan *et al.*, "Phosphine gas in the cloud decks of venus", Nature Astronomy pp. 1–10 (2020).

Grenfell, J. L., "A review of exoplanetary biosignatures", Physics Reports **713**, 1–17 (2017).

Griffith, R. L., J. T. Wright, J. Maldonado, M. S. Povich, S. Sigurdsson and B. Mullan, "The $\hat{G}$ infrared search for extraterrestrial civilizations with large energy supplies. iii. the reddest extended sources in wise", The Astrophysical Journal Supplement Series **217**, 2, 25 (2015).

Gutierrez, J. M. P., T. Hinkley, J. W. Taylor, K. Yanev and L. Cronin, "Evolution of oil droplets in a chemorobotic platform", Nature communications **5**, 1, 1–8 (2014).

Hanson, J. R. and S. I. Walker, "Integrated information theory and isomorphic feed-forward philosophical zombies", Entropy **21**, 11, 1073 (2019).

Hanson, J. R. and S. I. Walker, "Formalizing falsification of causal structure theories for consciousness across computational hierarchies", arXiv preprint arXiv:2006.07390 (2020).

Hanson, R., "The great filter-are we almost past it", preprint available at http://hanson. gmu. edu/greatfilter. html (1998).

Harnad, S., "Why and how we are not zombies", Journal of Consciousness Studies **1**, 164–167 (1995).

Harnad, S., "The annotation game: On turing (1950) on computing, machinery, and intelligence", in "The Turing test sourcebook: philosophical and methodological issues in the quest for the thinking computer", (Kluwer, 2006).

Hartmanis, J., *Algebraic structure theory of sequential machines (prentice-hall international series in applied mathematics)* (Prentice-Hall, Inc., 1966).

Haun, A. and G. Tononi, "Why does space feel the way it does? towards a principled account of spatial experience", Entropy **21**, 12, 1160 (2019).

Hazen, R. M., "Chance, necessity and the origins of life: a physical sciences perspective", Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **375**, 2109, 20160353 (2017).

Hobson, J. A., C. C.-H. Hong and K. J. Friston, "Virtual reality and consciousness inference in dreaming", Frontiers in psychology **5**, 1133 (2014).

Hoel, E. P., "When the map is better than the territory", Entropy **19**, 5, 188 (2017).

Hoel, E. P., L. Albantakis, W. Marshall and G. Tononi, "Can the macro beat the micro? integrated information across spatiotemporal scales", Neuroscience of Consciousness **2016**, 1 (2016).

Hohwy, J., "The predictive processing hypothesis", in "The Oxford handbook of 4E cognition", pp. 129–145 (Oxford University Press, 2018).

Hopcroft, J. E., R. Motwani and J. D. Ullman, "Automata theory, languages, and computation", International Edition **24**, 19 (2006).

Horneck, G., N. Walter, F. Westall, J. L. Grenfell, W. F. Martin, F. Gomez, S. Leuko, N. Lee, S. Onofri, K. Tsiganis *et al.*, "Astromap european astrobiology roadmap", Astrobiology **16**, 3, 201–243 (2016).

Joyce, G., "Origins of life: The central concepts, eds. dw deamer and gr fleischaker", (1994).

Joyce, J., "Bayes' Theorem", in "The Stanford Encyclopedia of Philosophy", edited by E. N. Zalta (Metaphysics Research Lab, Stanford University, 2019), spring 2019 edn.

Juel, B. E., R. Comolatti, G. Tononi and L. Albantakis, "When is an action caused from within? quantifying the causal chain leading to actions in simulated agents", in "Artificial Life Conference Proceedings", pp. 477–484 (MIT Press, 2019).

Kamminga, H., "Historical perspective: the problem of the origin of life in the context of developments in biology", Origins of Life and Evolution of the Biosphere **18**, 1, 1–11 (1988).

Karnaugh, M., "The map method for synthesis of combinational logic circuits", Transactions of the American Institute of Electrical Engineers, Part I: Communication and Electronics **72**, 5, 593–599 (1953).

Kim, H., H. B. Smith, C. Mathis, J. Raymond and S. I. Walker, "Universal scaling across biochemical networks on earth", Science advances **5**, 1, eaau0149 (2019).

Kim, H., G. Valentini, J. Hanson and S. I. Walker, "Informational architecture across non-living and living collectives", Theory in Biosciences pp. 1–17 (2021).

Kim, J., "Concepts of supervenience", in "Supervenience", pp. 37–62 (Routledge, 2017).

Kirchhoff, M. D. and T. Froese, "Where there is life there is mind: In support of a strong life-mind continuity thesis", Entropy **19**, 4, 169 (2017).

Kirk, R., *Mind and body*, vol. 11 (McGill-Queen's Press-MQUP, 2003).

Kite, E. S., M. Manga and E. Gaidos, "Geodynamics and rate of volcanism on massive earth-like planets", The Astrophysical Journal **700**, 2, 1732 (2009).

Kleiner, J., "Mathematical models of consciousness", Entropy **22**, 6, 609 (2020).

Kleiner, J. and E. Hoel, "Falsification and consciousness", arXiv preprint arXiv:2004.03541 (2020).

Kopparapu, R. K., R. Ramirez, J. F. Kasting, V. Eymet, T. D. Robinson, S. Mahadevan, R. C. Terrien, S. Domagal-Goldman, V. Meadows and R. Deshpande, "Habitable zones around main-sequence stars: new estimates", The Astrophysical Journal **765**, 2, 131 (2013).

Korpela, E. J., S. M. Sallmen and D. L. Greene, "Modeling indications of technology in planetary transit light curves—dark-side illumination", The Astrophysical Journal **809**, 2, 139 (2015).

Koshland, D. E., "The seven pillars of life", Science **295**, 5563, 2215–2216 (2002).

Krissansen-Totton, J., S. Olson and D. C. Catling, "Disequilibrium biosignatures over earth history and implications for detecting exoplanet life", Science advances **4**, 1, eaao5747 (2018).

Krohn, K. and J. Rhodes, "Algebraic theory of machines. i. prime decomposition theorem for finite semigroups and machines", Transactions of the American Mathematical Society **116**, 450–464 (1965).

Krohn, S. and D. Ostwald, "Computing integrated information", Neuroscience of consciousness **2017**, 1, nix017 (2017).

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

Lammer, H., J. Bredehöft, A. Coustenis, M. Khodachenko, L. Kaltenegger, O. Grasset, D. Prieur, F. Raulin, P. Ehrenfreund, M. Yamauchi *et al.*, "What makes a planet habitable?", The Astronomy and Astrophysics Review **17**, 2, 181–249 (2009).

Lee, C.-T. A., S. Thurner, S. Paterson and W. Cao, "The rise and fall of continental arcs: Interplays between magmatism, uplift, weathering, and climate", Earth and Planetary Science Letters **425**, 105–119 (2015).

Lin, H. W., G. G. Abad and A. Loeb, "Detecting industrial pollution in the atmospheres of earth-like exoplanets", The Astrophysical Journal Letters **792**, 1, L7 (2014).

Lizier, J. T., M. Prokopenko and A. Y. Zomaya, "Local information transfer as a spatiotemporal filter for complex systems", Physical Review E **77**, 2, 026110 (2008).

Lizier, J. T., M. Prokopenko and A. Y. Zomaya, "Local measures of information storage in complex distributed computation", Information Sciences **208**, 39–54 (2012).

Luger, R., M. Sestovic, E. Kruse, S. L. Grimm, B.-O. Demory, E. Agol, E. Bolmont, D. Fabrycky, C. S. Fernandes, V. Van Grootel *et al.*, "A seven-planet resonant chain in trappist-1", Nature Astronomy **1**, 6, 1–8 (2017).

Machery, E., "Why i stopped worrying about the definition of life... and why you should as well", Synthese **185**, 1, 145–164 (2012).

Maler, O., "A decomposition theorem for probabilistic transition systems", Theoretical Computer Science **145**, 1-2, 391–396 (1995).

Mallon, E., S. Pratt and N. Franks, "Individual and collective decision-making during nest site selection by the ant leptothorax albipennis", Behavioral Ecology and Sociobiology **50**, 4, 352–359 (2001).

Marcus, E., "Why zombies are inconceivable", Australasian Journal of Philosophy **82**, 3, 477–490 (2004).

Mariscal, C. and L. Fleming, "Why we should care about universal biology", Biological Theory **13**, 2, 121–130 (2018).

Marshall, S. M., A. R. Murray and L. Cronin, "A probabilistic framework for identifying biosignatures using pathway complexity", Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **375**, 2109, 20160342 (2017a).

Marshall, W., H. Kim, S. I. Walker, G. Tononi and L. Albantakis, "How causal analysis can reveal autonomy in models of biological systems", Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **375**, 2109, 20160358 (2017b).

Maturana, H. R. and F. J. Varela, *The tree of knowledge: The biological roots of human understanding.* (New Science Library/Shambhala Publications, 1987).

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

Mayor, M., X. Bonfils, T. Forveille, X. Delfosse, S. Udry, J.-L. Bertaux, H. Beust, F. Bouchy, C. Lovis, F. Pepe *et al.*, "The harps search for southern extra-solar planets-xviii. an earth-mass planet in the gj 581 planetary system", Astronomy & Astrophysics **507**, 1, 487–494 (2009).

McKay, D. S., E. K. Gibson, K. L. Thomas-Keprta, H. Vali, C. S. Romanek, S. J. Clemett, X. D. Chillier, C. R. Maechling and R. N. Zare, "Search for past life on mars: Possible relic biogenic activity in martian meteorite alh84001", Science **273**, 5277, 924–930 (1996).

McQueen, K. J., "Interpretation-neutral integrated information theory", Journal of Consciousness Studies **26**, 1-2, 76–106 (2019).

Meadows, V. S., "Planetary environmental signatures for habitability and life", Exoplanets pp. 259–284 (2008).

Meadows, V. S., "Reflections on o2 as a biosignature in exoplanetary atmospheres", Astrobiology **17**, 10, 1022–1052 (2017).

Meadows, V. S., C. T. Reinhard, G. N. Arney, M. N. Parenteau, E. W. Schwieterman, S. D. Domagal-Goldman, A. P. Lincowski, K. R. Stapelfeldt, H. Rauer, S. DasSarma *et al.*, "Exoplanet biosignatures: understanding oxygen as a biosignature in the context of its environment", Astrobiology **18**, 6, 630–662 (2018).

Metzinger, T., *Neural correlates of consciousness: Empirical and conceptual questions* (MIT press, 2000).

Montanes-Rodriguez, P., E. Palle, P. R. Goode and F. J. Martin-Torres, "Vegetation signature in the observed globally integrated spectrum of earth considering simultaneous cloud data: Applications for extrasolar planets", The Astrophysical Journal **651**, 1, 544–552, URL `https://doi.org/10.1086/507694` (2006).

Moon, K., "Exclusion and underdetermined qualia", Entropy **21**, 4, 405 (2019).

Moore, E. F., "Logical design of digital computers", Journal of Symbolic Logic **23**, 3, 363–365 (1958).

Murphy, G. L. and D. L. Medin, "The role of theories in conceptual coherence.", Psychological review **92**, 3, 289 (1985).

Nagel, T., "What is it like to be a bat?", Philosophical Review **83**, October, 435–50 (1974).

NASA, "President clinton statement regarding mars meteorite discovery", URL `https://www2.jpl.nasa.gov/snc/clinton.html` (1996).

Negro, N., "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** (2020).

Niizato, T., K. Sakamoto, Y.-i. Mototake, H. Murakami, T. Tomaru, T. Hoshika and T. Fukushima, "Finding continuity and discontinuity in fish schools via integrated information theory", PloS one **15**, 2, e0229573 (2020).

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

Oizumi, M., S.-i. Amari, T. Yanagawa, N. Fujii and N. Tsuchiya, "Measuring integrated information from the decoding perspective", PLoS computational biology **12**, 1, e1004654 (2016a).

Oizumi, M., N. Tsuchiya and S.-i. Amari, "Unified framework for information integration based on information geometry", Proceedings of the National Academy of Sciences **113**, 51, 14817–14822 (2016b).

Pallé, E., M. R. Z. Osorio, R. Barrena, P. Montañés-Rodríguez and E. L. Martín, "Earth's transmission spectrum from lunar eclipse observations", Nature **459**, 7248, 814–816 (2009).

Popiel, N. J., S. Khajehabdollahi, P. M. Abeyasinghe, F. Riganello, E. Nichols, A. M. Owen and A. Soddu, "The emergence of integrated information, complexity, and 'consciousness' at criticality", Entropy **22**, 3, 339 (2020).

Pratt, S. C. and D. J. Sumpter, "A tunable algorithm for collective decision-making", Proceedings of the National Academy of Sciences **103**, 43, 15906–15910 (2006).

Putnam, H., "Minds and machines", in "Dimensions of Minds", edited by S. Hook, pp. 138–164 (New York, USA: New York University Press, 1960).

Putnam, H., *Representation and reality* (MIT press, 1988).

Putnam, H., "The nature of mental states", The philosophy of mind: Classical problems/contemporary issues pp. 51–58 (1992).

Reardon, S., "Rival theories face off over brain's source of consciousness", (2019).

Rees, G., G. Kreiman and C. Koch, "Neural correlates of consciousness in humans", Nature Reviews Neuroscience **3**, 4, 261–270 (2002).

Rhodes, J., C. L. Nehaniv and M. W. Hirsch, *Applications of automata theory and algebra: via the mathematical theory of complexity to biology, physics, psychology, philosophy, and games* (World Scientific, 2010).

Roe, G. H., K. X. Whipple and J. K. Fletcher, "Feedbacks among climate, erosion, and tectonics in a critical wedge orogen", American Journal of Science **308**, 7, 815–842 (2008).

Rosenthal, D. M., "How many kinds of consciousness?", Consciousness and cognition **11**, 4, 653–665 (2002).

Rubner, Y., C. Tomasi and L. J. Guibas, "The earth mover's distance as a metric for image retrieval", International journal of computer vision **40**, 2, 99–121 (2000).

Schaefer, B. E., R. O. Bentley, T. S. Boyajian, P. H. Coker, S. Dvorak, F. Dubois, E. Erdelyi, T. Ellis, K. Graham, B. G. Harris *et al.*, "The kic 8462852 light curve from 2015.75 to 2018.18 shows a variable secular decline", Monthly Notices of the Royal Astronomical Society **481**, 2, 2235–2248 (2018).

Schneider, J., C. Dedieu, P. Le Sidaner, R. Savalle and I. Zolotukhin, "Defining and cataloging exoplanets: the exoplanet. eu database", Astronomy & Astrophysics **532**, A79 (2011).

Schreiber, T., "Measuring information transfer", Physical review letters **85**, 2, 461 (2000).

Schrödinger, E., *What is life?: With mind and matter and autobiographical sketches* (Cambridge University Press, 1992).

Schwieterman, E. W., C. S. Cockell and V. S. Meadows, "Nonphotosynthetic pigments as potential biosignatures", Astrobiology **15**, 5, 341–361 (2015).

Schwieterman, E. W., N. Y. Kiang, M. N. Parenteau, C. E. Harman, S. DasSarma, T. M. Fisher, G. N. Arney, H. E. Hartnett, C. T. Reinhard, S. L. Olson *et al.*, "Exoplanet biosignatures: a review of remotely detectable signs of life", Astrobiology **18**, 6, 663–708 (2018).

Seager, S., W. Bains and J. Petkowski, "Toward a list of molecules as potential biosignature gases for the search for life on exoplanets and applications to terrestrial biochemistry", Astrobiology **16**, 6, 465–485 (2016).

Seager, S., E. L. Turner, J. Schafer and E. B. Ford, "Vegetation's red edge: a possible spectroscopic biosignature of extraterrestrial plants", Astrobiology **5**, 3, 372–390 (2005).

Searle, J. R., "Minds, brains, and programs", Behavioral and brain sciences **3**, 3, 417–424 (1980).

Sergent, C. and S. Dehaene, "Neural processes underlying conscious perception: experimental findings and a global neuronal workspace framework", Journal of Physiology-Paris **98**, 4-6, 374–384 (2004).

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

Sevenius Nilsen, A., B. E. Juel and W. Marshall, "Evaluating approximations and heuristic measures of integrated information", Entropy **21**, 5, 525 (2019).

Shagrir, O., "The rise and fall of computational functionalism", Hilary Putnam pp. 220–250 (2005).

Shannon, C. E., "A mathematical theory of communication", Bell system technical journal **27**, 3, 379–423 (1948).

Shannon, C. E. and J. McCarthy, *Automata Studies.(AM-34)*, vol. 34 (Princeton University Press, 2016).

Sleep, N. H. and K. Zahnle, "Carbon dioxide cycling and implications for climate on ancient earth", Journal of Geophysical Research: Planets **106**, E1, 1373–1399 (2001).

Smith, K. C., "Life is hard: countering definitional pessimism concerning the definition of life", International Journal of Astrobiology **15**, 4, 277–289 (2016).

Snellen, I., L. Guzman-Ramirez, M. Hogerheijde, A. Hygate and F. van der Tak, "Re-analysis of the 267 ghz alma observations of venus-no statistically significant detection of phosphine", Astronomy & Astrophysics **644**, L2 (2020).

Spiegel, D. S. and E. L. Turner, "Bayesian analysis of the astrobiological implications of life's early emergence on earth", Proceedings of the National Academy of Sciences **109**, 2, 395–400 (2012).

Stanley, R. P., "Enumerative combinatorics volume 1 second edition", Cambridge studies in advanced mathematics (2011).

Sterelny, K. and P. E. Griffiths, *Sex and death: An introduction to philosophy of biology* (University of Chicago press, 2012).

Stevens, A., D. Forgan and J. O. James, "Observational signatures of self-destructive civilizations", International Journal of Astrobiology **15**, 4, 333–344 (2016).

Stewart, J., "Cognition= life: Implications for higher-level cognition", Behavioural processes **35**, 1-3, 311–326 (1995).

Tegmark, M., "Improved measures of integrated information", PLoS computational biology **12**, 11, e1005123 (2016).

Toker, D. and F. Sommer, "Moving past the minimum information partition: how to quickly and accurately calculate integrated information", arXiv preprint arXiv:1605.01096 (2016).

Tononi, G., "An information integration theory of consciousness", BMC neuroscience **5**, 1, 42 (2004).

Tononi, G., "Consciousness as integrated information: a provisional manifesto", The Biological Bulletin **215**, 3, 216–242 (2008).

Tononi, G., "The integrated information theory of consciousness: an updated account", Archives italiennes de biologie **150**, 2/3, 56–90 (2012).

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

Tononi, G., M. Boly, M. Massimini and C. Koch, "Integrated information theory: from consciousness to its physical substrate", Nature Reviews Neuroscience **17**, 7, 450–461 (2016).

Tononi, G. and G. M. Edelman, "Consciousness and complexity", science **282**, 5395, 1846–1851 (1998).

Trifonov, E. N., "Vocabulary of definitions of life suggests a definition", Journal of Biomolecular Structure and Dynamics **29**, 2, 259–266 (2011).

Turing, A., "Computing machinery and intelligence", Mind **59**, 236, 433–433 (1950).

Unterborn, C. T., S. J. Desch, N. R. Hinkel and A. Lorenzo, "Inward migration of the trappist-1 planets as inferred from their water-rich compositions", Nature Astronomy **2**, 4, 297–302 (2018).

Valentini, G., N. Masuda, Z. Shaffer, J. R. Hanson, T. Sasaki, S. I. Walker, T. P. Pavlic and S. C. Pratt, "Division of labour promotes the spread of information in colony emigrations by the ant temnothorax rugatulus", Proceedings of the Royal Society B **287**, 1924, 20192950 (2020).

Von Neumann, J. and R. Kurzweil, *The computer and the brain* (Yale University Press, 2012).

Walker, S. I., "Origins of life: a problem for physics, a key issues review", Reports on Progress in Physics **80**, 9, 092601, URL `https://doi.org/10.1088/1361-6633/aa7804` (2017).

Walker, S. I., W. Bains, L. Cronin, S. DasSarma, S. Danielache, S. Domagal-Goldman, B. Kacar, N. Y. Kiang, A. Lenardic, C. T. Reinhard *et al.*, "Exoplanet biosignatures: future directions", Astrobiology **18**, 6, 779–824 (2018).

Walker, S. I. and P. C. Davies, "The algorithmic origins of life", Journal of the Royal Society Interface **10**, 79, 20120869 (2013).

Walker, S. I., P. C. Davies and G. F. Ellis, *From matter to life: information and causality* (Cambridge University Press, 2017a).

Walker, S. I., N. Packard and G. Cody, "Re-conceptualizing the origins of life", (2017b).

Wheeler, M., "Cognition's coming home: The reunion of life and mind", in "Proceedings of the fourth European conference on artificial life", pp. 10–19 (Citeseer, 1997).

Whipple, K. X. and B. J. Meade, "Controls on the strength of coupling among climate, erosion, and deformation in two-sided, frictional orogenic wedges at steady state", Journal of Geophysical Research: Earth Surface **109**, F1 (2004).

Wilkinson, S., G. Deane, K. Nave and A. Clark, "Getting warmer: predictive processing and the nature of emotion", in "The value of emotions for knowledge", pp. 101–119 (Springer, 2019).

Wolszczan, A. and D. A. Frail, "A planetary system around the millisecond pulsar psr1257+ 12", Nature **355**, 6356, 145–147 (1992).

Wright, J. T., K. M. Cartier, M. Zhao, D. Jontof-Hutter and E. B. Ford, "The $\hat{G}$ search for extraterrestrial civilizations with large energy supplies. iv. the signatures and information content of transiting megastructures", The Astrophysical Journal **816**, 1, 17 (2015).

Zeiger, H. P., "Cascade synthesis of finite-state machines", Information and Control **10**, 4, 419–433 (1967a).

Zeiger, H. P., "Cascade decomposition using covers", in "Algebraic Theory of Machines, Languages, and Semigroups", edited by A. M. Arbib, chap. 4, pp. 55–80 (Academic Press, 1968).

Zeiger, P., "Yet another proof of the cascade decomposition theorem for finite automata", Theory of Computing Systems **1**, 3, 225–228 (1967b).

APPENDIX A

ON THE COMPUTATIONAL COMPLEXITY OF Φ

For a subsystem of size n, the computational complexity scales as follows. First, one must calculate the cause-effect structure (CES) for every possible partition of the subsystem. If the partition is a bipartition, as is typically assumed (Mayner *et al.*, 2018; Krohn and Ostwald, 2017), the number of ways to do this is $S(n, 2)$, where n is the size of the subsystem and $S(n, m)$ are Stirling numbers of the second kind Stanley (2011). However, two small corrections should be considered: first, partitions are unidirectional, and second, the unpartitioned system must also be addressed. The former consideration results in twice as many bipartitions, while the latter results in a single additional partition. Combining these results in a total of $2S(n, 2)+1$ partitions which, for large n, is well approximated as $2S(n, 2)$. For each CES, there are $2^n - 1$ potential mechanisms, corresponding to the size of the powerset of elements excluding the empty set. For each potential mechanism, there are $\binom{n}{k}$ purview elements of size k, each of which can be partitioned $S(k, 2)$ times. Therefore, there are $2\sum_k \binom{n}{k} S(k, 2) = 2(3^n)$ elementary distance calculations that must be performed to calculate a single CES, where the additional factor of two is due to the need to optimize ϕ^{max} over both past and future purviews. Putting this together, there are a total of $2S(n, 2 + 1) \times (2^n - 1) \times 2(3^n) \approx 12^n$ elementary distance calculations required to get the system-level integrated information Φ^{MIP} for a given subsystem. For the global system, this calculation must be embedded in an additional optimization corresponding to maximizing over the powerset of all possible subsystems (i.e. $\Phi^{max} = \max\{\Phi^{MIP}\}$). For a global system of size m, there are $\binom{m}{n}$ subsystems of size n, each with 12^n elementary distance calculations. Therefore, there are a total of $\sum_n \binom{m}{n} 12^n = 13^m$ elementary calculations required to find Φ^{max} for a global system of size m. For all but the smallest m values, the computational resources required to actually calculate Φ^{max} are impossible to realize.

Interestingly, the $\mathcal{O}(13^m)$ scaling derived here is in tension with the previously published value of $\mathcal{O}(53^m)$ (Mayner *et al.*, 2018). This could be due to the possibility that the $\mathcal{O}(53^m)$ scaling considers all possible partitions, rather than strict bipartitions, or perhaps it resolves the elementary computation in terms of some more fundamental operation (e.g. bit flips). Without additional information, it is difficult to say whether or not either of these considerations could resolve the tension between values. We do note, however, that the published values of $t = 1$, $t = 16$, and $t = 9900s$ for $n = 3$,$n = 5$, and $n = 7$ (Mayner *et al.*, 2018) are within an order of magnitude of the predicted $\mathcal{O}(13^m)$ scaling while off by 40 orders of magnitude from an $\mathcal{O}(53^m)$ scaling, though the use of parallelization complicates this point.

APPENDIX B

CALCULATING AN UPPER BOUND ON Φ

It is relatively straightforward to calculate a loose upper bound on Φ^{MIP} for a subsystem of size n. To do so, one need only understand the extension of the earth mover's distance D that is used in the calculation $\Phi^{MIP} = D(C||C_{\rightarrow})$. By definition, the "earth" being moved is ϕ^{Max} between concepts in the unpartitioned CES (C) and the partitioned CES ($C_{\rightarrow}$), while the "distance" it is moved is measured by the regular earth mover's distance between concepts (Oizumi *et al.*, 2014). In light of this, a straightforward upper bound on Φ^{MIP} can be found by asking what the maximum value of ϕ^{Max} is for each concept and moving that amount as far away as possible. For a mechanism of size m, its ϕ^{Max} value is bounded from above by the maximum value of the regular earth mover's distance, which is $EMD^{Max}(m) = m$. It is easy to see this is the case, as EMD^{Max} is achieved when all the probability ($p = 1$) is moved the maximum Hamming distance (H^{Max}), which is m for a mechanism comprised of m bits. For example, EMD^{Max} for a three-bit mechanism is achieved when $p = 1$ is moved from state 000 to state 111 ($H^{Max} = 3$), so $\phi^{max} = EMD^{Max} = 3$. Next, we must ask what the maximum distance D^{Max} in conceptual space is that this amount of ϕ^{Max} can be moved. Since this distance is again a regular earth mover's distance, we have $D^{Max} = EMD^{Max}(m) = m$. Thus, the maximum contribution a mechanism of size m can make to the extended earth mover's distance D is upper bounded by $\phi^{Max}(m)D^{Max}(m) = [EMD^{Max}(m)]^2 = m^2$. Of course, not all mechanisms are the same size, so the total contribution is bounded by the sum of the maximum contribution from mechanisms of each size, namely:

$$\Phi^{MIP}(n) \leq \sum_{m=1}^{n} \binom{n}{m} m^2 = 2^{n-2} n(n+1)$$

To date, this is the only known upper bound on Φ^{Max} that we are aware of (though bounds on ϕ^{max} and IIT 2.0 are readily available (Krohn and Ostwald, 2017; Oizumi *et al.*, 2016a; Arsiwalla and Verschure, 2016; Tegmark, 2016; Toker and Sommer, 2016)), and it is a very loose bound. For a subsystem of size $n = 2$, as is the case for the `AND+OR` system we consider in the main text, we have $\Phi^{MIP}(2) \leq (1)^2+(1)^2+(2)^2 = 6$ bits. In practice, we cannot reasonably expect $\phi^{Max} = EMD^{Max}$ for all mechanisms, as the existence of $\phi^{Max} = EMD^{max}$ for one mechanism almost certainly precludes the existence of $\phi^{Max} = EMD^{Max}$ for another. Likewise, cutting a CES cannot possibly result in a distance of $D^{Max} = EMD^{Max}$ for all concepts, as additional noise cannot be used to increase the fidelity of constraints. At best, it is likely that concepts map to the null concept in the CES of the MIP, corresponding to a maximum distance $D^{Max}(n) = n/2$. In this case, the bound that results is $\Phi^{MIP} \leq 2^{n-3} n(n+1)$, which is still likely loose. To tighten it, one must consider the ϕ^{Max} values that can result for a system of mechanisms as an ensemble, rather than individually, which is a task that we found quickly became intractable.

Numerical Approach

Fortunately, for our purposes a numerical approach will suffice. Given a small enough system, it is possible to calculate the Φ values for every possible transition probability matrix (TPM) that results from Boolean logic on a two-bit system. Namely, each bit (A and B) takes one of two possible states in response to the global state of the system. This means there are $2^4 = 16$ possible state transitions for each

coordinate, for a total of 16^2 unique TPMs. For each, it is possible to calculate the Φ spectrum that results using the algorithm we describe in the main text. Then, the upper bound on Φ^{MIP} is simply the maximum Φ^{MIP} value over all possible TPMs in all possible initial states. Since the system is only two bits, this bound on Φ^{MIP} is equivalent to the bound on Φ^{Max}, as subsystems must be comprised of at least two bits to generate $\Phi^{MIP} > 0$. Performing this exercise results in the bound $\Phi^{Max} \leq 1.5$, which is interestingly exactly one-fourth the analytical bound derived in the previous section; as discussed, it is likely that a factor of $1/2$ is accounted for if $D^{Max}(n) = n/2$, while the other factor of $1/2$ may be accounted for by the same type of argument applied to ϕ^{Max} (rather than D^{Max}). If so, the upper bound on Φ^{MIP} would be $2^{n-4}n(n+1)$ and is potentially more tractable to derive than previously believed.

APPENDIX C

BASIC USAGE FOR PYPHI-SPECTRUM

A pseudo-code overview of the `PyPhi-Spectrum` wrapper is shown in Algorithm 2, while basic usage is shown in Algorithm 3. Note, the `get-phi-spectrum` call returns the Φ values that result from all possible concepts for each cut, while the `get-phi-MIP` call returns the Φ values corresponding to the minimum information partition (i.e. Φ^{MIP}). Optimizing the latter over all possible subsystems would provide Φ^{Max} for a given system. To install the code, download or clone the entire `PyPhi-Spectrum` repository (which includes core `PyPhi` functionality) from github.com/jakehanson.

Algorithm 2: Pseudocode overview of the `PyPhi-Spectrum` wrapper

```
## Return the spectrum of Phi values
def get_phi_spectrum(subsystem):
        ## Initialize an empty list to store all Phi values for all cuts
        Phi_Spectrum = []
        ## Find all concepts for the specified subsystem
        all_concepts = get_all_concepts(subsystem)
        ## Create all possible CES via the Cartesian product of all concepts
        original_CES = get_all_CES(all_concepts)
        print("\tNumber of Non-unique Constellations =",len(original_CES))
        ## Cut the TPM and find all concepts. Get the new Phi value and repeat.
        bipartitions = get_all_bipartitions(cut_indices, cut_node_labels)
        for cut in bipartitions:
                print("\nEvaluating Cut ",cut)
                new_subsystem = subsystem.apply_cut(cut)
                ## Find all concepts for the specified subsystem
                new_concepts = get_all_concepts(new_subsystem)
                new_CES = get_all_CES(new_concepts)
                print("\tNumber of Non-unique Constellations =",len(new_CES))
                ## Now store all possible Phi values for this cut
                Phi_cut = []
                for original in original_CES:
                        for new in new_CES:
                                Phi = ces_distance(original,new)
                                if Phi not in Phi_cut:
                                        Phi_cut.append(Phi)
                ## Append the list of Phi values to the spectrum
                Phi_Spectrum.append(Phi_cut)
        return(bipartitions,Phi_Spectrum)
```

Algorithm 3: Basic Usage for the `PyPhi-Spectrum` Wrapper

```
import pyphi
import numpy as np
from pyphi import phi_spectrum

# TPM (little-end notation)
tpm = np.array([
    [0.,0.,0.],
    [0.,0.,0.],
    [1.,0.,0.],
    [1.,0.,1.],
    [0.,1.,0.],
    [0.,1.,0.],
    [1.,1.,0.],
    [1.,1.,1.]
])

# Set up network object
network = pyphi.Network(tpm, node_labels=['A','B','C'])
print("Network = ",network.node_labels)

# Put the system into a given state
state = (0,0,0)
nodes = ['A','B','C']

## Get the requisite Subsystem
subsystem = pyphi.Subsystem(network, state, nodes)

## Calculate all Phi values
display_CES= False  # if True, output will display constellations
solution = None # How to handle degeneracy ('Smallest','Largest', or 'Moon')
Phi_Spectrum = phi_spectrum.get_phi_spectrum(subsystem,display_CES,solution)

print("\nCuts = ",Phi_Spectrum[0])
print("\nPhi Spectrum = ",Phi_Spectrum[1])

Phi_MIP = phi_spectrum.get_Phi_MIP(Phi_Spectrum)
print("Phi MIP = ",Phi_MIP)
```

APPENDIX D

ADDITIONAL DETAILS RELATED TO THE CALCULATION OF Φ VALUES

In this section, we provide the transition probability matrices and initial states necessary to replicate our results. The same data can be found in downloadable form via the GitHub repository: https://github.com/jakehanson/pyphi-spectrum.

Photodiode (Chalmers and McQueen, 2014; Oizumi *et al.*, 2014)

A photodiode is a simple system of two interacting `COPY` gates, taking input from one another. It is arguably the simplest "integrated" system one can study, and has been studied in the context of IIT at least twice (Chalmers and McQueen, 2014; Oizumi *et al.*, 2014). Following Chalmers and McQueen Chalmers and McQueen (2014), we set the initial state of the system to be $s_0 = 10$. The transition probability matrix is given below.

Table D.1: The Transition Probability Matrix for a Simple Diode Comprised of Two Interconnected `COPY` Gates Taking Input from One Another Such as That Described in Chalmers and McQueen (2014)

s(t)	s(t+1)
00	00
10	01
01	10
11	11

`AND+OR` (Hanson and Walker, 2019; Albantakis *et al.*, 2019)

Like the photodiode, the `AND+OR` system has been studied in the context of IIT at least twice prior to the current work (Hanson and Walker, 2019; Albantakis *et al.*, 2019). However, a concrete Φ value has yet to be published. Therefore, we take the "published value" to be that of the PyPhi value found in Section 2.3. Similarly, we take the initial state to be $s_0 = 00$ in accordance with Section 2.3. The transition probability matrix is given below.

Table D.2: The Transition Probability Matrix for an `AND+OR` System Such as That Described in Section 2.3

s(t)	s(t+1)
00	00
10	01
01	01
11	11

Hanson and Walker (2019)

This system is a three bit digital counter in the initial state '101'. The initial state is selected somewhat arbitrarily, since any initial state will work, but $s_0 = 101$ results in a particularly fast evaluation. The TPM, from Figure 4 of the original publication, is as follows:

Table D.3: The Transition Probability Matrix for the Simple Electronic Counter From Hanson and Walker (2020)

s(t)	s(t+1)
000	110
100	000
010	101
110	010
001	100
101	111
011	001
111	011

Majority Gate System

This system is comprised of three interconnected majority gates, each with three inputs, as shown in Figure 2.8. If the majority of inputs to a given node are 0 the state of the node at the next timestep is 0 and if the majority of inputs to a given node are 1 the state of the node at the next timestep is 1. In the main text, the system is evaluated in initial state $s_0 = 000$. The transition probability matrix is provided below.

Table D.4: The Transition Probability Matrix for the `MAJ`+`MAJ`+`MAJ` System (Figure 2.8)

s(t)	s(t+1)
000	000
100	000
010	000
110	111
001	000
101	111
011	111
111	111

Gomez *et al.* (2021)

This papers studies the p53–Mdm2 biological regulatory network. Typically, this network is multivalued, but there are two possible binarizations that make standard Φ calculations possible. Of these, we chose the Fauré and Kaji binarization as it is much faster to analyze than the Tonello binarization. Following the authors, we choose an initial state $s_0 = 0001$ and use the following TPM. Note, the PyPhi value we compute for this TPM differs from that published by the authors due to their use of several non-standard configuration settings, such as Krohn and Ostwalds definition of Φ as a difference in integrated conceptual information rather than the IIT 3.0 definition.

Table D.5: The Transition Probability Matrix for the Fauré-Kaji Binarization of the p53-mdm2 Biological Regulatory Network From Gomez *et al.* (2021)

s(t)	s(t+1)
0000	1101
1000	1100
0100	1100
1100	1110
0010	1101
1010	1101
0110	1101
1110	1111
0001	0001
1001	0000
0101	0000
1101	0010
0011	0001
1011	0001
0111	0001
1111	0011

Farnsworth (2021)

In this paper a virocell (virus infected cell) is introduced into a Boolean network model of host cell dynamics. There are two network models provided, the first consists of five nodes and is the "full system", while the second consists of three nodes and is the "reduced system". For both systems, we study the case where all the nodes are 'ON' (i.e. $s_0 = 11111$ and $s_0 = 111$, respectively). Following the Supplementary Material provided by Farnsworth, the transition probabilities matrices are given below. Note, in the full system, the second node is an `AND` gate (as shown in his Figure 6) rather than a `COPY` gate (as shown in Figure 8 of his Supplementary Material).

Table D.6: The Transition Probability Matrix for the Entire Boolean Network Model of Virus-host Dynamics From Farnsworth (2021)

s(t)	s(t+1)
00000	00000
10000	00000
01000	10000
11000	10000
00100	01000
10100	01000
01100	11000
11100	11000
00010	00100
10010	00101
01010	10100
11010	10101
00110	01100
10110	01101
01110	11100
11110	11101
00001	00000
10001	00000
01001	10010
11001	10010
00101	01000
10101	01000
01101	11010
11101	11010
00011	00100
10011	00101
01011	10110
11011	10111
00111	01100
10111	01101
01111	11110
11111	11111

Table D.7: The Transition Probability Matrix for the Reduced System From Farnsworth (2021)

s(t)	s(t+1)
000	000
100	000
010	100
Continued on next page	

Table D.7 – continued from previous page

s(t)	s(t+1)
110	101
001	010
101	010
011	110
111	111

Oizumi *et al.* (2014)

This is the canonical `OR+AND+XOR` system that is often used in demonstrating how to calculate Φ (Tononi, 2015; Oizumi *et al.*, 2014; Mayner *et al.*, 2018). Following Oizumi et al., we take the system to be in the initial state $s_0 = 100$. The transition probability matrix is given below.

Table D.8: The Transition Probability Matrix for the `OR+AND+XOR` System From Oizumi *et al.* (2014)

s(t)	s(t+1)
000	000
100	001
010	101
110	100
001	100
101	111
011	101
111	110

Tononi *et al.* (2016)

This paper demonstrates the calculation of Φ for a simple system of four interacting logic gates: `MAJORITY+OR+AND+AND`. Following the authors, we use the initial state $s_0 = 1110$. The transition probability matrix is given below.

Table D.9: The Transition Probability Matrix for the `MAJ+OR+AND+AND` System From Tononi *et al.* (2016)

s(t)	s(t+1)
0000	0000
1000	0100
0100	0000

Continued on next page

Table D.9 – continued from previous page

s(t)	s(t+1)
1100	1110
0010	0000
1010	1100
0110	1000
1110	1110
0001	0100
1001	0100
0101	0100
1101	1110
0011	0101
1011	1101
0111	1101
1111	1111

Hoel *et al.* (2016)

This paper examines several small Boolean networks at both micro and macro scales. We choose to analyze the smallest of microsystems here, which is a system of four interconnected `AND` gates with noisy input. Following the authors, we analyze the system in initial state $s_0 = 0000$. Due to the noisy input, the TPM is not deterministic and therefore cannot be written as an N by 2 matrix. Instead, it must be written as an N by N matrix where entry (i, j) specifies the probability of state i transitioning to state j at timestep $t+1$ (a standard transition probability matrix). The transition probability matrix is given below.

Table D.10: The Transition Probability Matrix for the Noisy `AND`+`AND`+`AND`+`AND` System From Hoel *et al.* (2016)

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
0	0.24	0.10	0.10	0.04	0.10	0.04	0.04	0.02	0.10	0.04	0.04	0.02	0.04	0.02	0.02	0.01
1	0.24	0.10	0.10	0.04	0.10	0.04	0.04	0.02	0.10	0.04	0.04	0.02	0.04	0.02	0.02	0.01
2	0.24	0.10	0.10	0.04	0.10	0.04	0.04	0.02	0.10	0.04	0.04	0.02	0.04	0.02	0.02	0.01
3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.49	0.21	0.21	0.09
4	0.24	0.10	0.10	0.04	0.10	0.04	0.04	0.02	0.10	0.04	0.04	0.02	0.04	0.02	0.02	0.01
5	0.24	0.10	0.10	0.04	0.10	0.04	0.04	0.02	0.10	0.04	0.04	0.02	0.04	0.02	0.02	0.01
6	0.24	0.10	0.10	0.04	0.10	0.04	0.04	0.02	0.10	0.04	0.04	0.02	0.04	0.02	0.02	0.01
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.49	0.21	0.21	0.09
8	0.24	0.10	0.10	0.04	0.10	0.04	0.04	0.02	0.10	0.04	0.04	0.02	0.04	0.02	0.02	0.01
9	0.24	0.10	0.10	0.04	0.10	0.04	0.04	0.02	0.10	0.04	0.04	0.02	0.04	0.02	0.02	0.01
10	0.24	0.10	0.10	0.04	0.10	0.04	0.04	0.02	0.10	0.04	0.04	0.02	0.04	0.02	0.02	0.01
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.49	0.21	0.21	0.09
12	0.00	0.00	0.00	0.49	0.00	0.00	0.00	0.21	0.00	0.00	0.00	0.21	0.00	0.00	0.00	0.09
13	0.00	0.00	0.00	0.49	0.00	0.00	0.00	0.21	0.00	0.00	0.00	0.21	0.00	0.00	0.00	0.09
14	0.00	0.00	0.00	0.49	0.00	0.00	0.00	0.21	0.00	0.00	0.00	0.21	0.00	0.00	0.00	0.09
15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Marshall *et al.* (2017b)

This model system is the fission yeast cell cycle from Marshall *et al.* (2017b). As mentioned in the main text, we study the three node subsystem rather than the full

eight node subsystem (plus one external node) studied in the original publication. To calculate the spectrum of Φ values for this subsystem (or just the PyPhi Φ value), the TPM for the entire system (all nine nodes) is required. Therefore, there are 512 states in the TPM. Following the authors, the initial state of the system is set to $s_0 = 000110011$. In little-end binary notation (most significant bit on the right), the TPM used is as follows:

Table D.11: The Transition Probability Matrix for the Three-node Fission Yeast System From Marshall *et al.* (2017b)

s(t)	s(t+1)	**s(t)**	s(t+1)	**s(t)**	s(t+1)	**s(t)**	s(t+1)
0	2	**128**	162	**256**	78	**384**	110
1	2	**129**	162	**257**	66	**385**	98
2	130	**130**	162	**258**	2	**386**	162
3	130	**131**	162	**259**	2	**387**	162
4	4	**132**	132	**260**	76	**388**	76
5	0	**133**	128	**261**	68	**389**	68
6	128	**134**	128	**262**	4	**390**	132
7	128	**135**	128	**263**	0	**391**	128
8	8	**136**	136	**264**	76	**392**	76
9	0	**137**	128	**265**	72	**393**	72
10	128	**138**	128	**266**	8	**394**	136
11	128	**139**	128	**267**	0	**395**	128
12	12	**140**	140	**268**	76	**396**	76
13	0	**141**	128	**269**	76	**397**	76
14	128	**142**	128	**270**	12	**398**	140
15	128	**143**	128	**271**	0	**399**	128
16	256	**144**	384	**272**	332	**400**	332
17	256	**145**	384	**273**	320	**401**	320
18	384	**146**	384	**274**	256	**402**	384
19	384	**147**	384	**275**	256	**403**	384
20	260	**148**	388	**276**	332	**404**	332
21	256	**149**	384	**277**	324	**405**	324
22	384	**150**	384	**278**	260	**406**	388
23	384	**151**	384	**279**	256	**407**	384
24	264	**152**	392	**280**	332	**408**	332
25	256	**153**	384	**281**	328	**409**	328
26	384	**154**	384	**282**	264	**410**	392
27	384	**155**	384	**283**	256	**411**	384
28	268	**156**	396	**284**	332	**412**	332
29	256	**157**	384	**285**	332	**413**	332
30	384	**158**	384	**286**	268	**414**	396
31	384	**159**	384	**287**	256	**415**	384
32	18	**160**	178	**288**	82	**416**	114
33	18	**161**	178	**289**	82	**417**	114
34	146	**162**	178	**290**	18	**418**	178

Continued on next page

Table D.11 – continued from previous page

s(t)	s(t+1)	s(t)	s(t+1)	s(t)	s(t+1)	s(t)	s(t+1)
35	146	**163**	178	**291**	18	**419**	178
36	16	**164**	144	**292**	84	**420**	84
37	16	**165**	144	**293**	80	**421**	80
38	144	**166**	144	**294**	16	**422**	144
39	144	**167**	144	**295**	16	**423**	144
40	16	**168**	144	**296**	88	**424**	88
41	16	**169**	144	**297**	80	**425**	80
42	144	**170**	144	**298**	16	**426**	144
43	144	**171**	144	**299**	16	**427**	144
44	16	**172**	144	**300**	92	**428**	92
45	16	**173**	144	**301**	80	**429**	80
46	144	**174**	144	**302**	16	**430**	144
47	144	**175**	144	**303**	16	**431**	144
48	272	**176**	400	**304**	336	**432**	336
49	272	**177**	400	**305**	336	**433**	336
50	400	**178**	400	**306**	272	**434**	400
51	400	**179**	400	**307**	272	**435**	400
52	272	**180**	400	**308**	340	**436**	340
53	272	**181**	400	**309**	336	**437**	336
54	400	**182**	400	**310**	272	**438**	400
55	400	**183**	400	**311**	272	**439**	400
56	272	**184**	400	**312**	344	**440**	344
57	272	**185**	400	**313**	336	**441**	336
58	400	**186**	400	**314**	272	**442**	400
59	400	**187**	400	**315**	272	**443**	400
60	272	**188**	400	**316**	348	**444**	348
61	272	**189**	400	**317**	336	**445**	336
62	400	**190**	400	**318**	272	**446**	400
63	400	**191**	400	**319**	272	**447**	400
64	66	**192**	194	**320**	78	**448**	78
65	66	**193**	194	**321**	66	**449**	66
66	130	**194**	130	**322**	66	**450**	194
67	130	**195**	130	**323**	66	**451**	194
68	68	**196**	196	**324**	76	**452**	76
69	64	**197**	192	**325**	68	**453**	68
70	128	**198**	128	**326**	68	**454**	196
71	128	**199**	128	**327**	64	**455**	192
72	72	**200**	200	**328**	76	**456**	76
73	64	**201**	192	**329**	72	**457**	72
74	128	**202**	128	**330**	72	**458**	200
75	128	**203**	128	**331**	64	**459**	192
76	76	**204**	204	**332**	76	**460**	76
77	64	**205**	192	**333**	76	**461**	76
78	128	**206**	128	**334**	76	**462**	204

Continued on next page

Table D.11 – continued from previous page

s(t)	s(t+1)	s(t)	s(t+1)	s(t)	s(t+1)	s(t)	s(t+1)
79	128	**207**	128	**335**	64	**463**	192
80	320	**208**	448	**336**	332	**464**	332
81	320	**209**	448	**337**	320	**465**	320
82	384	**210**	384	**338**	320	**466**	448
83	384	**211**	384	**339**	320	**467**	448
84	324	**212**	452	**340**	332	**468**	332
85	320	**213**	448	**341**	324	**469**	324
86	384	**214**	384	**342**	324	**470**	452
87	384	**215**	384	**343**	320	**471**	448
88	328	**216**	456	**344**	332	**472**	332
89	320	**217**	448	**345**	328	**473**	328
90	384	**218**	384	**346**	328	**474**	456
91	384	**219**	384	**347**	320	**475**	448
92	332	**220**	460	**348**	332	**476**	332
93	320	**221**	448	**349**	332	**477**	332
94	384	**222**	384	**350**	332	**478**	460
95	384	**223**	384	**351**	320	**479**	448
96	82	**224**	210	**352**	82	**480**	82
97	82	**225**	210	**353**	82	**481**	82
98	146	**226**	146	**354**	82	**482**	210
99	146	**227**	146	**355**	82	**483**	210
100	80	**228**	208	**356**	84	**484**	84
101	80	**229**	208	**357**	80	**485**	80
102	144	**230**	144	**358**	80	**486**	208
103	144	**231**	144	**359**	80	**487**	208
104	80	**232**	208	**360**	88	**488**	88
105	80	**233**	208	**361**	80	**489**	80
106	144	**234**	144	**362**	80	**490**	208
107	144	**235**	144	**363**	80	**491**	208
108	80	**236**	208	**364**	92	**492**	92
109	80	**237**	208	**365**	80	**493**	80
110	144	**238**	144	**366**	80	**494**	208
111	144	**239**	144	**367**	80	**495**	208
112	336	**240**	464	**368**	336	**496**	336
113	336	**241**	464	**369**	336	**497**	336
114	400	**242**	400	**370**	336	**498**	464
115	400	**243**	400	**371**	336	**499**	464
116	336	**244**	464	**372**	340	**500**	340
117	336	**245**	464	**373**	336	**501**	336
118	400	**246**	400	**374**	336	**502**	464
119	400	**247**	400	**375**	336	**503**	464
120	336	**248**	464	**376**	344	**504**	344
121	336	**249**	464	**377**	336	**505**	336
122	400	**250**	400	**378**	336	**506**	464

Continued on next page

Table D.11 – continued from previous page

s(t)	s(t+1)	s(t)	s(t+1)	s(t)	s(t+1)	s(t)	s(t+1)
123	400	**251**	400	**379**	336	**507**	464
124	336	**252**	464	**380**	348	**508**	348
125	336	**253**	464	**381**	336	**509**	336
126	400	**254**	400	**382**	336	**510**	464
127	400	**255**	400	**383**	336	**511**	464

APPENDIX E

STATEMENT OF COAUTHOR PERMISSIONS

All coauthors have granted their permission to use articles Hanson and Walker (2019) and Hanson and Walker (2020) for Chapters 3 and 4 respectively.