Reducing Uncertainty

Understanding the Information-Theoretic Origins of Consciousness

By

Garrett Mindt

Submitted to
Central European University
Department of Philosophy

In partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Philosophy

Supervisor: Tim Crane

Budapest, Hungary
2019

I hereby declare that the dissertation contains no material accepted for the completion of any other degrees in any other institutions and no materials previously written and/or published by another person unless appropriate acknowledgement is made in the form of biographical reference.

Budapest, 31st July, 2019 Garrett Mindt

Abstract

Ever since the hard problem of consciousness (Chalmers, 1996, 1995) first entered the scene in the debate over consciousness many have taken it to show the limitations of a scientific or naturalist explanation of consciousness. The hard problem is the problem of explaining why there is any experience associated with certain physical processes, that is, why there is anything it is like associated with such physical processes? The character of one's experience doesn't seem to be entailed by physical processes and so an explanation which can overcome such a worry must (1) explain how physical processes give rise to experience (explain the entailment), (2) give an explanation which doesn't rely on such physical processes, or (3) show why the hard problem is misguided in some sense.

Recently, a rather ambitious and novel theory of consciousness has entered the scene – Integrated Information Theory (IIT) of Consciousness (Oizumi et al., 2014; Tononi, 2008; Tononi et al., 2016) – and proposes that consciousness is the result of a specific type of information processing, what those developing the theory call *integrated information*. The central aim of this dissertation is to philosophically investigate IIT and see whether it has the ability to overcome the hard problem and related worries. I then aim to use this philosophical investigation to answer a set of related questions which guide this dissertation, which are the following: Is it possible to give an information-theoretic explanation of consciousness? What would the nature of such an explanation be and would it result in a novel metaphysics of consciousness?

In this dissertation, I begin in chapter one by first setting up the hard problem and related arguments against the backdrop of IIT (Mindt, 2017). I show that given a certain understanding of structural and dynamical properties IIT fails to overcome the hard problem of consciousness. I go on in chapter two to argue that a deflationary account of causation is the best view for IIT to overcome the causal exclusion problem (Baxendale and Mindt, 2018). In chapter three, I explain IIT's account of how the qualitative character of our experience arises (qualia) and what view of intentionality (the directedness of our mental states) IIT advocates. I then move on in chapter four to show why the hard problem mischaracterizes structural and dynamical properties and misses important nuances that may shed light on giving a naturalized explanation of consciousness. In the last and fifth chapter, I outline a sketch of a novel metaphysics of consciousness that takes the conjunction of Neutral Monism and Information-Theoretic Structural Realism to give what I call Information-Theoretic Neutral-Structuralism.

Acknowledgements

I have over the course of my studies been overgenerously blessed with a wide-range of influences and people in my life over these many years, so much so, their friendship and intellectual influences overshadows the work of this dissertation by a longways. I would first like to thank the person who initially got me started in studying philosophy of mind and continuing my study of philosophy past the initial stages of my BA, Terrence Horgan. Although I suspect he would faintly remember me from my short time at the University of Arizona so many years ago, the initial spark he struck in me created a lifelong passion for philosophy and the study of consciousness. This is something I will forever be grateful for, as it's taken me to places I couldn't have imagined going all those years ago, so thank you, Terry. Thank you to Iris Oved, who during the last couple of years at U of A shared with me a generous amount of her time and invaluable guidance in deciding whether I should continue studying philosophy. Acting as preceptor for her course on the philosophy of cognitive science was my first glimpse into what being a professional philosopher and academic entailed and was a big motivator in my decision in choosing to continue my training in philosophy after my BA. During my time at U of A I also met some incredible people studying philosophy, thank you to Steven Gubka, Scott Plummer, and Jay Smith, I had a fun time stumbling through philosophy with all of you during those early years. Scott Plummer was one of the first people with whom I discovered the joy of discussing philosophy, although his intellect and ability were light-years past my own. I always felt like I was three steps behind in our arguments. In many ways, I still feel like I'm playing catch up with him even after all these years. Scott you will be missed, you left us too early, but I hope you've found some peace.

I met Philip Goff during my time at the University of Arizona at the 5th Southwest Undergraduate Philosophy Conference, where he was the keynote speaker and I was giving a rather shaky talk on the Chinese room argument and connectionist AI. He convinced me shortly after to apply for an MA at the University of Liverpool, where he was a lecturer. I decided to pack my bags and make my first trip across the pond (for the first time of two) to study for a MA in Liverpool. It was during this time that I met some truly remarkable people that have made a lasting impression on my life, both philosophically and personally. I would like to thank Gregory Miller, Raven Skobe, Kaleena Stoddard, Dan Whistler, and Grace Whistler, for their enduring friendship over the years since, I didn't know you before I came to Liverpool, though now, I can scarcely imagine what life would have been like if y'all hadn't come into it, thank you. Gregory Miller, you've proven to be one

of my closest friends and philosophical sparring partners, your inability to let an argument go or a premise unturned has made me a better thinker and philosopher. Thank you, Greg, for the years of friendship. Thank you to Grace Whistler for the many fun trips and travels we've gone on over the years, it's rare that I meet someone who enjoys spontaneous adventures and travels as much as I do. Even that random road trip to Escanaba in Upper Michigan that one time (though I could tell you weren't pleased with just how spontaneous and random that trip was). You're always up for an adventure and I couldn't ask for a better friend.

Thank you to Dan Whistler for continuing to keep a watchful eye on me even after our time at Liverpool. You've always supported me and given me invaluable advice, you're a good mentor and an even better friend, thank you for everything. While at Liverpool, I completed my MA under the supervision of Barry Dainton. Barry steered me in the direction of writing my MA on Integrated Information Theory (IIT) of Consciousness, and so, to a large extent, was the impetuous for pursuing the research that was the central focus of this PhD dissertation. If there's anyone to blame for my current research interests and my subsequent writing on it, it's Barry. Thank you, Barry, for pushing me in that direction, guiding me through the MA, and the discussions we've subsequently had since over the years.

Philip left Liverpool in the middle of my MA (still bitter about this, Philip) and took a post at Central European University (CEU) in Budapest, Hungary, where he once again convinced me to apply for a PhD to work with him (I'm apparently a glutton for punishment). I applied and soon was packing my bags once again for the next trip across the pond, this time to spend four years in Budapest. It was in Budapest that I was once again exceptionally lucky to meet the people I did. Thank you to Antonio Alfieri, Damian Aleksiev, Matthew Baxendale, Zoltan Brys, Mathieu Charbonneau, Arianna "Pickle" Curioni, Caglan Dilek, Jamie Elliot, Zhiwei Gu, Rob Hoveman, Anna Kocsis, Michele Luchetti, Heather Mackenzie Morris, Carlos Montemayor, Marta Santuccio, James Strachan, in some way you all made an impression on me personally and/or philosophically.

I would especially like to thank Matthew Baxendale and Michele Luchetti, the other two members of the Flying Tigers (the name Philip gave our little trio), your friendship was a truly remarkable gift. I honestly don't know what the PhD or my life in Budapest would have been like if you two hadn't been a part of it, and words cannot describe the joy it has brought me knowing you

both over these past few years. Thank you to Mathieu Charbonneau, you've given me some invaluable advice and encouragement over the years, I cannot thank you enough for those words of encouragement and support. Thank you to Heather Mackenzie Morris, you taught me that there can be a life outside of philosophy, which is a valuable lesson and gift. I would also like to thank Jamie Elliot who entered the scene at the beginning of my second year of the PhD, you are one of the most genuine and true friends anyone could ask for, thank you for your friendship. Thank you to Carlos Montemayor for your lending your ear for numerous discussions, writing and rewriting our collaborative paper, and for all the fun times in San Francisco, your generosity and kindness has not gone unappreciated.

Marta Santuccio, I'm not usually at a loss for words but when it comes to you there isn't enough to capture the impact you've had on me and my life, so I will keep it short and to the point. Marta, thank you, I carry it (all) within.

The faculty and staff at the philosophy department at CEU have had a great deal of impact on me over these years and for that I am thankful. Thank you to Maria Kronfeldner for invaluable advice about research and publishing at the earliest stages of my PhD and for sparking my interest in the philosophy of science, this guidance was immensely beneficial to me. Thank you to Kriszta Biber and Zsofi Jeney for always helping me whenever I had an issue (or more often than not, when I made a mistake), the many times which had you going above and beyond the call of duty. Thank you to Hanoch Ben-Yami for always steering me in the direction of taking on various responsibilities in the department and pointing out opportunities, although at the time they may have felt like burdens, they were a necessary part of my training and I'm a better philosopher for them. Thank you to Philip Goff for once again taking me on as a student and supervising me during the first three years of my PhD (he left in the middle of my studies for a second time, can you believe that?!). You recognized something in me when I was an undergraduate (I still don't know what), you have been an incredible mentor over these years at the beginning of my philosophical career, I hope that one day I might be able to repay this weighty intellectual debt. Thank you, Philip. Thank you to Tim Crane who generously took over my supervision when I was so unceremoniously abandoned by Philip, I am extremely grateful for your guidance in the last year of the PhD and for thoughtful comments on the work that composes this dissertation. In the short time we've worked together you've helped improve my philosophical writing and ability and for that I will always be thankful.

At the start of my fourth year I was awarded a grant to visit the Wisconsin Institute for Sleep and Consciousness under the supervision of the lab's director, Giulio Tononi. Thank you to Central European University for awarding me the Doctoral Research Support Grant that made the visit possible. Thank you to Giulio for allowing me to hang around the lab with everyone for a few months, it was an exhilarating and invaluable period of learning and discussion, it's a time I will always look back on fondly. Giulio was generous with his time, knowledge, and hospitality, and I cannot thank him enough. We had philosophical disagreements, the most important of which I still maintain I am correct about, the phenomenal character of Sauternes is still most like the aftertaste of a Connecticut wrapped cigar and melted lemon otter pops. I will hear nothing to the contrary. Thank you to the other members of the lab that made my stay in Madison so spectacular, Larrissa Albantakis, Leonardo Barbasso, Melanie Boly, Erick Chastain, Renzo Comolatti, Graham Findley, Matteo Grasso, Andrew Haun, Elsa Juan, Bjørn Erik Juel, Jonathan Lang, Sophia Loschky, Rong Mao, Will Mayner, Michael Payton, and Giovanna Spano. I would like to especially thank Larissa Albantakis, who after I met her at a workshop during the first year of my PhD at NYU on IIT, was extremely kind and helpful over the years answering my questions about IIT. Larissa also provided comments on much earlier versions of the first two essays in this dissertation and they were all the better because of those comments, thank you. I would especially like to thank the other two philosophers at WISC during my time there, Matteo Grasso and Jonathan Lang. Thank you Matteo for immediately including me in the group and making my stay so enjoyable, I hope one day I can return the kindness. Thank you, Jonathan for helping to make my stay a reality and helping me to navigate the task of moving to Madison for a short stint, your help didn't go unnoticed and I can't thank you enough.

Thank you to my mother, Nichole Mindt, you were the best teacher I've ever had, and you've always supported me in whatever I do, I will always be thankful for everything you've done for me. Thank you to my sister Natalie Varela and my brother-in-law Johnny Varela, the life of a philosopher can be rather wanderlust, but you've always offered your home as my own when I needed a break away from everything and for that I am always grateful. I would also like to thank my father, Randy Mindt, for instilling a passion for reading at an early age. Running around a warehouse full of books as a child had an impact on my love and passion for reading and learning and I've lived a life surrounded by books and knowledge ever since as a result. I may no longer be playing in steel shipping containers filled with books but sometimes it still feels like I am, thank you.

I would like to end by thanking my internal examiner Howard Robinson, who during my time at Central European University was always a wonderful source of guidance and also a friendly and insightful presence in all discussions. Howard's comments on the thesis were helpful, insightful, and gave me much food for thought. Thank you to my external examiner, Kelvin McQueen, who provided some critical feedback and suggestions, I'm incredibly thankful for his thorough and thoughtful comments on my dissertation and the fun and enlightening discussion during the defense.

Table of Contents

List of Figures

General Introduction:

Consciousness, what is it? Where does it come from? Why is it so hard to think about? To a large extent this dissertation is a catalogue of my attempts to come to an understanding on consciousness. The questions which drive this dissertation are the following:

- Is it possible to give an information-theoretic explanation of consciousness?
- What would the nature of such an explanation be and would it result in a novel metaphysics of consciousness?

These questions are the central focus of this dissertation and the discussion takes place against the backdrop of Integrated Information Theory (IIT) of Consciousness. There are a couple of reasons for doing this which I want to explain now. My MA dissertation at the University of Liverpool was on IIT and the hard problem of consciousness. The research then was preliminary, and I wanted to delve a bit deeper into IIT and what the theory was actually claiming. At the same time, I had become interested in the ontology and metaphysics of information and the question that stuck with me after my MA was whether it was even possible to give an information-theoretic explanation of consciousness. I decided to treat IIT as something of a sample case for asking these questions. I set out to examine its claims and see if that could reveal anything about the nature of information and whether it was or was not related to phenomenal experience in any relevant or important ways.

Ultimately this is as much a dissertation on the nature of information as it is one about the nature of consciousness. Though, I think it will become clear to the reader as they read this thesis that I think both are intimately related to one another.

I want to briefly comment on the nature and structure of this dissertation, as it deviates slightly from the typical philosophical monograph. I have opted to write five essays, all on questions that I think are of central importance to coming to grips with what exactly the nature of consciousness is from an information-theoretic perspective. I wrote the essays, for the most part, to stand on their own as self-contained works. I also chose to write independent essays since a monograph on this topic would constitute a tome of a work which would have far exceeded the bounds and limitations of a PhD dissertation.

The first two essays were both published after my first and second year of the PhD. The first essay, "The Problem with the 'Information' in Integrated Information Theory," was published in the Journal of Consciousness Studies (2017) vol. 24 No. 7-8. I wrote this essay as a criticism against IIT and anticipated at the time that this would be the direction my dissertation took. I was pleasantly surprised when I discovered over the course of my research that things took a more positive and constructive turn. The first essay is best read as a conditional, if we take the hard problem of consciousness, the structure-dynamics argument, the explanatory gap argument, and IIT at face value, would IIT be able to overcome these arguments and account for phenomenal experience? In the essay I argue that IIT as it is currently conceived fails to solve the hard problem and falls victim to both the structure and dynamics and explanatory gap arguments. The first essay reads rather negatively but I meant it to be a kind of challenge to myself for writing the thesis. If I couldn't find a way to overcome those criticisms, then IIT wouldn't be a full account of the nature of consciousness and phenomenal experience. I then endeavored over the course of the next few years to develop arguments as to why IIT might be able to overcome the worries in the first essay.

The second essay, co-written with Matthew Baxendale[1], was published in the special issue Causality in the Sciences of Brain and Mind, edited by Lise Marie Anderson, Jonas Fogedgaard Christensen, Samuel Schindler, and Asbjørn Steglish-Peterson in Minds and Machines vol. 28 issue 2, pp 331-351, titled "Intervening on the Causal Exclusion Problem for Integrated Information Theory." The essay looked at the implications of the causal exclusion problem for IIT. Matthew and I placed IIT in the context of the causal exclusion problem and then investigated whether IIT had a solution to the issue. We argued that with slight modification IIT overcomes the causal exclusion problem and this is a point in favor of IIT as an explanation of consciousness. Ultimately, we argue that IIT should disentangle its use of information and causation and opt for a deflationary view of causation. We propose that IIT should adopt a broadly manipulationist brand of interventionist causation.

[1] To comment briefly on the division of labor for the work, the essay was from start to finish, a collaborative effort written and re-written a number of times. However, we each came to the project with our own set of strengths, mine being the IIT and philosophy of mind side of things, Matthew with his knowledge of interventionist causation. And so, in those sections where we each had our own firmer body of knowledge a heavier hand from the respective author can be seen.

The next three essays were written in succession and so have some reference to each other, the 5th essay is something of a conclusion and so has the most direct engagement with the previous essays than the rest do with each other.

The third essay, "Integrated Information Theory and the Marks of the Mental," is more expository about IIT than the other essays. My aim with the third essay was to give an exposition on what IIT's implications are for what has become known as the two candidates for the mark of the mental, those being (1) phenomenal experience (qualia) and (2) intentionality. There were a couple of reasons for wanting to do this. Firstly, given that IIT can be rather technical given its mathematical framework, philosophers can sometimes be daunted by the task of searching through the formal aspects for the philosophical insights. I wanted to give a philosopher friendly exposition of IIT's claims about qualia and intentionality, and I attempted to do this with the third essay. The topic of qualia is something which IIT has directly engaged whereas on the question of intentionality IIT has stayed rather silent. I took this as an opportunity to give a philosophical treatment of what IIT claims about qualia and at the same time offer some speculations about what the view theory might claim with regard to intentionality.

The fourth essay, "Not all Structure and Dynamics are Equal," is focused on the structure and dynamics argument, which is the main argument behind the first essay. My goal with this essay was to (1) attempt to knock down the arguments I gave as a conditional in that first essay and (2) to show that the structure and dynamics argument only works if we have a limited conception of structural and dynamical (S & D) properties. I pulled from recent work in complexity sciences, concerned with giving a semantic notion of information by understanding the meaningful connections a system has to its environment relevant for that system to causally maintain its existence over time. The goal was to make a distinction between the external and internal structure and dynamics of a system, to show that not all notions of information and S & D are equal and there are important differences to be gleaned from looking to the intrinsic properties of complex systems. The aim in this essay was to put to rest the hard problem and the structure and dynamics argument to hopefully open the door on tackling consciousness from a natural and information-theoretic perspective.

The last essay is a speculative foray into metaphysics. I offer what I see as one possible metaphysics that falls out of the previous four essays. I maintain in the essays three important theses

which I think result in the view I call Information-Theoretic Neutral-Structuralism (ITNS). Those are (1) phenomenal realism – that the phenomenal character we are presented with in our experience is real – (2) informational realism – information is a real thing and not merely an abstraction from certain physical or mental processes – and (3) scientific realism – our natural investigations of the world reveal to us the nature of reality, albeit, sometimes indirectly (as is argued in chapter five). I think the conjunction of these three theses result in a view which claims that information is the ultimate ontological category, what Russell (1927) calls the *common ancestor* of both mind and matter. I argue that information should be treated as the common ancestor and is the basis of the metaphysical view that falls out as a result. I offer ITNS as the view that best encapsulates the conjunction of an information-theoretic neutral monism and information-theoretic structural realism. There is of course much work to be done after this point, but I hope to at least motivate that the general direction of this research is fruitful for finding consciousness' place in nature.

Overall, I hope this dissertation accomplishes two modest tasks. Firstly, that it shows the viability of looking at consciousness and how it comes about from an information-theoretic perspective. There is a powerful toolset to be found in investigating how phenomenal experience may come about through novel information-theoretic properties of certain systems. If there is one lesson that one might draw from looking at IIT, I think it should be this. That there may yet be aspects of information which we have yet to understand and looking towards measures which track a systems complexity is a route that needs to be explored further, particularly to crack the problem of consciousness. Secondly, that we might dispel at least some of the force of the arguments that stem from a certain limited conception of structural and dynamical properties. If this can be done (as I argue in the fourth essay) then we can open a whole new breadth of possible explanations for how the qualitative aspect of experience arises. I think this will have two illuminating results, it'll help us in understanding consciousness from a naturalistic perspective and might provide insights into the nature of information. If I can show at least these two things I will be satisfied with the work contained in this dissertation.

"Thus, there were perceptions that we did not consciously perceive right away, the apperception in this case arising only after an interval, however brief. In order better to recognize [juger] these tiny perceptions [petites percetions] that cannot be distinguished in a crowd, I usually make use of the example of the roar or noise of the sea that strikes us when we are at the shore. In order to hear this noise as we do, we must hear the parts that make up the whole, that is, we must hear the noise of each wave, even though each of these small noises is known only in the confused assemblage of all the others, and would not be noticed if the wave making it were the only one. For we must be slightly affected by the motion of this wave, and we must have some perception of each of these noises, however small they may be, otherwise we would not have the noise of a hundred thousand waves, since a hundred thousand nothings cannot make something. Moreover, we never sleep so soundly that we not have some weak and confused sensation, and we would never be awakened by the greatest noise in the world if we did not have some perception of its beginning, small as it might be, just as we could never break a rope by the greatest effort in the world, unless it were stretched and strained slightly by the least efforts, even though the slight extension they produce is not apparent.

These tiny perceptions are therefore more effectual than one thinks. They make up this I-know-not-what, those flavors, those images of the sensory qualities, clear in the aggregate but confused in their parts; they make up those impressions the surrounding bodies make on us, which involve the infinite, and this connection that each being has with the rest of the universe.

It can even be said that as a result of these tiny perceptions, the present is filled with the future and laden with the past, that everything conspires together, and that eyes as piercing as those of God could read the whole sequence of the universe in the smallest of substances.

The things that are, the things that have been, and the things that will soon be brought in by the future."

- G.W. Leibniz
Preface to the New Essays (1703-1705)

"Against the door he leans and starts a scene,
And his tears fall and burn the garden green

And so castles made of sand,
Fall in the sea, eventually"

- Jimi Hendrix
"Castles Made of Sand"
Axis Bold As Love (1967)

1. The Problem with the 'Information' in Integrated Information Theory

Introduction:

Ever since David Chalmers (1995) first introduced what he called the hard problem of consciousness it has been seen as a goal for a theory of consciousness to meet. The hard problem is the problem of explaining why there is any experience associated with all the physical processes going on inside our brains. There may be an elaborate story to explain *how* this might occur, such an explanation would consist of elaborating the structure and dynamics of that physical system (what Chalmers calls the easy problems), but such an explanation doesn't seem capable of answering the *why* question – why it feels like something for our brains to carry out all these physical processes (the hard problem). This essay will be looking at one attempt to explain the *how* and *why* questions of experience, *Integrated Information Theory (IIT) of Consciousness*.

In this essay I will be examining the foundations of IIT, specifically, how IIT defines and utilizes the notion of information as a base for a theory of consciousness. It has been argued (Chalmers, 2003, 1996) that physicalist accounts – those accounts which say the brain is wholly physical, and thus describable purely in terms of structural and dynamical features – are unable to offer a solution to the hard problem, since at most they will only ever explain more structure and dynamics, but fail to give an explanation of *why* there is any phenomenal experience associated with those physical processes. Through my discussion of IIT's use of information I will show that IIT is committed to a structural-dynamical (physicalist) notion of information and so falls victim to a number of anti-physicalist arguments.

I first introduce IIT in §1 and give a short account of the basic essence of the theory. Then I move on to give an overview of the account of information given within IIT, and suggest that in its

current formulation it is a purely physical notion of information, and because of this IIT faces a number of problems commonly raised against physicalist accounts (§2). I argue that this account of information is solely structural and dynamical, and so has the same explanatory power as other physicalist accounts of consciousness. In the next section (§3) I elaborate the explanatory gap argument and show how IIT succumbs to this argument as well. Therein, I also call into question some of the predictions (Tononi et al., 2016) of IIT based on the aforementioned explanatory gap worry.

I conclude by rephrasing the hard problem of consciousness in terms of information. Since information-theoretic theories, such as IIT, think consciousness is the result of information (specifically how information integrates and is carried through a system), we might call the resulting problem for such theories – *the hard problem of information: why is it the case that there is any experience associated with the informational processes occurring in our brain?* For information-theoretic accounts like IIT this is the heart of their hard problem, one must explain: (i) why particular organizations of information produce phenomenal experience in the brain, while other organizations of information, such as the laptop I am currently writing this essay on, produce none; and furthermore (ii) such explanations, to address the hard problem, must do so not merely through a purely structural-dynamical explanation. I do not think this is solely an issue for IIT, but rather for any account of consciousness that attempts to explain phenomenal experience as an information-theoretic phenomenon. If it can be shown that IIT's notion of information is insufficient to provide a foundation for a theory of consciousness, then IIT should revise the notion of information it utilizes in constructing the theory.

Section 1.1: What is Integrated Information Theory?

IIT proposes that consciousness is integrated information in a system, the degree of which is signified by the Greek letter, Φ[2]. The quantity of integrated information – or consciousness – present in a system is quantified by Φ which is "the amount of information generated by a complex of elements, above and beyond the information generated by its parts" (Tononi, 2008, p. 216). The substantial difference between IIT and other philosophical or neuroscientific theories of consciousness is that it recognizes the significant amount of data given to us in our everyday experience. According to IIT we can use this data in constructing an account of consciousness, one that gives us a physically realizable model of consciousness. Having such a model would be a giant leap forward in our understanding of the mind, as it would give us the ability to quantify consciousness and so measure and study it scientifically. This would presumably lead to us having the ability to detect and predict when consciousness is present in a system (Tononi et al., 2016). Aside from the ability to quantify the degree of consciousness present in a system, IIT might have interesting implications for certain empirical cases, such as, split-brain cases (Tononi & Koch 2015, Tononi et al., 2016), and dissociative and conversion disorders (Oizumi et al. 2014). The predicative and explanatory power of IIT gives one strong motivation to take IIT seriously, but only if integrated information is indeed identical to consciousness.[3] IIT makes the claim that consciousness can be captured in terms of varying quantities of integrated information, then we must be certain that the thing being quantified is indeed consciousness, and not merely integrated information itself. If one has good reason to think that when

[2] Tononi writes, "Integrated information is indicated with the symbol Φ (the vertical 'I' stands for information, the circle 'O' for integration)" (Tononi 2008: 220).

[3] It is important to note here that Cerullo (2015) calls into question the explanatory power of IIT, regardless of its ability to tackle the hard problem of consciousness. Cerullo argues that IIT is really a theory of proto-consciousness, and so any explanations it might offer regarding consciousness are really explanations of proto-consciousness. According to Cerullo, this doesn't seem to provide us any answers to the so called easy problems of consciousness (easy problems are things such as: attention, the directedness of behavior, the correspondence between memory and cognition, etc.).

one quantifies integrated information, one fails to quantify the degree of consciousness present in a system, then we have reason to suppose IIT is not a full explanation of consciousness.

IIT is constructed by first outlining what Tononi takes to be the five undeniable attributes of conscious experience (phenomenological axioms): *intrinsic existence*, *composition*, *information, integration*, and *exclusion*. According to IIT these axioms are evident to us through our experience, and so the theory takes them as axiomatic in constructing a theory of consciousness.[4] Tononi thinks that these phenomenological axioms are evidence enough to then derive a set of physical postulates which explain how these aspects of our phenomenology can be realized through a physical system, e.g. the brain. Since IIT is a neuroscientific theory its aim is to provide a detailed account of how consciousness is brought about by physical systems, it is the job of the physical systems postulates to give such an account. How might physical systems have the ability to bring about the essential aspects of our phenomenology (phenomenological axioms)? Presumably, according to IIT, this question is answered by the constraints detailed in the physical systems postulates. To give an example of how the axioms relate to the postulates let us take a look at the second axiom and postulate of IIT – *composition*:

> Consciousness is structured: each experience is composed of phenomenological distinctions, elementary or higher-order, which *exist* within it.
> (Tononi & Koch 2015: 7)

This axiom is meant to express the essential property of our conscious experience, that there are many phenomenal aspects to our experience at any given time. For example, say you are sitting at your favorite local coffee shop. Within your experiential field is a white coffee cup in front of you with a latte steaming inside. Within that experience you have the phenomenal distinctions of white-

[4] That is not to say that these axioms are exhaustive, Tononi and colleagues admit that there may be more than the current five in IIT as it stands now. I will be taking these for granted as they are not the focus of this essay, but one could find disagreement in the axioms and postulates.

cup, white, cup, in front of, table, steam, etc., all creating a *composition* of phenomenal distinctions. According to IIT, for physical systems to be able to instantiate this *composition:*

> The system must be structured: subsets of system elements (composed in various combinations) must have cause-effect power upon the system.
> (Tononi & Koch 2015: 7)

Understood this way, the composition which is given to us in our everyday experience is the result of cause-effect powers of elements in a system, which are able to bring about change to one another and the system as a whole, thereby revealing phenomenal distinctions. By 'cause-effect power' Tononi means the way in which those various elements interact with other elements in the system, and so *causes* state changes to those elements and the system as a whole; and how other elements in turn bring about *effects* on a particular element, thus changing the overall state of the system. For an element to 'intrinsically exist,' as Tononi puts it, an element must have cause-effect power upon itself, and must make a difference to the overall character of the state of the system as its states evolve and change over time.

One may disagree with the translation of these axioms into postulates, whether generally about the move from these axioms to postulates or the way in which they are translated, but I will set aside these disagreements to bring into focus the problem being discussed in this essay. For now, this will serve to give a general idea of how IIT is developed. IIT begins with *identifying* the *essential properties* of our experience – phenomenological axioms – and derives postulates that explain how physical systems might realize these axioms – physical systems postulates.

The five features of phenomenology and their corresponding postulates lead Tononi to posit a central identity of IIT, this will be of particular interest in §2 & §3 in examining IIT's use and definition of information:

> According to IIT there is an identity between phenomenological properties of experience and informational/causal properties of physical systems…The maximally irreducible conceptual structure (MICS) generated by a complex of elements is

> identical to its experience… An experience is thus an intrinsic property of a complex of mechanisms in a state.
> (Oizumi et al., 2014, p. 3)

What exactly does Tononi mean by "maximally irreducible conceptual structure"? According to IIT, the brain is composed of billions upon billions of elements (neurons/neuronal groups) and these elements take the place of information states – states which express some degree of information in their processing through the system as they fire, activating various regions of the brain. These elements do not exist distinct from one another. Rather they form integrated complexes that express information greater than the information generated by those elements independently of each other. According to IIT they would be maximally irreducible, as separating any of those elements from one another would decrease the amount of information which it is able to express. In this sense the whole is greater than the sum of its parts. This is what IIT means by integrated information, information which through integration with other informative elements in the system achieves a state that expresses more information than those elements did independently from one another.

To summarize thus far, experience according to IIT is identical to a MICS, those conceptual structures are composed of integrated information states, *ipso facto* experience is identical to integrated information states. Given that this is what IIT is arguing, the case must be made that the integration of information can give one a thorough account of phenomenal experience. If IIT can make such a case it would need to propose a direct response to the why-question of experience: why is it that integrated information states have a what-it's-like-for-me associated with their instantiation?

The identity of experience with the MICS is of central importance as it is due to this, depending on what information is according to IIT, that the theory may face a number of objections commonly raised against physicalism. An important thing to keep in mind from the short explanation of IIT which I have given in this section is that IIT bases its theory of consciousness on the notion of

information. We now need to make clear what exactly information is and how IIT defines and utilizes the notion in constructing a theory of consciousness.

Section 1.2: What is Information?

It is by no means uncontroversial what exactly is meant by invoking the notion of 'information,' since there is an ambiguity in what exactly one means by 'information'. Does one mean the common sense understanding of information as something which informs, and so gives one meaning or understanding, i.e. a semantic notion of information? Or do we understand information in terms of syntax, i.e. *how*, in the sense of what way does, information flows through a system, rather than the *meaning* of that information? Or do we mean some combination of the two? And, what exactly would such a combination look like? I suspect this ambiguity has something to do with a widely-held assumption that things which can be said to contain information must have some sort of meaning associated with them.

For the purposes of my argument, however, we need to understand what Tononi means by 'information' as he defines it in explicating IIT. Accordingly, I argue that if Tononi's use of information is as I have outlined it in the following subsection (§2.1), then his brand of IIT is committed to a physicalist position and so succumbs to the same problems as physicalist accounts more generally.

Section 1.2.1: What is Information According to IIT?

Claude Shannon, arguably the father of modern information theory/communication theory, in his paper *A Mathematical Theory of Communication*, is concerned with what he calls "the engineering problem" in communication. This problem can be summed up as: how does a particular state of the

system specify a particular message from the range of all possible messages expressible by that system? Since a system incapable of producing a vast array of possible messages to be transmitted would have very little use in communication, the system must be able to instantiate different possible messages. For example, when you type a message into your smartphone, it is able to transmit messages such as: "hey, what's up?", "what time for dinner?", "should I bring wine?", etc., and that is because: (i) it is a system that is able to transmit a vast array of possible messages; (ii) the system on the receiving end is one that is able to receive a vast array of possible messages which are sent to it; and, (iii) these particular possible messages are unknown at the time of design and so must be able to discriminate between a large number of eventual possibilities. As Shannon says,

> … semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design. (Shannon, 1948, p. 379)

For instance, take a six sided die (D6) as an example. This system which is composed of six possible states (the possibilities ranging from 1-6) all of which convey a particular message to whoever rolls the die. That particular system has the possibility of communicating a number of states, and so can be said to instantiate information equal to $\log_2(6) = 2.59$ bits of information.[5] For systems with more possible states, when a particular state is achieved they convey more information expressible as bits, i.e. a twenty-sided die (D20) is $\log_2(20) = 4.32$ bits. Another way to understand what it means for something to convey more 'bits' of information, is to say that information is the reduction of uncertainty, and the more uncertainty is reduced by the system the more information that system expresses. In the case of the D6 and D20 there is more uncertainty in the D20 system than in the D6,

[5] $\log_2$ express that there are two possible outcomes, in other words 'is' the case or 'is not' the case, 'yes' or 'no'. In the case of the dice there are six possible outcomes, and whatever number the dice shows, say it lands on the 6 side, can be said to be a 'yes' response to 6 and a 'no' response to 1-5. Another way to think of it is that it expresses the likelihood of 6 being chosen out of all possible options. This is expressible in terms of bits of information, which in this case, the physical system of the dice produces 2.59 bits of information.

and so when one possible state is picked out of either system, the system with more possible states has a higher degree of uncertainty reduced by instantiating a particular state of the system, i.e. 4.32 bits > 2.59 bits.

Before the die is rolled in either case the system is in a maximal state of uncertainty[6] as it's jumping around in your hand; but once it is thrown and lands on any of the possible states 1-6 or 1-20, in either case, the overall uncertainty of the system is reduced and that particular state (message) conveys information equal to 2.59 or 4.32 bits respectively. When uncertainty has been reduced it is the same as information being express by that system. And the more uncertainty which is reduced the more information that is expressed.

Tononi diverges from Shannon's definition of information when it comes to physical systems instantiating information, as his conception concerns information integration. Tononi says his definition is vastly different from how information is used in common language and communication theory (communication theory is in reference to Shannon's notion of information), and merely stays true to the etymology of the term 'information' (Tononi and Koch, 2015, p. 8). Rather, for a physical system to instantiate information, he thinks it must "*specify* a cause-effect structure that is *the particular way it is*: a specific set of specific cause-effect repertoires – thereby differing from other possible ones (differentiation)"(Tononi and Koch, 2015, p. 8). Cause-effect repertoires are all the possible ways a particular element or set of elements can bring about changes to a system, or be affected by other elements, or sets of elements, in that system; thereby differentiating themselves causally from other elements that have an effect and can be affected by other elements. If elements have a cause and effect that is different from other elements' cause and effect on that system, then it can be said to have a

[6] To clarify the meaning of 'maximal state of uncertainty' I mean maximal as in the maximum amount of *possible states*, not infinitely possible outcomes, since the dice is not a system that can express an infinite amount of possible states but as in the examples above, only 6 or 20, respectively.

cause-effect repertoire. With this in mind, Tononi proposes a modified version of information that has a causal notion built in:

> [I]nformation refers to how a system of mechanisms in a state, through its cause-effect power, specifies a form ('informs' a conceptual structure) in the space of possibilities.
> (Tononi and Koch, 2015, p. 8)

Accordingly, information must be able affect the system and in turn be effected by other elements in that system. Tononi adopts *differentiation* to articulate this, thus showing the way in which the elements in the system specify particular cause-effect structures differing from other elements in the system. This form of information is reminiscent of Gregory Bateson's definition of information from his *Steps to an Ecology of Mind*, in which he gives the following causal definition of information:

> [T]he world of form and communication invokes no things, forces, or impacts but only differences and ideas. (A difference which makes a difference is an idea. It is a "bit," a unit of information.)
> (Bateson, 1972, p. 276)

To unpack the quote, what Bateson means by "differences that make a difference" is that some system can be said to convey information if it can bring about a change of state in another system. If some difference in the state of one system can bring about a difference in another system, then information has been conveyed. Understood this way, information is instantiated in a system when it is able to make a difference to that system, and constrain the possible past and future states of that system; only differences which make a difference count as information. This definition of information gives one a causal notion of information, one that characterizes information in terms of how it flows through and brings about changes to other elements in the system. But what about this leads to the phenomenology? Tononi defines the phenomenological axiom of *information* as:

> Consciousness is *specific*: each experience is *the particular way it is* – it is composed of a specific set of specific phenomenal distinctions – thereby differing from other possible experiences (*differentiation*).[7]
> (Tononi and Koch, 2015, p. 6)

What is it about information integrating that gets one the phenomenal distinctions which we experience? Consider this example, what is the difference between my experience of the view from the Chain Bridge in Budapest, overlooking the Danube, and from the Tower Bridge in London, overlooking the Thames? Aside from the obvious geographical difference between the two, they both afford a unique set of possible experiences. In one I have the *possibility* of seeing the London Shard, in the other I have the *possibility* of seeing the Hungarian Parliament. My neurophysiology has to be a system that can at any point in its operation discern the *differences* between these two vistas and any innumerable amount of other objects of experience. I experience the view from the bridge and discriminate in my environment a vast amount of small differences, which overall reduce the amount of uncertainty in my experiential field. Distilled to its core, according to IIT, phenomenology is a complex field of difference relations. This differentiation is thus accounted for by the differentiation of the internal elements from one another, according to IIT. Our experience of the world presents us with a large array of information, and if the system which produces consciousness is able to do that, it must in some sense, be capable of accounting for the informational states instantiated in experience.[8]

The key thing to take from Tononi's use of information is the notion of '*differentiation*'. The question then becomes is it possible to give an account of information as '*differentiation'* that can capture

---

[7] Emphasis in original

[8] Cerullo (2015) characterizes integrated information in what he calls the *principle of information exclusion* which is that the "level of consciousness is directly related to the amount of perceptual possibilities ruled out by the system" (Cerullo, 2015, p. 3). The characterization of phenomenology being a "complex field of difference relations" and Cerullo's principle above are subtly different. Cerullo's relies on the exclusion of "perceptual possibilities", whereas, I take it IIT is concerned with the internal differentiation of the mechanisms which compose the system from one another. In this sense, I take it that Cerullo's principle doesn't quite capture what is meant by IIT's notion of integrated information. It is about the differentiation of the elements themselves from one another, and that becomes reflected in our phenomenal experience, not merely what is excluded from our perceptual experience.

the 'what-it's-like' of our experience of the world? The following subsection (§2.2) is an examination of this question.

Section 1.2.2: The Problems with IIT's use of Information

If Tononi thinks that the way elements in a system express information is by way of differentiation, then one must be certain that differentiation, as Tononi has defined it, really captures what we want in explaining consciousness. Consider the example above once again. If, according to IIT my experience is composed of a highly organized collection of difference relations, and this is all done in my brain through integrated neurons and neuronal groups (those mechanisms which instantiate information states), where in this story of differentiation does the phenomenal character, or what it's like, of experience come from? Understanding information states as differentiation gives one the difference relations which exist between various elements in a complex. In other words, one can track the change of the mechanisms in a global context of the system, by tracking their differentiation from one another. This, presumably, will be instantiated by different neurons/neuronal groups firing in particular locations in our brain/nervous system – firings in different spatial locations – and neurons/neuronal groups firing at different times – firings in different temporal locations. Differentiation gives one an effective way to understand the relationship between these various elements in the system, which stand in unique informational relationships to other elements in the system.

If we are to understand 'information' in terms of "differences which make a difference" then these various informational elements standing in a unique set of spatial-temporal relationships will also have various causal relationships, i.e. how those elements affect and are effected by other elements in the system, bringing about a range of possible states of the system. If IIT is claiming we should

understand information in this way, then one is given a structural[9] and dynamical[10] account of information.

Are structure and dynamics alone enough on which to construct a theory of consciousness? It has been argued by David Chalmers (2003) that structure and dynamics alone will not suffice in giving a satisfactory account of consciousness. The problem with a physicalist account of the world is that it solely relies on structure and dynamics to construct a theory of consciousness; presumably such an account would give a detailed description of how physical elements and their spatial-temporal organization, along with how those elements evolve dynamically through the system, causes other elements to change. Such accounts are only able to appeal to more structure and dynamics, essentially providing a detailed explanation of *how* consciousness comes to be, but failing to provide an equally thorough explanation of *why* consciousness comes about in the first place. However, there doesn't appear to be any good reason to think that truths about consciousness are fully captured through appeal to only structure and dynamics (Chalmers, 2003, p. 120). This has become known as the structure and dynamics argument (Chalmers, 2003), namely that structure and dynamics alone are not enough to account for consciousness[11].

To put the structure and dynamics issue specifically in terms for IIT, why should it be the case that there is anything it is like associated with the relevant structural and dynamical properties of information presented by IIT? If IIT is to be considered a full-blooded account of consciousness it should be able to offer a response to this question.

In essence, there is a gap between explaining how integrated information states, which express difference relations, give rise to phenomenology. One should not just take for granted that

[9] Spatial-temporal relationships between physically instantiated information states.
[10] Range of possible cause-effects on that system, i.e. the states evolve and change dynamically over time given the cause-effect relationships of other elements those information states stand in a relation to.
[11] For a thorough overview of the structure and dynamics argument see Torin Alter (2016).

"differences which make a difference" can account for our everyday experience. If IIT is making the claim that the structure and dynamics of integrated information states in a system can account for experience, then it would appear that IIT encounters a serious problem to which it must have a response. As David Chalmers has put it, in expressing the hard problem of consciousness:

> …the structure and dynamics of physical processes yield only more structure and dynamics, so structures and functions are all we can expect these processes to explain. The facts about experience cannot be an automatic consequence of any physical accounts, as it is conceptually coherent that any given process could exist without experience. Experience may *arise* from the physical, but it is not *entailed* by the physical. (Chalmers, 1995, p. 12)

To apply the above quote directly to IIT, experience may *arise* from integrated information states, but it is not *entailed* by them. IIT seems to make the argument that if experience arises from the structure and dynamics of integrated information states, then it is entailed by those integrated information states, and so posits an identity to explain that entailment. Yet, this move should give one pause; just because consciousness might arise from integrated information does not mean that it is identical to integrated information. To echo the concerns raised by Chalmers with regard to physicalist accounts of consciousness, and apply them to IIT; experience may arise from integrated information states, but that does not necessarily mean experience is entailed by integrated information states.

For example, recall from §1 that IIT posits a central identity that experience is identical to the MICS. It may be the case that the MICS is a result of how physical elements in a system that express information are integrated, but it is another thing entirely for that MICS to be *identical* to experience. If one is convinced that structure and dynamics alone are not enough to explain consciousness, and that IIT's definition of information is a purely structural and dynamical one, then there cannot be an identity between experience and the MICS. This is because one is left with a gap from the structural and dynamical properties of integrated information and those properties of experience. IIT as it is currently explicated seems to skip a step in positing this identity. IIT has given us a detailed account

of *how* experience might arise from integrated information, but has yet to provide a convincing reason to suppose that experience is *identical* to integrated information. This leaves open the question of *why* experience is the result of integrated information, and so leaves open the hard problem of consciousness.

If we accept the view of information that Tononi appears to be advocating in IIT – a modified and further developed form of Bateson's definition – then because of its use of information, we are left with a dilemma of how integrated information accounts for the hard problem, as it tells us nothing of the story of how one gets from structure and dynamics to our everyday experience[12]. I have argued that this is a consequence of IIT's use and definition of information and not at all in spirit with the goal of IIT more generally – namely the goal of being a theory of consciousness that attempts to tackle the hard problem of consciousness (Tononi and Koch, 2015, p. 5). If IIT maintains a structural & dynamical notion of information, it doesn't appear likely that IIT will be able to account for the hard problem of consciousness. In §3 I bolster the structure and dynamics argument against IIT that I have made in this section by appeal to the explanatory gap argument.[13] Before I move on to §3, I first want to discuss some criticisms which have been raised by Cerullo (2011) and Searle (2013) against IIT's use of information to explain consciousness.

---

[12] In a recent blog post by Scott Aaronson (2014), in discussion with Giulio Tononi's reply to the post, David Chalmers and Scott Aaronson came to a consensus that IIT might offer a response to what they called the Pretty-Hard Problem (PHP). The PHP is the problem of picking out and predicting when consciousness is present in a system. Of course, this would mean it doesn't answer the traditional hard problem, but it would still put IIT a bar above other theories of consciousness, in so far as it would provide a powerful predictive tool in the scientific study of consciousness.

[13] It has been suggested that IIT might interpreted as a kind of emergentism. This may help IIT avoid the charge of being a purely physicalist account, but at a prohibitively high cost. Most physicalist accounts would deny strong emergence, since strong emergence is arguably inconsistent with the causal closure of the physical (Kim, 2005). I take it this would be a less desirable position for a defender of IIT. In his blog, Peter Hankins (2014) suggests that one of the defenders of IIT, Christof Koch, should rather hold an emergentist IIT, than a panpsychist one (as Koch claims himself to be (Koch, 2012)). Even if IIT were seen as an emergentist theory, then one trades avoiding my argument against IIT for a brute fact of nature, and still IIT would not be a robust explanation of consciousness in any useful sense.

Section 1.2.3: Understanding the Distinctions – Syntax vs. Semantics and Structure & Dynamics vs. Phenomenal

Much of the debate regarding notions of information have centered around the distinction between mathematical formulations of information, such as Shannon's notion of information (such notions we can refer to as syntactic notions), and semantic notions of information that attempt to understand how information acquires/expresses meaning. For the purposes of understanding the notion of information as it relates to consciousness I find this distinction inadequate, as it fails to make clear what is important about information as it relates to consciousness. Rather I have opted to frame the discussion in the previous sections in terms of structure & dynamics vs. phenomenal. I have done this for two reasons. Firstly, thinking of information in a purely mathematical/syntactic sense leaves out a vital notion of causation from understanding the dynamic quality of information that is required for understanding consciousness. I take it that the structural/dynamical features of a system can be quantified mathematically, and its syntactic structure mapped, but merely mapping out syntactic structure seems to leave out the meaning of the causal claims which are more naturally discussed with regard to structure & dynamics. Secondly, it's not clear that semantics fully captures what we mean when we want to understand the phenomenal aspect of information, since it is not at all certain that semantics is all there is to the phenomenal, thus leaving an important feature of what we are attempting to describe unrecognized. For example, it's not clear that the phenomenal experience of colors, shapes, etc., have any semantic features which are essential to their being experienced.

I think an important distinction to bring up is one IIT uses itself, IIT stresses that it is necessary to distinguish between *extrinsic* notions of information and *intrinsic* notions (Oizumi et al., 2014, p. 6). Here I think a parallel can be drawn between the syntax vs. semantics and the structure & dynamics vs. phenomenal distinctions: *extrinsic information* is concerned with syntax & semantics - how

information can be *quantified* from an extrinsic perspective and what that information *means* from an extrinsic perspective – versus – *intrinsic information* which is concerned with structure/dynamics & phenomenal – how information is *organized spatial-temporally* and *evolves dynamically* over time, and *what that is like* for the element in the system from the internal perspective. Clearly, the structure and dynamics of information alone would not be enough to capture the intrinsic perspective, since ultimately structure and dynamics can be quantified extrinsically. IIT seeks to explain how a system might gain an intrinsic perspective given a sufficient degree of integration, but it's not clear that as a result of sufficiently complex structural and dynamical properties of information an intrinsic perspective necessarily pops up. This is why the distinction is so important, and for the purposes of my argument so damning. Since IIT's causal notion of integrated information as *differentiation* is purely structural-dynamical, it fails to fully capture the *intrinsic* perspective, but merely quantifies the *extrinsic characteristics* of that physical system.

The best way to come to grips with engaging with the conception of information within IIT, whether to defend or critique it, is thus to adopt the right distinction. In the case of this essay, that is structure & dynamics vs. phenomenal, rather than, the traditional syntax vs. semantics. To make the need to focus on the right distinction more apparent, I would now like to discuss two objections which have been raised against IIT for explaining consciousness in terms of information, each of which has taken a syntax vs. semantics approach to the debate. I endeavor to show how attacking IIT on the grounds of syntax vs. semantics fails to: (i) meet IIT on its own terms, and so fails to argue against IIT's causal notion of information; and, (ii) further highlight the need to adopt the structure & dynamics vs. phenomenal distinction, over the classic syntax vs. semantics distinction, when discussing the relationship between information and consciousness.

Cerullo (2011) and Searle (2013) have raised worries for IIT with regard to using information to explain consciousness. Searle argues in his review of Christof Koch's (2012) book, *Confessions*: *of a*

Romantic Reductionist, that information cannot be used to explain consciousness, because information is an observer-dependent phenomenon, rather than an observer-independent phenomenon. Observer-independent phenomena would be things like electrons, rocks, galaxies, etc., those things which exist that do not require an observer, but would rather quite naturally exist despite humans observing them. This is in contrast to observer-dependent phenomena such as sonnets, novels, or papers on IIT, etc., that require an observer to realize their existence. Searle takes it that explaining consciousness in terms of an observer-dependent notion, such as information, would inevitably lead to such an explanation being circular in nature.

Ultimately though, this fails to take into account what IIT's project is attempting to do, it looks to describe the intrinsic features of a physical system, i.e. characterize information from an internal perspective. Searle's conception of information is an extrinsic one, since his objection concerns the dependence/independence of objects/systems relative to an observer. Koch and Tononi respond to Searle's objection to their use of information to explain consciousness thusly:

> IIT introduces a novel, non-Shannonian notion of information – integrated information – which can be measured as "differences that make a difference" to a system from its intrinsic perspective, not relative to an observer. Such a novel notion of information is necessary for quantifying and characterizing consciousness as it is generated by brains and perhaps, one day, by machines." (Koch and Tononi, 2013)

Koch and Tononi are correct to point out that criticisms on the grounds that information is an extrinsic phenomenon are only appropriately brought against Shannonian notions of information. Objections on these grounds fail to argue against IIT's causal notion of information, since objections regarding the observer relevance of information are only concerned with extrinsic notions of information. Because of this, Searle's argument fails to argue against IIT's notion of information, and thus fails to bring into question IIT's account of consciousness on these grounds.

Cerullo (2011) criticizes IIT's notion of information for similar reasons as Searle. Cerullo argues that it is not clear how invoking the notion of information[14] in IIT should be at all useful for IIT in the way Tononi wants it to be. Cerullo argues that if IIT is going to be a contender to account for the challenges for a theory of consciousness outlined by Chalmers (1996, 1995) then it must meet the constraints of structural coherence[15] and organizational invariance[16]. Cerullo concludes that IIT fails to meet these two constraints, and thus integrated information does not do the job which Tononi suggests it does. As Cerullo says, "A purely data-defined theory of information such as Shannon's lacks the ability to link information with the causal properties of the brain… Only by including syntactic, and most importantly semantic, concepts can a theory of information hope to model the causal properties of the brain" (2011, p. 58). If IIT had a purely Shannonian notion of information in the theory, I suspect Cerullo would be correct, but as I explained in §2.1 & §2.2, IIT has a causal notion of information, one that is much more reminiscent of Gregory Bateson's (1972) notion. As a result of this, I take it that the spirit of Cerullo's critique of IIT is on the right path, in so far as it points out that IIT's notion of information is problematic, but ultimately the critique is unsuccessful because IIT does not have a Shannonian notion of information.

Whilst Cerullo's and Searle's criticisms are on the right lines, in so far as, they point towards an issue with the use of information in IIT, their critiques overlook the non-Shannonian notion of information in the theory. By engaging with IIT's non-Shannonian notion of information the arguments which I have advanced herein constitute an improvement of those of Cerullo and Searle. Thus, perhaps the most important reason to diverge from the syntax vs. semantics distinction, is

[14] Cerullo claims that IIT is employing a notion of information such as C.E. Shannon, but as was explained in §2.1, IIT does not hold a Shannonian notion of information, so this might be an uncharitable characterization of IIT's notion of information. Here I wish to point out that although I also agree with Cerullo that there is an issue with IIT's notion of information, I disagree on what that notion of information is and why IIT's notion of information it is unsuitable to base a theory of consciousness.

[15] This constraint is meant to express that there is a correspondence between awareness and experience.

[16] This constraint is meant to express that systems with the same functional organization will have identical experience.

because of the notion of information at work in IIT. Syntax vs. semantics discussions are more applicable to Shannon's notion of information (an extrinsic notion of information) which IIT claims it does not have, and I have attempted to show that in this section. Since IIT argues it has a causal notion of information, I have chosen to frame the issue in terms of structure & dynamics vs. phenomenal, which I think more accurately gets at the heart of the issue for IIT's notion of information (an intrinsic notion of information). The following section (§3) bolsters the structure and dynamics argument against IIT that I have made in §2.2 by appeal to the explanatory gap argument.

Section 1.3: The Gap Between the Physical and Phenomenal

The explanatory gap argument takes the form of highlighting the epistemic gap between physical facts and phenomenal facts – to put it another way, they try to show that knowledge of all the physical facts does not lead one to knowledge of facts about our phenomenology. Generally, once the epistemic gap has been secured, those arguing against physicalism then infer an ontological gap. I take it that even just securing an epistemic gap between IIT's notion of information and experience will be enough to show the seriousness of the problem for IIT. In particular, showing an epistemic gap between physical facts and phenomenal facts would be particularly detrimental to IIT given that Tononi begins with evidence from our experience (phenomenological axioms) and translates those into how physical systems could bring about said experience (physical systems postulates). As was explained in §1, IIT begins with the evidence from our own experience and uses that evidence to develop its 'phenomenological axioms', which it then uses to derive a set of corresponding physical postulates for how physical systems realize those phenomenological aspects of our experience. If there is an epistemic gap resulting from IIT's use of information (one of the axioms and postulates), then

there may be good reason to doubt whether the other four axioms/postulates would hold as well, as these axioms and postulates are defined in terms of information.

Section 1.3.1: The Explanatory Gap Argument against IIT

The explanatory gap argument goes as follows:

1) Physical accounts explain at most structure and function.
2) Explaining structure and function does not suffice to explain consciousness.

---

3) No physical account can explain consciousness.[17]

Any physical account will involve an explanation of consciousness in terms of structure and functions because that is the purview of the physical sciences, and according to physicalism, all facts about consciousness are accounted for by physical facts. There are of course certain things which can have a full explanation in terms of structure and function, such as the fact that water is H_2O. Presumably an explanation of water as H_2O in terms solely of structure and function would be an exhaustive explanation of water. Such an explanation would be satisfactory since it tells us exactly why every instance of water is H_2O, and conversely why every instance of H_2O is water. Furthermore, such explanations of water and H_2O will also tell us at what temperature water/H_2O reaches a boiling point, at which point it freezes, what particular conditions must obtain for it to go through state changes, e.g. from a solid to a liquid, etc. None of these explanations require, nor hint towards, a grander explanation than the purely structural and functional one provided to us.

The problem with consciousness is: it doesn't seem to be the case that such an explanation purely in terms of structure and function would give us such an analogously exhaustive explanation. Recall that IIT posits an identity between phenomenal experience and the 'informational/causal properties of physical systems' – the MICS. If one is to get an exhaustive explanation of consciousness

[17] The argument, as it is formulated here, comes from Chalmers (2003), the original argument is given by Levine (1983).

in these terms, it should be analogous to the case of water being H_2O; one should be satisfied with the explanation that experience is information/causal properties, and conversely that information/causal properties is experience.

If we are to understand function as 'causal roles in the production of a system's behavior' as Chalmers suggests, then I take it that 'intrinsic cause-effect structures of certain mechanisms in a state' (as explained by IIT) satisfy the relevant causal role in producing the behavior of a system, as it is the intrinsic cause-effect structures that constrain the possible states mechanisms within which a system can instantiate, i.e. neurons in the brain. Furthermore, if we are to understand structure as spatiotemporal structures, then the overall "space of possibilities in their past and future" – all those possible mechanisms arranged spatially and temporally (neurons to other neurons) – is the entire structure of the overall system. In its entirety this definition of what information is, and thus consciousness, consists in a specification of the structural and functional properties of a system. The structural and functional properties of a system are not enough to explain consciousness. If information according to IIT is about how a mechanism through its "cause-effect power" and "space of possibilities" is nothing over and above structure and function, then IIT is committed to being a physicalist account of consciousness. If this is so, IIT succumbs to the same explanatory gap argument against physicalism. The existence of an epistemic gap due to IIT's use of a physicalist construal of information is an undesirable consequence to say the least.

Now let us give a revised explanatory gap argument specifically for IIT:

1) Integrated Information Theory explains at most structure and function.
2) Explaining structure and functions does not suffice to explain consciousness.

---

3) Integrated Information Theory cannot explain consciousness.

One might object that IIT is not solely a theory based on its construal of information, it is just attempting to make sense of our phenomenology and apply that to how physical systems might

instantiate phenomenal experience. So it may be objected that the theory is not lead by its construal of information, but rather the character of our own experience. The issue with this response is that IIT posits an identity between one's integrated information structures and conscious experience – which means it should cut both ways. Tononi's view is set up by taking as evidence our phenomenology and then positing physical systems postulates that are able to realize those phenomenological axioms in a physical system. But if one is not able to go the other way, start with the physical system postulates and derive the phenomenological aspects of experience, then something is terribly amiss. If there is an *identity* between the integrated information states and phenomenal experience, there should be no gap whatsoever. I fail to see how the austere physical language used to describe the physical postulates lead one naturally to the phenomenological aspects of our experience.

To motivate the explanatory gap, and to further call into doubt the explanatory power of IIT if the argument I have just proposed holds true, I would now like to turn to one of the possible predictions IIT argues is a consequence of the theory and show why it might not actually have this predictive power. In a recent paper by Tononi and colleagues (Tononi et al., 2016) they have argued that IIT offers explanations and predictions regarding the physical substrate of consciousness. Specifically, I am interested in one particular prediction they argue IIT makes regarding consciousness and its physical substrate: that "consciousness should split if a single major complex splits into two or more complexes" (Tononi et al., 2016, p. 10). Let us grant that because of how information integrates in the brain, the two hemispheres achieve a global maximum of Φ, and that when there is a bi-section of the corpus callosum this global maximum is separated into two distinct complexes. Despite this, there would still be an explanatory gap.

Such a prediction would seem to lend support for IIT as solving one of the so called 'easy problem', the easy problems of consciousness are those such as, the directedness of behavior, the relationship between language and thought, and more importantly, the integration of information in

the brain (Chalmers, 1995). IIT give us an explanation of how information integrates in the brain, and the fact that it explains and predicates the result of split-brain cases seems to provide strong support for that. Yet, it doesn't tell us why there is anything it is like associated with that information integration. Such an explanation/prediction of the theory still doesn't bridge that gap. It tells us *how* information integration occurs across the two hemispheres, but not *why* there is anything it is like associated with that information integration.

Furthermore, for the sake of argument, let us say that this prediction of IIT is tested empirically, and we find that when a major complex of integrated information splits into two separate complexes, that consciousness splits as well. This is essentially what occurs in split-brain cases, when there is a bi-section of the corpus callosum, leaving the two hemispheres of the brain detached from one another. Let's assume that IIT runs the experiments, and confirms this prediction on behalf of IIT, and finds that when one major complex is separated into two complexes, one has two local maximums of Φ, and thus a separation of consciousness. Does IIT really provide an explanation of this? Cerullo (2015, p. 5) calls into doubt the explanatory power of IIT in this regard, by showing that his own *faux* theory of consciousness would have the same prediction as IIT. Cerullo proposes a *faux* theory which he calls Circular Coordinated Message Theory (CCMT). Cerullo says "[t]he justification for CCMT is the self-evident property that consciousness is related to information traveling in feedback loops within a system (the principle of information circulation)" (2015, p. 5), the value of the degree of information circulation is signified by Omicron (O). Both have the same prediction, that when a major complex of Φ or O is separated, there will be two complexes each with a local maximum of Φ or O, respectively. Both seem to have equal explanatory power, one says that this can be explained because there is a great deal of information integration between the two hemispheres, the other because there are significant cortico-thalamic loops. Which explanation is better? It seems both IIT and CCMT have

equal predictive power. This would seem to at least call into question the weight behind such predictions of IIT.

Falling victim to the explanatory gap argument is a serious shortfall of IIT, as the theory is constructed with the hard problem in mind as the target. I don't think this is solely an issue with the theory, but rather with the definition of information utilized by the theory, because this is what commits IIT to giving a purely structural and dynamical explanation of consciousness. If one could change the definition of information according to IIT, that could avoid these obstacles, then IIT would be in a more robust position to tackle the hard problem.

If IIT falls so easily into a gap because of how information is defined according to the theory, the whole theory shouldn't be scrapped, but rather the definition of information. I leave open what such an account of information might be, as fully developing and defending such an account is outside the scope of this essay[18]. The goal has been to merely highlight the issue in IIT's definition and use of information. The next step for IIT, now that such worries have been raised, is either to show that the arguments I have given do not hold, or take on board the worries raised and offer a revised notion of information. I see no reason a more amenable notion of information cannot be developed. Such a notion of information will put IIT on a better track to solve the problem it intends to account for – the hard problem of consciousness.

[18] To at least indicate some possible notions of information that might be developed further to avoid these issues, one might look at Chalmers' (1996) dual-aspect account of information. Though it's not clear that a direct application of dual-aspect will avoid the worries raised in this essay, there may be some promising developments that could come from exploring dual-aspect as it relates to IIT. Another option, and one which Cerullo (2011) discusses, is the General Definition of Information (GDI) from Floridi (2009), though Cerullo dismisses it for the reason that it will face standard philosophical worries concerning meaning (Cerullo, 2011, p. 57). I share these concerns with Cerullo, as it doesn't appear GDI will be able to bridge the traditional syntax vs. semantics gap, or the structure & dynamics vs. phenomenal gap discussed in this essay, but nonetheless there may be some interesting developments that can come from further looking into the GDI as it relates to consciousness.

Conclusion:

In this essay I have shown that IIT is committed to a purely structural/dynamical notion of information, and because of this commits itself to a physicalist account of consciousness, thus leading IIT into a number of objections commonly brought against physical accounts. If IIT wishes to avoid these issues, which I argued there is good reason to think it should in §2 and §3, then it will need to rethink how it goes about defining information. The issues at play for IIT's definition of information are analogous to the issues at play in the hard problem of consciousness. One can capture the issues raised in this discussion of IIT's use of information as the *hard problem of information*: *why is it the case that there is any experience associated with the informational processes occurring in our brain?* The burden of proof falls to IIT and other information-based accounts of consciousness if they wish to avoid the issues raised in this essay.

I have endeavored to show that IIT should look at these issues in defining information when using it to construct a theory of consciousness. IIT would offer an incredible degree of explanatory and predictive power when it comes to consciousness, if integrated information is in fact quantifying consciousness. If IIT can come up with an alternative notion of information, then perhaps it may one day account for the hard problem of consciousness.

"In consequence, sensory experience presents itself to us as if it were the acquisition of information about intrinsic nature. But, very, obviously, it is not information about intrinsic physical nature, so the information Mary acquires presents itself to us as if it were information about something more than the physical. This is, I now think, the source of the strong but mistaken intuition that Mary learns something new about how things are on her release.

I still think though that we should take seriously the possibility that we know little about the intrinsic nature of our world, that we mostly know its causal cum relational nature as revealed by the physical sciences."

- Frank Jackson (2004)
"Postscript on Qualia"

"I've looked at clouds from both sides now
From up and down and still somehow
It's cloud's illusions I recall
I really don't know clouds at all"

- Joni Mitchell
"Both Sides Now"
(1969)

The first essay was something of a conditional, that is, if we accept the hard problem, structure and dynamics arguments, and explanatory gap argument, and IIT on their own terms, how would IIT fare against such worries. I argued that IIT would fail to be an adequate explanation of phenomenal experience as a result. However, to borrow from Joni Mitchell, having looked at "clouds" (consciousness) from both sides now, I think there's an issue with both sides of that debate. This resulted in the task of showing what I think is wrong with those anti-physicalist arguments and what is wrong on the side of the physicalist. The remainder of this dissertation is a an attempt at paving the way for this middle ground, in the hopes that what is developed will result in a new approach to the problem of consciousness and it's place in nature.

2. Intervening on the Causal Exclusion Problem for Integrated Information Theory

Introduction

Integrated Information Theory (IIT) of consciousness is a neuroscientific account of phenomenal experience, according to which conscious states are integrated information in a system (Oizumi et al., 2014; Tononi, 2008; Tononi et al., 2016; Tononi and Koch, 2015). Accordingly, IIT is a theory placed at the intersection between scientific and philosophical inquiries into consciousness. In this paper, we contribute to this ongoing dialogue by examining the causal framework within which IIT makes its claims. As a theory of consciousness IIT takes up the challenge of tackling traditional metaphysical problems concerning the complex causal and non-causal relationships between conscious experience and the processes through which it comes about. One such issue is the causal exclusion problem, according to which mental properties are systematically excluded from standing in causal relations to other properties due to their physical supervenience bases (Kim, 2005).

Recently, attempts have been made by proponents of IIT to rectify this potentially damaging problem for the causal claims of the theory (Hoel et al., 2013; Hoel et al., 2016). In this paper we argue, rather than resolving the causal exclusion problem for IIT, these attempts reveal problematic aspects of IIT's causal framework. Specifically, that the *informational* account of causation they adopt renders their response damagingly circular, chiefly due to the account's claim that the causal properties of a system are identical with its informational properties. Our goal in this paper is thus to provide a causal framework within which IIT can avoid the causal exclusion problem and maintain its unique central identity. The requirements on meeting this goal are tantamount to providing a causal framework that meets the following three conditions: (1) it has the resources to avoid the causal exclusion problem, (2) it does not inter-define causation and information; it is not *essentially* an informational account of causation, and (3) it remains compatible with the empirical data, methodology, and conceptual *aims* of IIT.

We argue that an interventionist causal framework can meet these three conditions for IIT. This paper is divided into two parts. The first (§1-§3) motivates both our claims that IIT must provide a solution to the causal exclusion problem and current attempts to do so reveal that IIT's reliance on an informational account of causation is deeply problematic. At this juncture, we will have provided

and explained the three conditions on a causal framework for IIT. In the second part (§4-§6) we show how interventionism can meet these three conditions. We detail how interventionism does not rely on a definitional relationship between information and causation (§4) and how the version of interventionism we present here has the resources to avoid the causal exclusion problem (§5). In the final section of the paper (§6) we show how this version of interventionism remains compatible with IIT.

Section 2.1: Introduction to IIT

Integrated Information Theory (IIT) of Consciousness proposes that consciousness is integrated information in a system, the degree of which is signified by the Greek letter, Φ. The quantity of integrated information – or consciousness – present in a system is quantified by Φ, which is the amount of information generated by a complex of elements, above and beyond the information generated by its parts (Tononi, 2008). IIT is developed two-fold, (i) it takes our everyday experience as evidence enough to posit five phenomenological axioms – those essential aspects of experience that are given to us in our everyday experience of the world. (ii) It takes these essential aspects of our experience to then derive a set of physical systems postulates – these are meant to describe how physical systems might realize such phenomenological aspects of experience.

IIT makes the claim that elements in a system, i.e. neurons in the brain, take on the role of information states. These information states can be said to express integrated information if the information generated by their integration is greater than the sum of their individual parts. To get clear on what IIT means by integrated information, let's take a look at the physical postulates of information and integration.

> **Information Postulate**: [T]he system must specify a cause-effect structure that is *the particular way it is*: a specific set of cause-effect repertoires – thereby differing in its specific way from other possible structures (*differentiation*). A *cause-effect repertoire* specifies the probability of all possible causes and effects of a mechanism in a state (Tononi and Koch, 2015, p. 7).

By this IIT claims that something is informative if it specifies a particular state out of the range of possible states in the system. For instance, a neuron can take on any number of possible states $S_{n1\ldots nx,}$ if it is able to *specify* some particular state of the system through its *cause-effect repertoire*, then it can be said to be informative. According to IIT, each element has a probabilistic role to play

in changing the character of the system as a whole, and this is done through its cause-effect power – the possible states that the element can take on, and by doing so influence the overall character of the system. For IIT it is not merely about the state of the system at a current time, it is also about what the current state of the system tells us about the possible past and future states of that system. Since according to IIT there are causal purviews[19] which a mechanism might have in a system, all those possible cause-effect repertoires. These indicate the possible past and futures states of those mechanisms and thus the system as a whole. This leads us into the next postulate.

> **Integration Postulate**: [T]he cause-effect structure specified by the system must be unified: it must be intrinsically *irreducible* to that specified by non-interdependent sub-systems (Φ>0) across its weakest (unidirectional) link: MIP = minimum information partition (Tononi and Koch, 2015, p. 7).

The integration postulate is meant to show that the amount of information expressed by a set of elements in a system express integrated information if those elements are irreducible to a smaller subset of elements. In other words, if an element is removed from the subset of elements and any degree of Φ is lost, then it can be said that those elements form integrated information and are thus *irreducible*. By making a 'cut'[20] one finds a partition of elements which cannot be cut further without loss of information, this is done to determine MIPs. The process of identifying MIPs reveals sets of irreducible elements, which according to IIT are *concepts*[21].

Next let's consider the exclusion postulate for IIT, which will be an important aspect of our discussion in §6. The exclusion postulate is meant to explain why one's experience is definite at any given spatio-temporal grain. That is to say, that an experience happens at a specific spatial grain (the spatial grain of, say, neurons/neuronal groups) and at a specific temporal grain (the time it takes for the neurons/neuronal groups to fire). IIT defines the exclusion postulate thusly:

> **Exclusion Postulate**: [T]he cause-effect structure specified by the system must be definite: specified over a single set of elements – not lesss[*sic*] or more – and spatio-

[19] Purview here is meant to capture the range of possible cause and effects various mechanisms play in a system, which can be captured by a mechanisms transitional probability matrix (TPM).
[20] 'Cut' refers to a partitioning of elements to determine which form subsets of integrated information.
[21] For the sake of completeness, IIT technically defines a concept as "[a] mechanism and the maximally-irreducible cause-effect repertoire it specifies, with its associated value of integrated information φ^{max}. The concept expresses the cause-effect power of a mechanism within a complex" (Tononi and Koch, 2015, p. 6).

> temporal grains – not faster or slower; this is a cause-effect structure that is maximally irreducible (Φ^{max}), called conceptual structure, made of maximally irreducible cause-effect repertoires (concepts) (Tononi and Koch, 2015, p. 7).

At any one point in a system, there will be a global value of integrated information, or rather Φ^{max}, and this conceptual structure will represent the global experience at a given time. All other conceptual structures with lower Φ-values will be excluded from the global experience. The exclusion postulate is meant to be an explanation of why that is.

Since IIT is a theory of consciousness, what is the way in which these *concepts* express themselves as one unified global experience? Take for instance the overall experience of looking out from a cruise liner and seeing the vast ocean in front. From that vista one is able to distinguish qualitative features of the various objects of experience, the water is blue, the deck is shiny from sea water, the wood of the deck is grainy, the air is cold, etc., each of these qualitative features of 'blue', 'shiny', 'grainy', 'cold', have a corresponding place in the cause-effect structure taken as a whole, and thus, a corresponding *concept*. The *concepts* would represent various features of that experience, and taken as a whole would be the totality of the character of that particular experience. Crucial to how IIT can account for the character of our global experience is through *the central identity*, which states:

> According to IIT there is an **identity between phenomenological properties of experience and informational/causal properties of physical systems**…The maximally irreducible conceptual structure (MICS) generated by a complex of elements is identical to its experience… An experience is thus an intrinsic property of a complex of mechanisms in a state.(Oizumi et al., 2014a, p. 3, emphasis added).

According to IIT, one's global experience at a given time is the MICS. The MICS is composed of all the *concepts* (complexes of mechanisms) and their constituent individual informative elements (neurons/neuronal groups). IIT claims that those properties of our experience are thus identical to the informational/causal proprieties of the system. If this is the case then the MICS itself, and consequently the *concepts* and the individual elements (IE) within those *concepts*, all exert causal influence on the system[22].

[22] A clarification is in order, one may well object that the MICS does not itself exert causal influence it merely as causal influence as a result of the individual mechanisms which compose the MICS. I think this matter has not been entirely

As a consequence of IIT's definition of integrated information as *differentiation*, mental causation is built into the identity of those phenomenological properties of experience and the information/causal properties of the system (MICS). Since macro-informational structures, such as the MICS, at least sometimes express cause-effect power greater than their micro-informational constituents, then the MICS can supersede the causal influence of the concepts/IE's (Hoel et al., 2013). Since the MICS is the global experience of the system at a given time IIT posits that the overall informational/causal properties of the system constrain the evolution of that system over time.

Section 2.2: The Causal Exclusion Problem & IIT

The causal exclusion problem comes in many forms (Bennett, 2007, pp. 324-328). Here, we draw on the most prominent version of the argument, presented by Jaegwon Kim (2005, 2011). The argument comes in two stages, the first denies the possibility of mental to mental causation. The second stage proceeds to deny the possibility of mental to physical causation, leaving mental properties with an, at best, epiphenomenal status.

The argument requires a few principles to get off the ground. The first, (a) is that the mental supervenes on the physical. For Kim (2005, p. 14) this principle is a requirement for any minimal form of physicalism and, as such, is supposed to be extremely compelling. It consists in the claim that the token instantiation of a mental property is metaphysically dependent on the token instantiation of a physical property, such that every change in a mental property is necessarily accompanied by a change in a physical property.

The second principle is (b) the 'principle of causal exclusion'. The causal exclusion principle posits that for any given event there can be no more than one distinct cause that is wholly responsible for the occurrence of that event, apart from in cases of 'genuine' over-determination. Rare instances of over-determination are presumed to be plausible enough, death by firing squad being an often-cited example, but the exclusion principle constitutes a constraint that over-determination is not *systematic*. The principle is supposed to capture the idea that the effects of *mental causes* could never be genuine cases of over-determination, or at least the conclusion that they are, would be wholly unsatisfactory (Bennett, 2007, p. 325).

---

settled by those developing IIT, but my understanding is that the MICS is what IIT considers the maximal and irreducible set of cause and effects of a particular system.

The final principle required (c) 'the causal closure principle', states that if a physical event has a cause, then it has a physical cause (Kim, 2005, p. 15). Again, this principle is supposed to be plausible, particularly in the conditional form presented here as it makes no claims about the causal relationship between non-physical events, and allows for the possibility of physical events that have no cause.

The argument then runs as follows: we start with the supposition that one mental property, M_1 causes the instantiation of another mental property M_2. Because of (a), M_2 must have a distinct physical supervenience base, P_2 upon which the instantiation of M_2 *depends*. As Kim notes, "Given that [P_2] is present on this occasion, [M_2] would be there no matter what happened before; as [M_2]'s supervenience base, the instantiation of [P_2] in and of itself necessitates [M_2]'s occurrence at *t*" (Kim, 2005, pp. 39–40). At this point it looks as though the instantiation of M_2 is guaranteed by two distinct events: its cause M_1 and its distinct supervenience base P_2. But, by (b) only one of these events can be responsible for the instantiation of M_2. The claim here is that it *must* be P_2 that is responsible because *regardless* of the occurrence of M_1, the very occurrence of P_2 *necessitates* the instantiation of M_2, thus the role of M_1 seems superfluous. This completes the first stage of the argument: mental properties do not cause the instantiation of other mental properties.

As for the second stage, consider that there may yet be a role for M_1 to play as long as something caused P_2. Let's suppose that the cause is in fact M_1. If M_1 is the cause of P_2 then a causal role is preserved for mental properties – they cannot stand in causal relationships to other mental properties directly, but only *indirectly* through causing the supervenience bases of mental properties. But this won't work either because of (c). If we are supposing that P_2 does have a cause, which *ex hypothesi* we are, then, because of (c), P_2 must have a *physical* cause, P_1 and once again M_1 is excluded from playing a causal role in this story.

The argument can be captured with a diagram as follows:

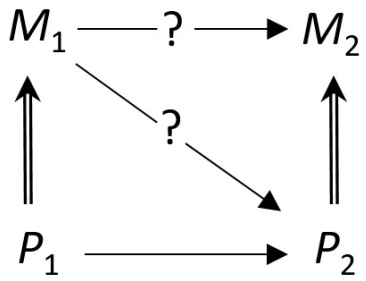

Figure 1 The Exclusion Problem

Each question mark represents a stage of the argument. The first considers the relationship between M_1 and M_2, and with the second concerning the relationship between M_1 and P_2. In both instances the potential causal role of M_1 is excluded, in the first instance by the supervenience relationship between M_1 and P_1, and secondly, by the causal relationship between P_1 and P_2.

How does the exclusion problem apply to IIT? Consider the following variation of *figure 1*, reformulated for IIT:

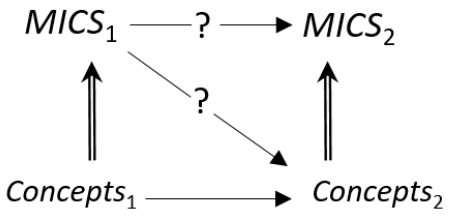

Figure 2 The IIT Causal Exclusion Problem

Since IIT argues that the MICS is the global experience of a system at a particular time, the MICS maps on to M_1 such that $M_1 = MICS_1$. Just as M_1 has a supervenience base, P_1, so too does the MICS, namely the collection of *concepts*, including their further subvenient base of individual informative elements (IEs). If this mapping of IIT's terms onto the traditional terms of the causal exclusion problem holds, then the argument runs analogously: according to (a) and (b) $concepts_2$ is sufficient for the instantiation of $MICS_2$, and because of (c) $concepts_1$ trumps $MICS_1$ when we turn to considering what property is responsible for the instantiation of $concepts_2$.

If this analogous argument holds then immediately IIT has a major problem for its central identity. Recall that IIT claims there is an "identity between phenomenological properties of experience and informational/causal properties of physical systems"(Oizumi et al., 2014, p. 3). The central identity claims that the phenomenological properties of experience, the total set of which at a given time amount to the MICS, are *identical* to informational/causal properties. However, if the exclusion problem holds for IIT, then the MICS *cannot be* causally efficacious. Thus, IIT has misidentified the properties of experience altogether, rendering the central identity false and placing the explanatory and predictive power of IIT in serious jeopardy. This provides our first condition to be met for a causal framework for IIT: (1) it has the resources to avoid the causal exclusion problem.

Section 2.3: IIT's Informational Account of Causation

Proponents of IIT agree that condition (1) is of genuine concern and have attempted to provide a solution (Hoel et al., 2013; Hoel et al., 2016). Here, we show how that proposed solution reveals a reliance on a problematic view of causal relationships. This in turn will lead us to the establishment of our second condition for a causal framework for IIT: (2) it does not inter-define causation and information; it is not *essentially* an informational account of causation.

IIT's approach to answering the causal exclusion problem has been to design a series of experiments showing that, at least sometimes, a system has more causal power when considered from a macro spatiotemporal grain than the micro scale upon which it supervenes. If this is correct then in contrast to the causal exclusion argument, proponents of IIT claim to have provided evidence for genuine causal emergence; situations in which macro states have more causal efficacy than their micro subvenient bases: 'the macro beats the micro' (Hoel et al., 2013). The methodology for these experiments is to show that macro structures express more *effective information* than micro structures, where EI is characterised as "a general measure for causal interactions…it uses perturbations to capture the effectiveness/selectivity of the mechanisms of a system in relation to the size of its state space" (Hoel et al., 2013, p. 19790).

The micro-structure is fixed and it's EI quantified, then the macro structure – the state that supervenes on that micro-informational structure – undergoes the same quantitative analysis and results are compared. If the experiments show that macro-informational structures exhibit higher amounts of effective information than their subvenient micro counterparts, then the causal power of the macro-informational structure surpasses the causal power of the micro-informational structures upon which they supervene (Hoel et al., 2016, p. 2). Once this is shown to be the case for sample systems, the conclusion drawn is that the causal exclusion problem is 'turned on its head': if events can only have one physical cause (causal closure) then it must be the micro state that is excluded rather than the macro state (Hoel et al., 2013, p. 19795). Why would measuring EI quantify causal power for states at various spatiotemporal scales? Because for IIT, "causation and information are necessarily linked" (Hoel et al., 2013, p. 19794).

Quantifying causal power as the amount of measurable effective information stems from IIT's central identity. Recall that the central identity posits an identity between "phenomenological properties of experience and informational/causal properties of physical systems" (Oizumi et al., 2014, p. 3). As noted above, it is this identity that allows proponents of IIT to claim that demonstrating that a system expresses more information at a macro scale than at a micro scale, is sufficient to show that phenomenological properties of experience (mental states) are causally efficacious over and above their subvenient micro states (physical bases). Crucially, establishing IIT's central identity requires adhering to a further identity: that the *informational* properties of a system are identical with its *causal* properties.

Without this identity, experiments showing the amount of information expressed within a system at a given scale would tell us nothing about the causal properties of that system at that scale. Thus, no conclusion about the causal properties of the system could be drawn from knowing the amount of information expressed and, in turn, no causal role can be attributed to the phenomenal properties of the system on the basis of IIT's experiments. More generally, then, IIT's central identity requires the adoption of an *informational* theory of causation. We can reconstruct IIT's argument as follows:

(1) If informational properties are identical to causal properties of a system, then the central identity holds,
(2) If the central identity holds, then IIT is not susceptible to the causal exclusion problem,
(3) Informational properties are identical to causal properties of a system,
(4) Thus, IIT is not susceptible to the causal exclusion problem

The recognition of the requirement for this second identity to hold – for an informational theory of causation to be true reveals the dependence of IIT's response to the causal exclusion argument on an informational theory of causation. Even if we accept that the central identity could show that IIT avoids the causal exclusion problem, we can now see that establishing the central identity itself requires the acceptance of a strong commitment to the relationship between information and causation, i.e., premise (3) must be correct.

The causal exclusion problem challenges a theory of consciousness to show how conscious or mental states can stand in causal relations over and above the causal relationships expressed by their subvenient physical bases. IIT builds its theory of consciousness upon the notion of *information* and more precisely the degree to which information is integrated in a system. So, the specific challenge for IIT is to show how states characterized by their degree of integrated information can have causal efficacy over and above their physical bases. But if informational properties are *identical* with causal

properties, then IIT gets its response by fiat – IIT has stipulated a solution by definition, rather than demonstrating one. If this identity holds then the claim that a state qualifies as conscious once it expresses a certain level of integrated information is validly re-interpreted as the claim that a state qualifies as conscious once it expresses a certain degree of causal efficacy. But of course, whether or not states with a specific level of integrated information can have unique causal capacities was precisely what we wanted to know in the first place when considering the causal exclusion problem. Thus, IIT's response to the causal exclusion problem seems to rely on the claim that conscious states are causally efficacious rather than demonstrate that they are.

To be clear, we are not arguing against those actively developing an informational theory of causation (Collier, 1999; Illari and Russo, 2014). Rather, our point is merely that such an option seems costly for IIT, specifically because IIT defines its causal relata – macro-information states and their constituents – in terms of information. Thus, for IIT, adopting an informational theory of causation leads to the circularity problems we have demonstrated here. Accordingly, we propose our second condition for a causal framework for IIT: (2) it does not inter-define causation and information; it is not *essentially* an informational account of causation.

IIT's use of a causal notion of information allows the theory to avoid one prominent objection from Searle (2013). Searle argues that information is an 'observer-relative' notion and one cannot use 'observer-relative' notions to explain consciousness. Proponents of IIT respond to Searle by arguing that the notion of information in IIT is causal and therefore not observer-relative (Koch and Tononi, 2013). However, if we are right, and IIT should avoid inter-defining information and causation, then IIT must modify its causal notion of information. Wouldn't this result in IIT being susceptible to Searle's objection? Although, this would mean a significant modification for IIT, this is not a new problem for IIT. IIT's notion of information has already been flagged as problematic (Cerullo, 2015, 2011; Mindt, 2017). Accordingly, we see our concerns regarding IIT's causal notion of information as contributing to this discussion, rather than creating a new problem for IIT. We contend that our proposed solution to the causal exclusion problem is more beneficial for IIT, than holding on to an already problematic notion of information and retaining the circularity which we described in this section. This would mean IIT needs to reassess their notion of information, but this is a task that those developing IIT must already do.[23]

[23] Here one might wonder, how can information be objective, if such objectivity is not grounded in causation? A possible solution could be to adopt Luciano Floridi's (2011) Information Structural Realism (ISR) in which information is objective in virtue of providing the interpretation of basic objects of reality – that is, structural objects. This is merely one possible avenue of investigation to see how IIT might rescue the objectivity of information. Suffice it to say, that if the arguments

Our final condition – (3) that the framework remains compatible with IIT's empirical data, methodology, and conceptual aims – follows from our own aims. We take it as a constraint on our proposal of a causal framework for IIT that it would enhance or lead to constructive modification of the theory in light of its own goals, rather than impose metaphysical restrictions on its development. With all three conditions in place, we will introduce the framework that we contend can meet them: the interventionist theory of causation.

Section 2.4: An Introduction to Interventionism

The interpretation of interventionism that we are concerned with here is closest to that provided by James Woodward (2003). This account is driven by the aim of bringing the causal modelling techniques developed by Judea Pearl (2000) and Spirtes et al., (2000), together with the conceptual framework of a broadly manipulationst approach to causation (Woodward, 2003, p. 38). This yields the following formulation of what it means for X to be a direct cause of Y:

> "A necessary and sufficient condition for X to be a (type-level) direct cause of Y with respect to variable set **V** is that there is a possible intervention on X that will change Y or the probability distribution of Y when one holds fixed as some value all other variables Z in **V**" (Woodward, 2003, p. 59).

Interventionism takes *variables* as causal relata as opposed to the standard properties or events found in many other accounts of causation. Variables represent magnitudes that are capable of taking more than one value, and have a well-defined notion of change (Woodward, 2003, p. 59). For example: temperature, T, can take the values $\{t_1, t_2, \ldots t_n\}$; Mass, M, can take the values $\{m_1, m_2, \ldots m_n\}$; and so on. Interventionism can easily represent more standard causal relata. Events can be understood as a variable that can take one of two values {occurred/not occurred}. Similarly, properties can be represented as {instantiated/not instantiated}. A variable set **V** is simply the collection of variables being represented as relevant to the claim at hand. Initially, then, under interventionism causal claims concern the relationship between changes in the values of variables, specifically that when one variable takes a value $X = x$ this results in $Y = y$.

---

we present hold in this essay, IIT shouldn't look for the objectivity of information in causation, but something else. What that something else is we leave open for further discussion and development, as it is outside the scope of this essay.

The change in the values of the variables under consideration must come about through an *intervention*. An intervention is a technical notion that stands for the manipulation of the value of a variable under specific conditions. Here we see the influence of the manipulationist approach to causation (Menzies and Price, 1993; Gasking, 1955; von Wright, 1974). Manipulationist accounts are motivated by the idea that we use causal claims, rather than mere correlations, to navigate the world around us via a process of *manipulating* and *controlling* our environment. In more abstract terms, it is precisely by actively adjusting the values of variables and tracking the results of those changes on other variables that we come to learn about the relationship between them. Interventionism takes this motivation firmly on board but seeks to refine that process of manipulation by supplanting it with a process that, its proponents argue, reliably gives rise to the sorts of relationships that we recognise as causal (Woodard, 2003, p. 28).

It is during the refinement process that we can see interventionism's other major influence: the work done in causal modelling to describe, formalise, represent and unpack just those sorts of relationships between variables that seem to fall under the category of causal. The work of Judea Pearl (2000) and Spirtes et al., (2000) has gone a long way towards refining the process of inferring or discovering causal relations from statistical data and developing a representational framework for these relations. Woodward's (2003, p. 38) overall project aims to build on Pearl's by providing: "an account of the meaning or content of just those qualitative causal notions that Pearl (and perhaps Spirtes et al.) take as primitive." At this stage, we can already see that the prospects for a complementary relationship between interventionism and IIT look good. Proponents of IIT readily employ and actively explore procedures that are "[…] akin to the calculus of interventions and the do(x) operator introduced by Pearl (2000), to identify causal relationships" (Hoel, et al., 2016, p. 2). A novel contribution of our argument to IIT will thus be to demonstrate how the conceptual tools provided by interventionism set within a broadly manipulationist framework, enables IIT to avoid the causal exclusion problem in a way that is compatible with the empirical data, methodology, and conceptual aims of IIT.

The process of defining an intervention begins with setting conditions for a manipulation to qualify as an *intervention variable* (**IV**). Woodward does so by offering four conditions that must be met, here we offer a truncated description of the conditions full details of which can be found in Woodward (2003, pp. 98–100).

(**IV**)

I1. *I* causes *X*

I2. *I* acts as a switch for all other variables that cause *X*, such that *X* ceases to depend on the values of any other variables.
I3. Any directed path that goes from *I* to *Y* goes through *X*.
I4. *I* is statistically independent of any other variable *Z* that causes *Y* and that is on a directed path that does not go through *X*.[24]

We noted above that interventionism attempts to unpack the meaning of causal claims represented by statistical modelling. One useful upshot of this approach is that we can use such modes of representation when discussing potential causal relations. As such, we will further explicate the conditions in **IV** with a directed causal graph[25]:

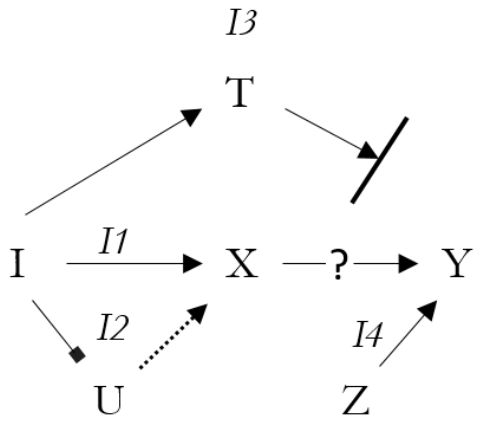

Figure 3 Directed Causal Graph of an Intervention

This directed graph represents an attempt to establish a causal relationship between *X* and *Y*, hence the question mark found in the arrow connecting these variables. *I* represents an intervention variable and does so because it meets all four conditions of **IV**. I1 is met because *I* causes *X*; that is,

[24] Technically, a directed graph is an ordered pair <**V, E**> with **V** being a set of vertices that represent the causal relata of the graph and **E** a set of directed edges connecting the vertices. Less technically, it's a graph with variables in it connected by arrows that stand for causal relations. A directed path is a route from one variable to at least one other via at least one directed edge, see Woodward (2003, p. 42 for more fine-grained details). All the graphs we work with in this paper are acyclic and we remain neutral on whether directed graphs have to be acyclic generally speaking.

[25] This is, in fact, a *toy graph*. Like a toy model, it is designed with a specific function in mind – here, to accessibly illustrate the conditions of **IV**. We have added some non-standard representative elements to do this work. The square-ended arrow is a preventer, illustrating that *X* is no longer causally dependent on *U because* of *I*, we feel this nicely captures Woodward's (2003, p. 100) sentiment that, ""any variable *U* (distinct from *I*) that was previously a cause of *X* is no longer such a cause," but with the proviso that it is not the *only* way the target variable, *X*, could be fully controlled by an intervention and thus not the only way to meet I2. Similarly, the solid line blocking the arrow exiting *T*, represents the idea that there *could be* paths for *I* to reach *Y* not via *X* (I3). However, in the present case that state-of-affairs does not occur, thus I3 is met.

there is a directed path from *I* to *X*. *U* is a variable upon which the value of *X would have* depended, so in accordance with I2, *I* acts as a 'switch' to break the dependency of the value of *X* on the value of *U*. *U* is an 'off-path' variable that has been controlled for in this graph. I3 is met because whilst *I* causes *T*, *T* is not on a path to *Y* (represented by the solid black line). *T* is on the graph to illustrate a potential way in which *I* could be on a path to *Y* that doesn't go through *X,* but in this case, it is not. Finally, I4 is met because *I* is not correlated with any other variable that causes *Y*, in this instance that is represented as *I* being statistically independent from *Z*.

Now that we have an intervention variable in place, we can understand an *intervention (***IN***)* as the following claim:

> "(**IN**) *I*'s assuming some value $I = z_i$, is an intervention on *X* with respect to *Y* if and only if *I* is an intervention variable for *X* with respect to *Y* and $I = z_i$ is an actual cause of the value taken by X" (Woodward, 2003, p. 98).

An intervention is the result of fixing the value of the variable that represents our potential cause, *X*, via the change in value of an *intervention variable*, which in turn is an exogenous process that meets the four conditions spelled out in **IV**. If the resultant change in the value of *X* changes the value of *Y*, then we can claim that *X* is a cause of *Y*.

Importantly for our purposes interventionism makes no *essential* reliance on the notion of information. In other words, condition (2) is met by interventionism. A variable must be capable of taking more than one value and there must be a 'well-defined' notion what such a change involves, but it by no means must represent the informational properties of a system. In §6 we will elaborate on how interventionism is nevertheless wholly compatible with the causal claims of IIT. Of course, that interventionism does not rely on an identity between the informational and causal properties of a system will not help IIT to retain its important central identity unless it can also meet condition (1): it has the resources to avoid the causal exclusion problem.

Section 2.5: Interventionism & the Causal Exclusion Problem

Consensus has been growing amongst proponents of interventionism that the theory of causation avoids the causal exclusion problem (List and Menzies, 2009; Raatikainen, 2010; Shapiro, 2010; Woodward, 2015a). However, an objection raised by Baumgartner (2009, 2010) has the potential to

threaten this consensus. In order to meet our second condition we will firstly illustrate how this dialectic has unfolded before providing our response to the latest objections raised by Baumgartner (2013).

Baumgartner presents a causal exclusion problem that applies specifically to interventionism. To see this, we can take the causal exclusion problem captured by *figure 2*, and apply it to an interventionist framework. Let $M_1 = X$, $M_2 = Y$, $P_1 = U$, and $P_2 = Z$. Accordingly, we're interested in whether *X* is a cause of *Y* and/or a cause of *Z*. To establish the nature of these relationships we need to perform an intervention on *X* as follows:

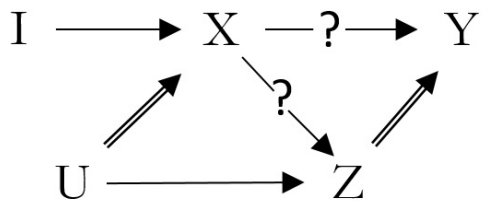

Figure 4 The Interventionist Exclusion Problem

To test these potential causal relationships, we must be able to intervene on *X*. However, Baumgartner notes that when intervening on *X* we intervene concurrently on the value of *U* and vice-versa, precisely because they are related by supervenience. This is a problem for interventionism because in order to test *X*'s potential causal relationship to *Y* or *Z*, we must be able to fix all off-path variables via interventions – including *U*. The supervenience relation between *X* and *U* will conceptually rule out being able to fix the values of these variables independently from one another and thus X will turn out to be neither a cause of Y nor Z. So, when physical properties and mental properties are represented on the same directed causal graph, it will be impossible for an interventionist to establish the causal efficacy of mental properties. If Baumgartner's objection holds, then IIT would be no better placed to respond to the causal exclusion problem for adopting interventionism.

Accordingly, the interventionist must say something about what to do in situations in which variables related by supervenience are under consideration on the same directed causal graph. There are (at least) two options pursued by interventionists:

(1) Consider variable sets to be 'well-defined' for the purposes of interventionist analysis, only if they do not contain variables related by supervenience.
(2) Refine the *theory* such that conditions governing interventions can accommodate instances of variables related by supervenience being represented in the same directed causal graph.

We will not be pursing (2) here and instead focus on (1).[26] This response denies that *figure 4* and the variable set it represents $\{M_1,P_1,M_2,P_2\}$ are well-defined for the purposes of interventionist analysis because interventionism does not extend to representing *non-causal dependency relations*. Non-causal dependency relations can be mathematical, definitional, logical, and, conceptual or 'metaphysical' (Woodward, 2015a, p. 326). This latter category includes supervenience relations. Thus, under interventionism we can remove the supervenience relations from the graph (variable set) and proceed to test sets that involve only potential *causal* relations, sets that are well-defined under interventionism (Eronen, 2012; Shapiro, 2010; Woodward, 2015a; Yang, 2013), for example

:

Figure 5 Potential Causal Relations

This response may strike some as question-begging, whilst interventionism may claim to be able to exclude supervenience relations from causal graphs and well-defined variable sets, the causal exclusion problem is designed precisely to ask the question *must* a theory of causation like interventionism represent non-causal dependence relations? Interventionists argue that this move is not question-begging because it derives from paying close attention to scientific practice. For example, Woodward (2015b, p.337-338) argues that if scientists were to control for supervenience bases when attempting to establish causal relations between variables through experimental interventions, the results would have absurd consequences such that scientists would not be able to establish causal relations between variables representing anything other than fundamental properties. Woodward's example is a variant of the 'expansion argument' according to which: if the causal exclusion problem is true for mental properties, then for any *X* and any *Y*, if *X* supervenes on *Y* then the causal relations in which *Y* stands exclude the potential for *X* to stand in any causal relations. In short: all non-fundamental properties (properties with no supervenience base) are excluded from standing in causal

[26] Woodward (2015b, pp. 332-335) explores (2) and Baumgartner (2013, pp. 16-24) provides a response. We will not be detailing that branch of the dialectic as we argue that option (1) has sufficient resources to avoid the exclusion problem.

relations, stripping sciences that make use of non-fundamental properties in their explanatory practices of true causal explanations.[27]

In reply, Baumgartner (2013) highlights a further stumbling block for this move. Whilst causal claims in interventionism are made only relative to variable sets (i.e., *X* is a direct cause of *Y* with respect to **V**) the conditions that comprise **IV** are *not* formulated relative to a variable set. So, if *I* does not qualify as an intervention variable for the set $\{M_1, P_1, M_2, P_2\}$ then it cannot qualify as an intervention variable for the set $\{M_1, P_2\}$. Thus, whilst interventionists can exclude non-causal dependency relations from 'well-defined' variable sets, they *still* cannot establish the potential causal relationships between the well-defined variable sets due to any possible intervention, *I*, failing to meet **IV** with respect to those sets.

What about modifying **IV** to relativize its conditions to variable sets such that *I* can qualify as an intervention variable *relative* to $\{M_1, P_2\}$, even if it fails to be an intervention variable relative to $\{M_1, P_1, M_2, P_2\}$? Baumgartner considers such a move but argues that it comes at a high cost, specifically that it renders interventionism unable to distinguish between difference making relations that stem from causal dependencies and difference making relations that stem from common causes. Consider a causal structure that exhibits a common cause:

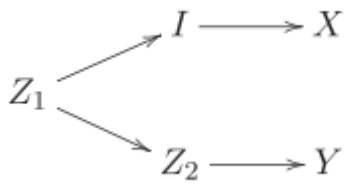

Figure 6 A common cause structure (Baumgartner, 2013, p. 12)

Next, restrict the set of variables under consideration only to {*I, X, Y*}. Baumgartner argues that *with respect* to the set {*I, X, Y*} *X* turns out to be a direct cause of *Y* because all four conditions of **IV** are met *relative* to {*I, X, Y*}, but clearly *X* is not a direct cause of *Y* for the structure represented in *figure 6*. In sum, moving to exclude non-causal dependency relations from what qualifies as a well-defined variable set for interventionist analysis will avoid the causal exclusion problem, but only at an unacceptably high cost for the theory of causation overall.

---

[27] For detailed presentations of the expansion argument see Burge (2003) as well as Baker (2003). For a response to the expansion argument see Kim (1998, p. 77*ff*) and for rejoinders to Kim's response see (Bontly, 2002; Ladyman et al., 2007; Marras, 2000)

We contend that Baumgartner's objection highlights an omission in Woodward's formalisation of **IV**, namely that nowhere in **IV** does it *explicitly* state that the value of *X* must be **set** by *I*, where 'set' is understood quite literally as 'manipulated via an exogenous process'. Rather, it states only that *X* must be caused by *I* and that changes in *Y* are correlated with changes in *X*. It is this discrepancy that allows Baumgartner's objection to work. In the common cause structure, changes in *Y* will correlate with changes in *X* whenever there are changes in Z_1 (and thus changes in Z_2). *I* and Z_2 will change as a result of changes in Z_1 and thus so will *X* and *Y*. But we take it that many proponents of interventionism (ourselves included) require something more than just passive observation of changes of the values of variables internal to the system. Consider that if the value of *X* were changed solely because an external manipulation of its value via *I*, then Z_1 would no longer be the cause of *I* (it's dependence on Z_1 would be broken) and in that case, changes in the value of *X* would *not* result in relevant changes in the value of *Y*, thus *X* would not be a direct cause of *Y*. We suspect that Woodward's (2003, p. 47) emphasis on interventions being an *exogenous* causal process is supposed to capture this idea:

> "interventions involve *exogeneous* changes in the variable intervened on […] It is important to understand that (i) the information that a variable has been set to some value is quite different from (ii) the information that the variable has taken that value as the result of some process that leaves intact the causal structure that has previously generated the values of that variable"

If we consider the variable set {*I, X, Y*} from Baumgartner's example as circumscribed from the common cause structure he presents, *I* is not an exogenous causal process with respect to *X* and *Y*. Furthermore, the example is based on the information that *X* and *Y* take their values as the result of some process that leaves the causal structure intact i,e., (ii). It is not based on information gained from the *setting* of *X* to a particular value i.e., (i). Why Woodward did not reflect this sort of requirement in the conditions **IV** is not clear to us, but clearly it *is* a requirement of interventionism as we understand it.

There are approaches to interventionism that do explicitly highlight the active setting of variables to specific values. For example, explicitly counterfactual approaches to interventionism. Take for example Raatikainen's (2010) 'active counterfactual' analysis of interventions. An active counterfactual takes the form:

> "If *X's* being x_1 were changed by an intervention to *X*'s not being x_1, then *Y* would change from being y_1 to not being y_1" (Raatikainen, 2010, p.354)

Consider Baumgartner's objection analysed with an active counterfactual: If *X's* being x_1 were changed by an intervention $I = i$ to *X*'s not being x_1, then *Y* would change from being y_1 to not being y_1. This counterfactual is false. If *X*'s value *was changed* from x_1 to not being x_1 by an exogenous manipulation of *I* to *i*, *Y's* value would not also change to a value other than y_1. Thus, *X* is not a direct cause of *Y* when analysed with an interventionist active counterfactual. Furthermore, one could discover this common cause structure by isolating well-defined variable sets and piecemeal testing these sets under active counterfactuals to build up a picture of the overall causal structure of the system. Of course, we do not suppose that the issue of interventionism's relationship to the causal exclusion problem ends here.[28] However, by considering some salient objections we have sought to refine and reinforce our claim that interventionism, set in a wider manipulationist context, has the requisite resources at its disposal to avoid the causal exclusion problem, thus meeting condition (2): it does not inter-define causation and information; it is not *essentially* an informational account of causation.

Section 2.6: Interventionism & IIT – The Perfect Fit

At this juncture, we have shown how interventionism can fulfil both condition (1) and condition (2): it avoids the causal exclusion problem and it does not identify causal properties with informational properties. What remains for our argument is to show that, as a causal framework, interventionism can meet condition (3): that it is compatible with the empirical data, methodology, and conceptual aims of IIT.

To begin with, let us examine the claims of IIT considered from the interventionist perspective we advocate. What does it mean for an IE/*concept*/MICS to be a variable in an interventionist

[28] To wit, recently Gebharter (2015) has demonstrated that from within a 'causal Bayes net theory of causation' (CBN), mental and physical properties can not only be modelled on the same graph, but in such situations, mental properties always fail to meet a productivity requirement, giving rise to an analogous causal exclusion problem. Gebaharter argues that because CBN can give a more detailed account of why the exclusion problem is valid, we ought to accept the exclusion problem. This argument paves the way for a more in-depth discussion of the conceptual and empirical reasons for, and against, choosing to model supervenience relations *as* causal relations. Whilst Gebharter's argument shows that CBNs can coherently choose to do so, our present purpose has been to defend the claim that interventionism can coherently chose not to.

framework? Woodward offers two criteria for a potential variable to meet within in an interventionist framework (i) that it can take on more than one value, and (ii) that there is a well-defined notion of change for the values of the variable (Woodward, 2003, p. 112). IE/*concepts*/MICS satisfy both these conditions because (i) they can all take on two or more values: IEs – {on/off}, *concepts* {ranges of values of φ^{max}}, MICS {ranges of values of Φ^{max}}. Such a value specifies the probability of how that system will evolve over time and influence other elements in the system to evolve probabilistically depending on the value of the variables around it, and (ii) there is a well-defined notion of change for these variables, which is given by the quantitative framework of integrated information in IIT.

Next, let's turn back to a now modified central identity under an interventionist framework:

> According to IIT there is an **identity between phenomenological properties of experience and informational properties of physical systems**…The maximally irreducible conceptual structure (MICS) generated by a complex of elements is identical to its experience… An experience is thus an intrinsic property of a complex of mechanisms in a state.

Notice that the central identity remains but with two important changes that result from IIT under an interventionist framework. The first is that the informational properties are not identical with the causal properties of the system. This is necessary to avoid our circularity objection detailed in §3 and is not a requirement of interventionism. The second is that IIT ought to construe its causal claims as *relational* rather than as *capacities*. Hoel et. al's (2013) experiments seek to show that at least sometimes the macro states have more *causal power* than their micro supervenience bases. However, such talk does not figure in interventionism. Recall that for interventionism causal claims are to be construed as *relations* between changes in the values of variables. This shift is not merely about how causal claims are described. Instead of arguing for a metaphysical thesis – downward causal exclusion, i.e., that the macro 'beats' the micro – IIT can harness the conceptual tools of interventionism, such as active interventionist counterfactuals, to avoid causal exclusion. For instance, recall our statement from §1:

> […] because of IIT's definition of integrated information as *differentiation*, mental causation is built into the identity of those phenomenological properties of experience and the information/**causal properties** of the system (MICS). Since macro-informational structures, such as the MICS, at least sometimes express

> **cause-effect power** greater than their micro-informational constituents, then the MICS can supersede the causal influence of the concepts/IEs.

Translated into an interventionist framework:

> Macro-informational structures, such as the MICS at least sometimes stand in causal relationships not found at the level of the micro-informational constituents (IEs/*concepts*). If this is the case, then the MICS can supersede the causal influence of the IEs/*concepts*.

Causal semantics such as "cause-effect power" sit less easily within an interventionist framework and so do not find their way into our translation.[29] Further, the problematic identity between "informational properties" and "causal properties" is no longer present under an interventionist causal framework. This modified statement does not conflict with IIT's empirical *aims* of showing novel causal relationships at the macro-level not found at the micro-level, with the added benefit of not needing to invoke the concept of downward causal exclusion.

This brings us to a final potential issue which concerns the exclusion postulate. Recall that the exclusion postulate is supposed to identify the spatio-temporal grain at which conscious experience arises – namely by identifying the state of the system with maximal Φ. However, in its current formulation it does so by availing of downward causal exclusion. Precisely because IIT inter-defines information and causation, identifying the state of the system with maximal Φ – the level at which consciousness occurs – just is to identify the state of the system that is causally efficacious and excludes all other states at a given time. Given that we have argued against IIT inter-defining causation and information – that the measure of Φ ought to be understood *non-causally* – what role is left for the exclusion postulate within the theory itself? Under the picture we offer here, conscious states of a system can be identified using measures of Φ and the potential causal relations that those states may

[29] To be clear here, we are not arguing that it is conceptually incoherent for interventionist causal claims to be grounded in causal powers. Rather, we think that avoiding metaphysically loaded semantics – like cause-effect power – will help proponents of IIT to make their position clear, without any loss of content. This is in keeping with our condition (3). We interpret the *strategy* utilised by proponents of IIT in avoiding the causal exclusion argument to be metaphysically neutral, rather than relying on metaphysical arguments that stem directly from adopting a causal powers view (eg., Gibb, 2013, 2015; Mumford and Anjum, 2011). Given that interventionism is a framework that also strives to remain metaphysically neutral (Woodward, 2015b) we are merely suggesting that IIT's causal semantics are adjusted to reflect this alignment of interests. Thanks to an anonymous reviewer for raising this objection.

stand in can be subsequently investigated using the particular interventionist framework we have demonstrated. Doesn't the need for the exclusion postulate simply dissolve?

Strictly speaking, we think that if the exclusion postulate *must* be understood as involving downward causal exclusion then it should be abandoned. If the exclusion postulate, in its current form, is to remain a central part of IIT, then those developing the theory must give independent justification for identifying the relevant value of Φ in overlapping systems; justification that does not rely on understanding Φ causally. The task of developing such a justification is left to those actively working on IIT, and unfortunately, lies outside the scope of this essay.

As we have demonstrated, the exclusion postulate is not required in order to avoid the exclusion problem, rather IIT can avail of the interventionist framework given in this essay to do this work. Either it should be abandoned or proponents of IIT leave need to develop new justification for the postulate in light of our main argument – that is, to provide independent justification for the exclusion postulate that doesn't rely on the circular definition of information and causation we've highlighted in §3. Importantly, what we present here remains consistent with our third condition; that what we propose is compatible with the *aims* of IIT. This condition is not that our proposal must satisfy all aspects of IIT as it is currently formulated, rather, what we propose conforms to the ultimate *aims* of IIT and its general approach to explaining consciousness. Thus, we can maintain that that condition (3) is met: interventionism remains compatible with the empirical data, methodology, and conceptual aims of IIT.

Conclusion

Our objective in this paper was to show that adopting a specific interventionist causal framework allows IIT to avoid the causal exclusion problem, whilst avoiding (i) the assumption of a damagingly circular identity between information and causation (§3) and (ii) defending a problematic notion of downward causal exclusion (§6). We do this by satisfying the three following conditions, which we set as constraints on a causal framework for IIT at the outset of this paper: (1) it has the resources to avoid the causal exclusion problem, (2) it does not inter-define causation and information; it is not *essentially* an informational account of causation, and (3) it remains compatible with the empirical data, methodology, and conceptual *aims* of IIT. We showed that condition (1) is satisfied through application of the specific version of interventionism we explained in §4 and applied to the causal exclusion problem in §5. Condition (2) is satisfied because the version of interventionism which we

advocate does not *essentially* rely on inter-defining information and causation (§4), thus avoiding the damaging circularity which we raise against IIT in §3. Finally, we satisfy condition (3) by detailing how IIT can be translated into the interventionist framework we advocate for in §6.

In Eudoxia, which spreads both upward and down, with winding alleys, steps, dead ends, hovels, a carpet is preserved in which you can observe the city's true form. At first sight nothing seems to resemble Eudoxia less than the design of the carpet, laid out in symmetrical motives whose patterns are repeated along straight and circular lines, interwoven with brilliantly colored spires, in a repetition that can be followed throughout the whole woof. But if you pause and examine it carefully, you become convinced that each place in the carpet corresponds to a place in the city and all the things contained in the city are included in the design, arranged according to their true relationship, which escapes your eye distracted by the bustle, the throngs, the shoving. All of Eudoxia's confusion, the mules' braying, the lampblack stains, the fish smell is what is evident in the incomplete perspective you grasp; but the carpet shows its true proportions, the geometrical scheme implicit in its every, tiniest detail.

It is easy to get lost in Eudoxia: but when you concentrate and stare at the carpet, you recognize the street you were seeking in a crimson or indigo or magenta thread which, in a wide loop, brings you to a purple enclosure that is your real destination. Every inhabitant of Eudoxia compares the carpet's immobile order with his own image of the city, an anguish of his own, and each can find, concealed among the arabesques, an answer, the story of his life, the twists of fate.

An oracle was questioned about the mysterious bond between two objects so dissimilar as the carpet and the city. One of the two objects – the oracle replied – has the form the gods gave the starry sky and the orbits in which the worlds revolve; the other is an approximate reflection, like every human creation.

For some time the augurs had been sure that the carpet's harmonious pattern was of divine origin. The oracle was interpreted in this sense, arousing no controversy. But you could, similary, come to the opposite conclusion: that the true map of the universe is the city of Eudoxia, just as it is, a stain that spreads out shapelessly, with crooked streets, houses that crumble upon the other amid clouds of dust, fires, screams in the darkness.

- Italo Calvino

"Invisible Cities" 1972, pg. 96-7

The night sets softly
With the hush of falling leaves
Casting shivering shadows
On the houses through the trees
And the light from a street lamp
Paints a pattern on my wall
Like the pieces of a puzzle
Or a child's uneven scrawl
…
And the pattern still remains
On the wall when darkness fell
And it's fitting that it should
For in darkness I must dwell
Like the color of my skin
Or the day that I grow old
My life is made of patterns
That can scarcely be controlled

- Simon and Garfunkel
"Patterns"
Parsley, Sage, Rosemary, and Thyme (1966)

The question we are left with is the following: which is the true pattern of our phenomenal experience? The world of which are experiences are about (Eudoxia)? Or rather the intricate threads of our own brain (the carpet)? In the following essay I explore what IIT's answer is to this question. I think treating the answer to these two questions as an either/or situation would be imprudent. Rather, I think the answer is best found in the relationship between the carpet and the city, the relationship between the threads and the streets, between the internal structure of our mind and the world of which that mind is engaged. My aim in the coming essay is to try and shed light on this matter.

3. The Marks of the Mental and Integrated Information Theory

Introduction:

There have traditionally been two contenders for the coveted title of *the mark of the mental*. The mark of the mental is supposed to capture the character of those states that differ significantly enough from the physical states they are associated with to be called mental. When looking for a mark of the mental one is concerned with finding those states which can be decidedly called 'mental' as opposed to say 'physical.' There have arisen two strong contenders for the mark of the mental over the discussion in philosophy of mind which are the following: (1) phenomenal experience, that is, to borrow a phrase from Nagel (1974) the what-it-is-like associated with our conscious mental states, and (2) the intentional nature of our mental states, that is their directedness at the objects (whether real or abstract) of our mental states.

This essay will be investigating one particular theory of consciousness – Integrated Information Theory (IIT) of Consciousness – and where it stands with regard to the two contenders for the mark of the mental. I will begin by first giving an explanation of IIT to the extent we need to get a grasp on IIT's account of how phenomenal experience (qualia) arises as the result of certain information-theoretic processes (§2), those being how integrated information gives us an explanation of how the qualitative aspect of consciousness arises. It's my hope that this explanation of how qualia arises as a result of integrated information will be useful for understanding the broader philosophical implications of IIT and inform the debates that have been ongoing in the philosophy of mind. In §3, I will then look at the issue of intentionality and what view IIT as a theory falls into; I will offer two views that I think are broadly implied by IIT, those being Horgan & Tienson (2002) and Mendelovici (2018). I will argue that although both are plausible interpretations of IIT's claims, Horgan & Tienson (2002) offer the most economical interpretation and the one that is most beneficial theoretically moving forward. My ultimate hope is to show that according to IIT thinking of qualia and intentionality as mutually exclusive is a mistake and rather IIT's view of how phenomenal experience arises is intimately related to what those conscious mental states are directed at (i.e. their intentional nature).

Section 3.1: The Preliminaries of IIT

Integrated Information Theory (IIT) of Consciousness proposes that consciousness, or rather phenomenal experience, is integrated information in a system, the degree of which is quantified by a function denoted by the Greek letter Φ. The quantity of integrated information – or consciousness – present in a system is quantified by Φ, which is the amount of information generated by a complex of elements, above and beyond the information generated by its parts (Hoel et al., 2016; Tononi, 2008). The rough idea is that one's hardware, in the case of humans the neurons/neuronal groups in the brain, "hang together" in unique ways. The structure those individual elements or groups of elements form amounts to a unique state, one state out of a mind-boggling large number of states that systems can have (given ~100 billion neurons in the brain, and their trillions of connections). There's nothing unique at this stage, as any information-theoretic or computational view of the mind would posit something similar to this. IIT's unique contribution is the degree and manner in which all these individual elements hang together, the unique relationships these elements stand in, relative to the rest of the elements in the system. IIT claims that it is not the reducible entities which figure into phenomenal experience, but rather the irreducible ones. Scientifically investigating consciousness then for IIT amounts to discovering how to characterize these relationships that emerge from the interaction of the parts distinct from the individual behavior of the elements themselves. To what extent are elements constrained by one another and what this tells us about pinpointing specific instances of phenomenal experience, both quantitatively and qualitatively. I hope the current work might go a bit of the way in convincing people that IIT is poised to tackle the qualitative aspect, perhaps not exhaustively as of writing this essay but in the future.

IIT makes the claim that elements in a system, i.e. neurons in the brain, take on the role of information states. These information states can be said to express integrated information if the information generated by their integration is greater than the sum of their individual parts. The precise mathematical way to determine this is still a matter of debate, regarding what may be the most computationally efficient method one can use to establish the extent to which various elements in the system are ultimately integrated. At the time of writing this essay, IIT 3.0 (Oizumi et al., 2014) is the canonical formalism.[30]

[30] See Tegmark (2016), Barrett & Mediano (2019), and Mediano, et al. (2018) for discussion on improvements or modifications to the formalism of measuring integrated information.

Perhaps an easier way of explaining how IIT argues phenomenal experience comes about is to paint a picture, or rather, by looking at one. The below image is a beautiful mess of confusion, as anyone that has stood staring at a Jackson Pollock can attest.

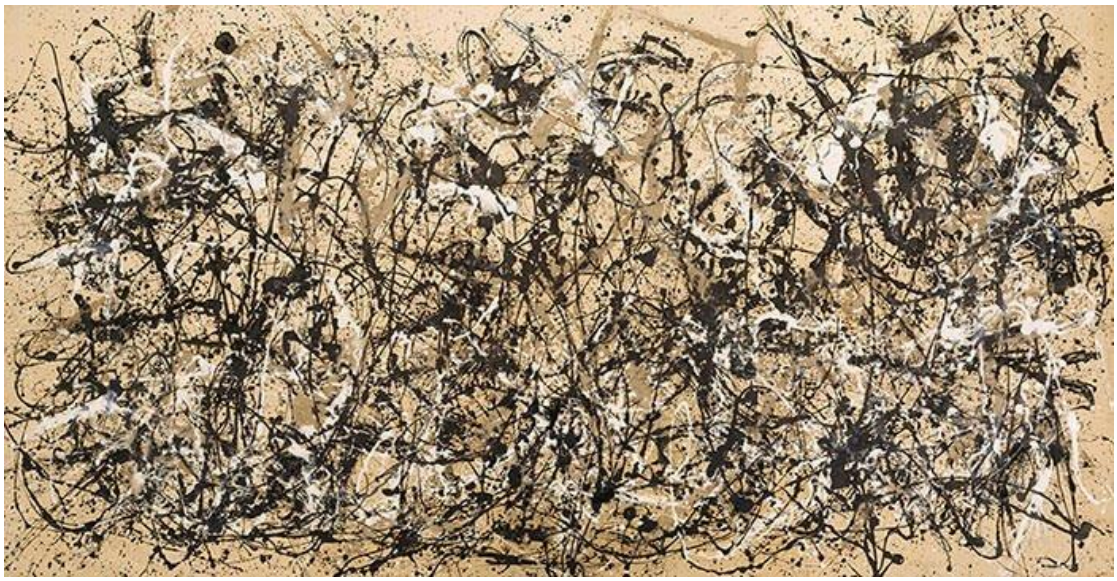

Figure 7 Autumn Rhythm (Number 30), Jackson Pollock

The question at hand is how the felt experience of looking at such a piece comes about in the first place? How does the squishy and pulsing grey matter of my brain produce such a vivid array of shapes, colors, and form, to produce the felt quality of such an artwork? We'll examine this painting using the same broadly Cartesian starting point and method which IIT itself employs.

The first starting point is that it is undeniable that you are having an experience of the painting at this moment, or more accurately that you are having an experience full stop. That is to say, there is something it is like for you to be reading the words scribbled on this page and to be examining the accompanying image, the system (that is you as a conscious subject) is currently undergoing a state for itself, that is to say you're having an experience from the intrinsic perspective (intrinsically). We thus are given the first axiom, that of intrinsic existence (axiom 1, intrinsic existence). Next, reflecting on the properties of that experience, it seems to be composed of a number of different phenomenal distinctions which are reflected in the phenomenology; all the colors, shapes, different objects, and spatial location in my visual field, etc., present a compositional structure to my experience (axiom 2, composition). Aside from presenting this compositional structure, the phenomenal distinctions in my experience are also informative, that is they're specific in form thus differentiating themselves from

all the innumerable other experiences I might be having at any given moment. We happen to be experiencing Jackson Pollack's Autumn Rhythm at this very moment, but the specific form of this particular experience, composed of its innumerable phenomenal distinctions, is only one of the innumerable other combinations of phenomenal distinctions one might experience (axiom 3, information). The system responsible (i.e. the human brain and its numerous neurons/neuronal groups and connections) is able to experience not only this particular painting, but any number of other possible paintings, or TV shows, and morning traffic jams, and spilled cups of coffee, etc., ad infinitum. My experience at any given time is as much about the experience I am currently having as it is about all the experiences that are ruled out by my having that specific one. This reduction in the amount of possibilities (or more precisely the uncertainty) is why one's experience is so informative. One's experience is also unitary, that is I am not having multiple distinct experiences at one given time. One does not experience the left visual field and the right visual field of the Pollock painting as two separate experiences, but rather one unified experience. This unitary experience is thus irreducible to some subset of its component phenomenal distinctions and their relations to one another (axiom 4, integration). Finally, one's experience is definite, that is to say it is neither less than what it is, say the Pollock painting minus the color or more, the experience of the paints on the canvas and the molecules that compose that paint. It is definite in spatio-temporal grain, the grain of a typical human experience. One might change the grain of their experience by using some tool, say a microscope or a telescope, but the spatio-temporal grain is only ever one definite grain at a given time. My experience also flows through time at a specific pace, roughly a hundred milliseconds for a particular experience, not faster or slower. I don't experience every visual scene of my life in a flash nor spread over an eternity. In other words, my experience at any given time is maximal in terms of what it contains and what it excludes (axiom 5, exclusion).

These axioms are only half the story, the bones of the skeleton so to speak, on coming to grips with explaining the nature of qualia according to IIT. We now need to put meat on them to see if they can stand the test of empirical observation and philosophical reflection. These axioms are taken as the undeniable attributes of our experience, there may be disagreement about the wording, but they're taken as the five features that any physical system must accommodate in order to give an explanation of phenomenal experience. The axioms, those essential feature of our phenomenology, are meant to constrain the possible ways the physical substrate (i.e. the human nervous system) must be like in order to produce such characteristics of our experience. The method of going from axioms to the physical systems postulates is best understood as an abductive exercise. One makes an inference to

the best explanation based on the evidence we have (our own phenomenology) and translates those into a language suitable for characterizing the physical substrate that realizes that phenomenology.[31]

To begin with, IIT only ascribes existence to those entities that can have cause-effect power, that is anything that can be said to exist must be able to make a difference to other things/entities and have a difference made to itself. Moreover, IIT makes a distinction between intrinsic and extrinsic existence. Intrinsic existence is reserved for an entity that not only has cause-effect power in the sense above as anything that makes a difference to other things, but in addition that it makes a difference to itself-for-itself, so to speak. Take for instance my phenomenal experience, that experience makes a difference to myself, but not for any other subject or entity in my vicinity. Extrinsic existence applies to those entities that have cause-effect power merely in terms of making a difference to other things and having a difference made to itself by other things. In this sense things can exist either purely extrinsically, say a rock, or intrinsically as is the case for a wakeful human experiencer (as a consequence, my dreamless body as I slumber is an extrinsic entity[32]). Let's take a neurotypical human as a way to illustrate this difference. In this sense of existence, I am both a conscious subject, there is something it is like for me to navigate the world, and thus have an intrinsic perspective (I exist intrinsically). Aside from this, my body when I am under general anesthesia still exists, but only in the extrinsic sense, as in doctors can manipulate my tissues and organs with a scalpel while I'm unconscious (make a difference to my body) and my body can make a difference to their scalpel (resist the pressure of the doctors cutting through tissue). Perhaps such a distinction of existence isn't as counter-intuitive as it first seems, it's just a distinction between the *way* in which two things exist, which seems to match the way we think of the existence of conscious and unconscious systems/organisms more generally. I have moral deference to other conscious agents (those that exist intrinsically) in a different way than I do to inanimate objects such as chairs (which exist extrinsically), furthermore, I have different moral commitments to things that have the potential for intrinsic

[31] Whether one agrees with such a method or what problems there are in relying on such an abductive method to determine the properties of the physical substrate is a discussion for another time. The target of this essay isn't to point out all the problematic things with IIT, but rather, to come to grips with understanding qualia according to IIT. I will thus be taking much for granted in this discussion and overlook opportunities to point out philosophically problematic aspects of the theory. For a good discussion of the problematic method of axiomatizing our phenomenology in the way IIT does c.f. (Bayne, 2018), in which Bayne recommends that IIT abandon the problematic axiomatic approach to developing a theory of consciousness and recommends a natural kind approach to developing a theory.

[32] I want to note here an interesting third distinction which doesn't appear in the work of those developing IIT and this is the potentiality issue. What about those systems which at some given time may not be undergoing a phenomenal experience at some time (think here my dreamless body as it sleeps), but nonetheless, have the potential, under the right conditions, to be conscious? Are these systems qualitatively different than systems that exist merely extrinsically? Are they qualitatively different than system that exist intrinsically? It's not clear to me that entities that have the potential to intrinsically exist should be regarded as extrinsic entities but also unsure how to classify such entities according to IIT.

existence as well. I don't change my ethical commitments to other human beings merely because they're in a coma or in dreamless sleep.

With that explained we can now understand the first physical systems postulate, that a ***mechanism***[33] exists if it has cause-effect power, and intrinsically so if it has cause-effect power on itself. One determines this by taking some variable (in this case whatever is taking the place of neurons/neuronal groups in the model, i.e. some logic gate with a particular weighting) placing it in a ***cause-effect space***[34] and performing interventions to see what causes (past states) lead to certain effects (future states) above chance. This tells us what cause-effect repertoires a mechanism has on itself and other mechanisms in the ***complex***[35].

Following the second axiom, composition, one must explain the internal structure of an experience and how that structure is reflected in the physical substrate of consciousness (PSC). If a mechanism in the system contributes to the structured features of a given experience, then its ***distinction/concept***[36] must be reflected in the phenomenal character of that experience. To determine which mechanisms contribute to experience let's consider an idealized system consisting of three individual elements A, B, and C. There may exist in the system the individual distinctions A, B, and C, their second-order conjunctions AB, AC, BC, and the third-order conjunctions that is the set as a whole ABC. In this sense "[c]omposition allows for elementary units to form distinct high-order mechanisms having internal cause-effect power" (Tononi, 2017a, p. 245). They won't necessarily form high-order distinctions, this all depends on intervening on the variables, in this scenario A, B, and C and seeing to what extent they are irreducible to their individual components.

This hierarchically organized grid-like structure, IIT argues, is reflected in early spatial vision in how we compose our experience of space from minimal features such as dots, lines, and edges. The compositional structure of experience should be evident at each step of visual processing from the lowest level to the top. The grid-like nature of early spatial processing gives good indications that IIT

---

[33] For words in ***bold italics*** I will have a definition of IIT's precise usage in a footnote. I do this for two reasons, (1) I don't want to burden the body of the text with excessive quotations of definitions and (2) IIT sometimes uses words with a dissimilar use to their conventional use in philosophy. With that being said, a mechanism is defined as "[a]ny subsystem of a system, including the system itself, that has a causal role within the system, for example, a neuron in the brain, or a logic gate in a computer." (Oizumi et al., 2014, p. 4)

[34] The probability space of all the possible cause and effect repertoires within a system that mechanisms might have in constraining the past and future states of a system of mechanisms.

[35] A complex is "[a] set of elements within a system that generates a local maximum of integrated conceptual information Φ^{max}. Only a complex exists as an entity from its own intrinsic perspective." (Oizumi et al., 2014, p. 4)

[36] A distinction/concept is "a set of elements within a system and the maximally irreducible cause-effect repertoire it specifies, with its associated value of integrated information φ^{Max}. The concept expresses the causal tole of a mechanism within a complex" (Oizumi, et al., 2014, p. 4).

is on the right track for explaining how low-level features compose high-order and more sophisticated characteristics of experience we normally associate with our phenomenal experience of the world.[37]

Every experience which one has is also informative, in the sense that it specifies a specific form of this cause-effect structure. This form is composed of all the specific phenomenal distinctions specified by all the individual mechanisms in this cause-effect space. Each corresponding to the unique distinctions they make in the overall structure specified by the collection of mechanisms taken together and their ***relations***[38] to one another in the complex. The specificity found in my experience is reflected in the relations between the distinctions in the cause-effect structure and their relations in this regard are irreducible to some subset of those mechanisms and relations (relations have their own Φ-value and thus exist intrinisically, i.e. they make a difference phenomenally). The cause-effect structure must reflect this irreducibility of my experience and so the PSC must form an irreducible set of these mechanisms. The way in which an irreducible set of elements is discovered is by performing interventions on the system, that is setting the system and its elements into a particular state and mapping their state-transitions as you perturb the system (turn elements on or off, introduce noise into the elements, and so on). By performing these 'cuts' on the system one discovers the weakest unidirectional link between all these elements (minimum information partition, or rather φ^{Max}). The point of this exercise is to see to what extent all the individual mechanisms in a particular system 'hang-together' so to speak. That is to what extent the cause-effect structure of a system is *integrated*. If a set of elements can be subdivided into two distinct sets of elements and there is no loss in the amount of Φ associated with those elements, then those sets of elements aren't integrated and are rather two distinct cause-effect structures. Since IIT is attempting to discover integrated systems, i.e. on conscious experience, performing these cuts on the system is essential in determining whether a cause-effect structure is integrated or merely an aggregate collection of parts (that is, a meaningless assemblage of distinct elements that just so happen to be spatially and temporally co-extensive).

Finally, we must account for the maximality of phenomenal experience in the PSC, that is how is it the case that only one global experience occurs in a system at a given spatial-temporal grain? How

---

[37] This work on early visual processing is still ongoing at the moment of writing this essay and was discussed during a visit at the Wisconsin Institute for Sleep and Consciousness in the Fall of 2018 with those developing IIT.

[38] As Tononi says about relations: "just like the existence of a concept can be established by determining to what extent a single cause-effect repertoire is irreducible to independent cause-effect repertoires, the existence of a relation can be established by determining to what extent a set of concepts is irreducible to independent concepts. One can thus consider a conceptual structure as a set of concepts bound by relations. According to IIT, an experience is identical to a conceptual structure, hence phenomenal distinctions correspond to the particular meaning of individual concepts, and how the various phenomenal distinctions are bound together corresponds to the relations (context) that bind concepts together." (Tononi, 2017b, p. 628)

is it the case that when I am undergoing an experience that I am having this one and maximal experience at a given moment rather than say my left-hemisphere having an experience of the right visual field and my right-hemisphere having an experience of the left visual field, distinct from one another? The PSC must be definite in the same way that my experience is definite, as Tononi says "…the cause-effect structure specified by the PSC must also be definite – neither less nor more, at a definite spatio-temporal grain, thereby singling out a single cause-effect structure and *excluding* overlapping ones" (Tononi, 2017a, p. 247). We can understand exclusion through a simple example, the left vs right hemispheres of the brain. Each hemisphere would have a correspondingly high quantity of Φ, and therefore a correspondingly complex cause-effect structure, but these two cause-effect structures taken separately according to IIT, would have a lesser degree of Φ than the system as a whole (that is the conjunction of the left and right hemispheres) and thus their distinct cause-effect structures taken separately are excluded by the system's (the brain's) cause-effect structure taken as a whole. This exclusion is necessary to account for why overlapping cause-effect structures don't appear in our phenomenal experience *distinct* from one another. Otherwise we'd end up with a kind of homunculi view of consciousness in which an infinite number of little consciousness would be screaming out for our attention (a horrifying prospect)! The exclusion postulate gives us the final value that our overall experience at a given time corresponds to Φ^{Max}; the cause-effect structure with the highest degree of Φ, and thus the one that wins the competition to present itself as the overall experience we enjoy at any given moment.

We are thus presented in IIT with the following central identity:

> "The maximally irreducible conceptual structure (MICS) generated by a complex of elements is identical to its experience." (Oizumi et al., 2014, p. 3)

If we are to find an explanation of qualia in IIT, it is through the physical systems postulates in conjunction with the axioms. Just what are those properties of the system that give rise to the phenomenal character of an experience? I'll reserve this for section two. The preceding discussion has provided us with the essential features of IIT necessary for understanding the picture of qualia generated by IIT. The following section will be a more in-depth explanation of how qualia is generated according to IIT.

Section 3.2: Qualia as the Overall Structure and Quale as Substructures

Qualia is one of those contentious concepts which seems to induce mud-slinging across the divide on those who take it to be a feature of reality and those who do not. I'm merely interested in qualia in a rather uncontroversial way as whatever property of a system gives that system its phenomenal character, that is whatever property of a system gives it its particular intrinsic feel. I by no means want to invoke something supernatural in its use, since my main concern in this work is understanding how phenomenal experience can come about as means of a natural process, that is, an information-theoretic one.

When I use qualia I only want to use it in this rather restrictive sense to indicate what properties a system might have which are necessary for the presence of phenomenal experience in said system, whether that be man, animal, machine, or something entirely alien. Turning back to IIT, when one wants to know what the theory says about qualia, it's important to make clear that IIT is committed to realism about phenomenal content; that is, qualia are real in so far as they are features of a system that are susceptible to naturalistic investigation.

According to IIT the brain is best understood, at least with regards to how consciousness originates through neuronal interactions, as a complex structure of mechanisms and their relations. The argument is the following: If one were to merely look at the physical substrate of consciousness (PSC), one would only find a number of more similar-than-not neurons which are connected in a complicated way, all with a feedforward architecture with recurrent feedback at various stages of processing. This seems to be the standard picture presented to us in neuroscience, whereby neural processing starts once our sensory organs receive some signal (the eyes for vision, ears for audition, etc.), the signal is then processed at each level in their respective cortices until boom! you have the experience of Jackson Pollock's Autumn Rhythm number 30! All of this is a matter of tracking the sensory signals from input to output and, through the magic of the neural soup that is the brain, one undergoes some experience of the world. IIT sees this picture as missing some rather important structural and dynamical features of the neurons and their relationships to one another.[39]

To begin, IIT claims that just looking at the substrate and their synaptic connections from an extrinsic perspective, one misses out on the complicated and *monstrous*[40] activity going on from the

[39] Chapter four discusses this subject in more detail.
[40] Tononi's (in correspondence) preferred expression for describing the intrinsic structure and dynamics of the elements in a complex. That is what one sees when the causal structure of a set of mechanisms is unfolded.

intrinsic perspective of the system (the perspective one should be concerned with in explaining experience according to IIT). The three essential concepts in IIT to explain qualia are as follows: distinctions/concepts[41], relations (the connections between distinctions), and Q-folds. Below I'll go over one-by-one each of these concepts and how the conjunction of the three give us an account of qualia[42].

A *distinction* recall is defined as "A set of elements within a system and the maximally irreducible cause-effect repertoire it specifies, with its associated value of integrated information φ^{Max}. The [distinction] expresses the causal role of a mechanism within a complex" (Oizumi et al., 2014, p. 4). Let's unpack this, a distinction is a unique set of elements which specifies a particular cause-effect repertoire, one that is irreducible to some subset of its individual components. For example, take the grid-like neurons found in the V1 area of the visual cortex. These neurons have strong bilateral connections with one another and form a lattice structure. IIT posits that these neurons and this rather simple lattice structure are responsible for the phenomenal aspect of space[43] that is, how our phenomenal experiences seem to be situated in space and the qualitative aspect of the spatiality of experience. IIT argues that space has a rich phenomenology of its own. Let's take a 1-D grid of eight cells all organized laterally with bi-directional connections between the elements. The basic idea is, how do individual neurons (i.e. individual elements) detect features which are presented spatially, say dots, lines, and edges and how to explain the phenomenal character associated with that? On their own, individual elements such as the individual simulated neurons that compose our eight-node grid are fairly "dumb" and they themselves don't indicate a dot, line, or edge spatially. Say some neuron in V1 fires, in what sense does that individual element "know" or more accurately indicate that what it is detecting is a dot, a line, or an edge, spatially in vision (remember again, we are asking from the intrinsic perspective of the system and its elements)? The answer is, it doesn't! what it does do is say

[41] "Concepts" used to be the terminology for this feature of the theory, but is being changed to "distinction", as calling these mechanisms "concepts" caused undue confusion. Throughout the rest of this explanation I will be changing "concept/s" to "distinction/s" in order to avoid these confusions. This will hopefully avoid any undue confusion about what IIT means when it says some element or collection of elements in the system are a "concept."

[42] I'll be using qualia to refer to the overall qualitative character of one's experience and quale to describe those sub-classes of qualitative distinctions, when I walk through the Wisconsin Dells in autumn my overall experience of the leaves changing color and all the qualitative features of my global experience are what one might call the **qualia**, the individual distinctions which compose it, say the red patches of color occupying some of the leaves would be a **quale**.

[43] This work is currently at the early stages and I learned of it during my visit at WISC. Any errors in exposition, inaccuracies, or mischaracterizations are solely on myself, though I've endeavored to faithfully reflect the work in this short explanation.

"this rather than that,"[44] the story gets more nuanced when we add another layer of elements above the first and expand the grid to 2-D, we now have elements which respond to the two cells below it which are directly connected to those elements. We can continue stacking layers on top of those, each detecting invariants from the layers below. What we begin to see is a pyramidal structure of grids (similar to the grids of neurons we see all sandwiched on top of one another in the various sensory cortices) stacked on grids, all with strong horizontal and vertical connections. These pyramids of neurons we are left with are able to detect invariances in our environment, the idea being that pyramids can detect dots (say locations of extension in our sensory fields) or other such invariances, the more elements and the more invariants picked out by elements at the top of the pyramids the more features that can be selected by those collections of elements. Here we end up getting a numerous amount of distinctions (that is irreducible mechanisms composed of individual elements, which specify a unique cause-effect repertoire).[45]

The distinctions in the PSC which specify a unique phenomenal distinction, say some local quality like a patch of red mapped onto a region of space in our experience, represents the first step in understanding how qualia is explained according to IIT. Each element contributes to the overall cause-effect structure specifying a gargantuan amount of distinctions, and even more so the larger and more nuanced the system. Elements specify distinctions of increasing order going up the pyramid, thus causally constraining how that system dynamically evolves over time.

> "[Distinctions] can be specified by *low-* or *high-order mechanisms*, over *low-* and *high-order purviews*, depending on the number of units involved. In addition, [distinctions] can be *low-* or *high-invariance*, depending on the size of the equivalence class of compatible states they specify. For example, that a particular spot in visual space is light (as opposed to dark) is a concept with low order-purview (or low-order [distinction] for short) and low-invariance (only one state out of two is compatible), presumably specified by a unit in a low-level visual area. By contrast, the letter "A" is a concept with high-order purview and high-invariance, since it specifies a sizable disjunction (logical OR) of compatible states over a large purview (a specific conjunction of

[44] Similar to say a photodiode that detects light, it only indicates the presence of a photon hitting its sensor, but less than that since all it indicates is *this* rather than *that*. A photodiode doesn't indicate what triggers it "as that thing" but merely something.

[45] Much of this work on IIT of SPACE, invariances, and pyramids is still ongoing or only planned for future work by those working on IIT at WISC. But this rough sketch at least begins to make clearer what role IIT takes distinctions to be playing in generating qualia.

> oriented edges at any large number of locations in the visual field)." (Tononi, 2017b, p. 627)

The idea here, is that there must be some way of accounting for the fact that rather undifferentiable (largely, from the extrinsic perspective) elements such as neurons can give rise to the varied and highly differentiable experience that a typical wakeful human-being experiences on a daily basis. So why is it the case that a particular subset of neurons fire whenever I come across the letter "A" in my visual field? According to IIT there is a set of distinctions (high-order in the case of the letter "A") which pick out "A" whenever it appears in my visual field. High-order since picking out the shape of an "A" actually takes a great deal of elements and binding between different orders of mechanisms in order to compose an "A". The invariance is high as well, since "A" might well be accompanied by an "N", as well as a "D", to compose "AN" or "AND", or any number of other letters to compose any number of words in the English language.

The high-order distinction of "A" would be specified by a collection of elements at different locations in the pyramid. The particular collection of distinctions which specifies "A" is composed of lower-order distinctions, these all taken in conjunction are a *mechanism Q-fold*. As Tononi writes, "The set of mechanisms to which a particular unit or subset of units contributes is called a mechanism Q-fold (a fold in qualia space, a synonym for cause-effect space), where the order of the Q-fold is given by the number of [distinctions] it specifies" (Tononi, 2017b, p. 627). There's the further matter of how such a mechanism Q-fold figures into other higher-order Q-folds. The mechanism that specifies the distinction "A" may figure into more abstract distinctions such as the third-order distinction "Air" and thus the mechanism Q-fold has a **purview**[46] of possible constraints it can have on the character of the overall system at any given time (and, as a result, what constraints can be made on itself by other mechanisms). Say one were to inactivate, or kill off, the Q-fold/mechanism that specifies "A," one would effectively be losing all those other higher-order mechanisms that have "A" as part of their purview "Air" and "the Air is fresh today", and so on, thus the system would lose its ability to recognize mechanisms that had "A" as a Q-fold within its higher-order Q-fold. We have good reason to already suspect this occurs, if I were to lesion parts of your fusiform face area (FFA) effectively eliminating your ability to recognize faces *as faces*, a condition called prosopagnosia. Say I killed the mechanism that acted as the invariant for the recognition of your partner's face. You would still see

[46] Tononi's definition of a purview is as follows: "[t]he subset of system elements whose past and future states are specified by a concept [distinction] is called its *purview*." (Tononi, 2017a, p. 247)

all the features of their face, their eyes, their dimples, etc., but you wouldn't see their face *as their face*. You've lost the invariant for that face, and thus, lost that composite and integrated entity of their face (there's a considerable amount of money to be made for services in lesioning the invariants of ex's faces and presents some interesting commercial applications of IIT).

Distinctions and their corresponding Q-fold mechanisms and purviews give's one an account of how (1) phenomenal distinctions contribute to one's overall qualia and (2) how their higher-order compositions develop into more robust phenomenal aspects of our experience (Q-folds). We can understand distinctions and Q-folds as giving us the first part in our story about qualia, what individual quales are – those individual features of our overall qualia. But where does the overall qualia of an experience come from?

When looking at the matter from this perspective it becomes clear to what extent even certain trivial (at least at first glance) aspects of our experience take a monstrous amount of activity to be realized. There is the further matter of how these distinctions and their high-order conjunctions bind to one another to get the compositional and integrated nature of our experience.

Here it would be worthwhile to understand what is at stake and what connection this has to broader discussions that have taken place in philosophy of mind and consciousness. There are those that argue that one of the aspects in need of explanation concerning our phenomenal experience is the seemingly unified nature of it, that is, how all the various aspects of our experience (say the green of some tree in an area of my visual field co-occurs or is unified with the sound of an owl to the left and the tactile feeling of the soil beneath my feet) unify into one global or maximal experience? Bayne (2010) argues that what we need is an explanation of phenomenal unity one that captures what it feels like for the contents of our phenomenal experiences to present themselves in the unified manner which they have in our own phenomenology. Bayne (2010) opts for a tripartite conception of how we are to individuate the contents of our experiences, as he says,

> "The business of counting experiences is a messy one, and there is more than one respectable way of going about it. That being said, I suggested that experiences should be thought of in tripartite terms: an experience is to be understood in terms of the instantiation of a phenomenal property by a subject at a time. We can think of these instantiations as phenomenal events. And in light of this, phenomenal unity can be understood in terms of mereological relations between phenomenal events. At any one point in time one's stream of consciousness takes the form of a single highly complex

> phenomenal event that subsumes a number of less complex phenomenal events. It is the fact that these less complex events are proper parts of more complex events that accounts for their unity. The mereological relations between phenomenal events might be reflected in mereological relations between their vehicles but they need not be." (Bayne, 2010, p. 28)

There are of course those who disagree, such as Tye (2005) who thinks it wrong to speak of experiences as having a part-whole character, rather Tye argues that there is just one given phenomenal experience that has no parts. The no-unity view, such as Tye though, is obviously at odds with IIT's axioms and postulates of composition and integration and as such can be left to the side. There are however views which take it that the unity relation isn't one of subsumption (as is argued by Bayne & Chalmers (2003) and in Bayne (2010)) but rather one of co-consciousness. Dainton (2000) argues there is no need for some ontologically costly explanation of what unifies the contents of experience, it's merely that they are co-conscious, that is various phenomenal properties which fill our experiential space are unified by merely being related to one another phenomenally. As Dainton puts it "…we simply accept that diverse experiences are experienced together, as co-conscious. Co-consciousness is a basic experiential relationship, one about which there is nothing more to be said, at least while we confine ourselves to describing how things seem" (Dainton, 2000, p. 84). We are left with a question though, which is, once we move past the phenomenological analysis of the unity of consciousness, how does the physical substrate of consciousness form the unity which is present in our phenomenal experience? Why is it that when I stare out into the night skyline of Budapest I experience, unified as one scene, the lights of the Buda hills glistening, the sound of a bus behind me screeching its brakes, the bead of sweat rolling down my cheek, and the warm breeze blowing through the hairs on my arm, all unified as one total experience?

To answer this question by returning to IIT, I don't experience all these phenomenal distinctions separate from one another, like oil separated in a glass of water, but rather combined and seamless, like oil mixed with pigment to make paint. It's one thing for a group of mechanisms to specify certain qualitative features of one's experience (quale), but how do these all get composed and integrated into one unified experience (qualia)?

Relations are what satisfy this role in IIT, effectively by mixing the distinctions and their Q-folds (pigment) with the oil (relations). Relations serve the role of binding the Q-folds and their component distinctions to create one overall integrated structure (the compositional and integrated

nature of our phenomenal experience). How does one mix the pigment (distinctions & Q-folds) with the oil (relations) to make the paint (qualia)? As Tononi writes,

> "Whenever there is an overlap of the purviews of different [distinctions], and therefore of the constraints they impose, their meanings are not independent, but *bound* together (*composed*). Phenomenally, this is evident because experience is rich in bindings (relations among [distinctions])… An even larger amount of bindings occurs when I see a person's face, with its outline, location, eyes, nose, mouth, and so on. These various [distinctions] are not independent, since their purviews overlap, establishing various kinds of relations among them." (Tononi, 2017b, p. 628)

And,

> "Just like the existence of a [distinction] can be established by determining to what extent a single cause-effect repertoire is irreducible to independent cause-effect repertoires, the existence of a relation can be established by determining to what extent a set of [distinctions] is irreducible to independent [distinctions]. One can thus consider a conceptual structure as a set of [distinctions] bound by relations. According to IIT, an experience is identical to a conceptual structure, hence the phenomenal distinctions correspond to the particular meaning of individual concepts, and how the various phenomenal distinctions are bound together to the relations (context) that bind [distinctions] together."(Tononi, 2017b, p. 628)

Let's return to the Jackson Pollock (Fig. 7) from the explanation of the postulates in section one. It's one thing for all the phenomenal distinctions to be present in my experience, just as it's one thing for all the pigment to be spatially located at different points on the canvas in Pollock's Autumn rhythm. But the pigment does not merely occupy a space on the canvas, it's blended so to speak, and related to all the other pigments, their contours, shapes, locations, etc., the scene is compositional. The pigment not only binds to the canvas, it binds to all those pigments it's related to, contours bind to other contours, shapes to other shapes, locations to other locations, etc. According to IIT then, we not only phenomenally experience the distinction themselves but their binding to one another. These *relations* then are present in our phenomenology, and thus have a value of Φ associated with them. The Φ of the relations is computed as the intrinsic distance between the probability distribution of the distinctions in a cause-effect space. The extent to which the purviews of low- and high-order

mechanisms irreducibly overlap with one another. What relations tell us then is the extent to which distinctions are bound to one another. One can think of this intuitively in terms of, how far away are the two states from being identical, that is how likely it is that these two states from time t1 and t2 are identical. The larger the difference between the likelihood of these two states is the distance between these two states.

With an explanation of distinctions, Q-folds, and relations we begin to see the outlines of IIT's account of qualia. This explanation is by no means finished, and a good deal of development is still required for IIT to give a robust explanation of qualia according to the theory. For now, the picture that is beginning to emerge is one in which the overall structure of the system (it's cause-effect space, or qualia space as it's referred to in Tononi & Balduzzi (2009) is the qualia and the substructures, the individual distinctions and their composite Q-folds, are the quale. We are of course left with some relevant and interesting questions, such as how do local contents bind to the elements that specify them? Are these local contents bound to individual collections of elements (say neuronal groups)? Individual elements themselves (i.e. specific neurons)? Or some other combination of elements? Answers to these questions will have to be left for future work, but they nonetheless point to some interesting theoretical and practical directions for future research.

Section 3.3: Qualia, Intentionality, and Phenomenal Intentionality

If we are concerned with what counts as the mark of the mental, phenomenal experience (qualia) is surely one of those features, but there is another facet of our mental states that we need to account for, and that is their directed nature, their *aboutness*.

This is the focus of this section what our mental states are *about*. This is usually referred to as the intentional nature of our mental states, more specifically our experiential states.[47] Lycan (2015) defines an intentional[48] state as "[a]n intentional state represents an object, real or unreal (say, I'll have another or Pegasus), and typically represents a whole state of affairs, one which may or may not actually obtain." As Searle puts it "[t]o say that a mental state has intentionality simply means that it is about something" (Searle, 1984, chap. 4). So, we can understand an intentional state as that state that

[47] One may well be interested in intentional states outside the scope of our experiential states. I will however be restricting the scope of the present work to conscious mental states, as IIT is a theory about phenomenal experience, and will thus stay silent on the nature of unconscious intentional states.

[48] I will be using 'intentional' and 'representational', as well as, 'intention' and 'representation' interchangeably from here on.

represents some object. In terms of IIT, the object in question would be the cause-effect structure at some time which may or may not correspond to something in the world. It has often been argued that intentional states are propositional in nature that is they take the form of *that*-statements, such as 'I believe *that* it is raining outside.' These types of propositions indicate that the mental state or episode of consciousness is *about* something (as described by Rosenthal (1994, p. 349) in his discussion of identity theory). I side with Crane (2003) in that intentional states need not be propositional, and that sensory states or perceptions can be intentional though they need not be propositional. I, however, wish to go a step further and side with those who take a certain class of intentional states to be wholly constituted of phenomenal states (c.f. Horgan and Tienson (2002)). I will defend this interpretation of IIT's claims regarding intentionality in the latter part of this section.

Searle (1984, 1983) argues there is a difference between original intentionality and derived intentionality. Original intentionality is of the kind that humans enjoy by being the type of creature that has conscious mental states. Derived intentionality is what one who has original intentionality might ascribe to another system, given a plausible interpretation of its behavior or actions as having intentionality. As an example, a sophisticated computer program, such as AlphaGo Zero (Silver et al., 2017), appears to display intentionality in that is successfully and masterfully executes the task of playing the game Go, but it would be an error to ascribe original intentionality to that system. Whatever intentionality it displays is merely derived from those that developed the program in the first place, i.e. the engineers and programmers that *consciously* built the system to play the game of Go masterfully.

There are two broad positions on might a take regarding intentionality, I will briefly define the views before moving on to discussing phenomenal intentionality and its connection to IIT view of qualia. The views are defined as follows:

> **Intentionalism**: all of one's mental states or experiences are intentional, meaning they are necessarily about some object.
> **Non-intentionalism**: not all of one's mental states or experiences are intentional, meaning they are not necessarily about some object.

To be up front, I think this is a matter that is largely open to interpretation for what IIT is actually committed to, I will be giving the interpretation which I think is most natural for IIT according to its canonical version. That being said, I think where you land on these particular distinctions is largely

concerned with matters that go beyond the scope of merely IIT itself, and lands one in the territory of what they think the nature of information is more generally as it concerns the mind. That being said, I will show what implications the most natural interpretation has for the theory's view of qualia and more generally what IIT says about our mental lives.

There is a thought experiment that I want to rely on to motivate my take on IIT's view of qualia and intentionality, and that's Putnam's (1981) brain in a vat thought experiment. His thought experiment is the following:

> "[I]magine a human being (you can imagine this to be yourself) has been subjected to an operation by an evil scientist. The person's brain (your brain) has been removed from the body and placed in a vat of nutrients which keeps the brain alive. The nerve endings have been connected to a super-scientific computer which causes the person whose brain it is to have the illusion that everything is perfectly normal. There seem to be people, objects, the sky, etc; but really all the person (you) is experiencing is the result of electronic impulses traveling from the computer to the nerve endings. The computer is so clever that if the person tries to raise his hand, the feedback from the computer will cause him to 'see' and 'feel' the hand being raised. Moreover, by varying the program, the evil scientist can cause the victim to 'experience' (or hallucinate) any situation or environment that evil scientist wishes. He can also obliterate the memory of the brain operation, so that the victim will seem to himself to have always been in this environment. It can even seem to the victim that he is sitting and reading these very words about the amusing but quite absurd supposition that these is an evil scientist who removes people's brains from their bodies and places them in a vat of nutrients which keep the brains alive. The nerve endings are supposed to be connected to a super-scientific computer which causes the person whose brain it is to have the illusion that…" (Putnam, 1981, pp. 5–6)

Putnam was concerned with a number of different issues in his *Brain in a Vat* thought experiment, namely with what our connection is to the external world, what is the nature of intentionality and representation, how reference is fixed, etc. Since our focus is on the intentional nature of mental states, specifically phenomenal experience, I'll be leaving the issues of reference and external world skepticism aside for the moment, to focus on intentionality and representation.

Now let us look at the following implication of IIT in light of what we've discussed thus far, consider the following modification to Putnam's original thought experiment: imagine that Giulio Tononi has decided he wants to test Putnam's brain in a vat thought experiment and grabs one of his unsuspecting colleagues and places them under anesthesia (I don't think this is something Tononi would actually do, but one can never be too cautious when theory is on the line…). Let's imagine Tononi takes his colleague that's unconscious places them in the Metropolitan Museum of Art where

one will find Pollock's painting from figure 7 hanging. He then wakes them up for but an instant, allows the colleague's MICS to take on the structure of that experience, and then immediately 'freezes' the system in that state (let's just imagine such a 'freezing' is possible for the sake of argument). According to IIT, that system would be having a phenomenal experience of the kind associated with viewing Jackson Pollock's Autumn Rhythm (Number 30), all the distinctions, relations, and Q-folds typically associated with such a stimulus, the Q-space[49] would be such that Tononi's colleague would have the experience of such a scene. Now, let us suppose that Tononi surgically removes the main complex from the colleague's skull, places it in a vat, just like in the original thought experiment, only this time the vat is not hooked up to a computer that simulates the environment (again for the sake of argument imagine such a fanciful scenario is possible). Rather, the system merely maintains that Q-space suspended in a nutrient-rich goo. The colleague is now a brain in a vat, or more accurately according to IIT as regards phenomenal experience, a maximally irreducible cause-effect structure in a vat. The only difference is, this vat is not connected to the external world in any relevant sense, it has been designed to merely maintain the current cause-effect structure that the colleague was in while viewing Pollock's painting. So long as that exact cause-effect structure is maintained, they will perpetually be having an experience of Autumn Rhythm Number 30, exactly of the kind they had before being transplanted. Now, what is such a phenomenal experience *about*?

It's not about some object in the world, as the actual stimulus of Pollock's painting is to be found in the Metropolitan Museum of Art out in the non-vat world. Currently they are only experiencing the painting as a fixed MICS in their brain, which is comfortably sitting in a vat of goo detached from the external world. According to IIT this is perfectly coherent, and more so a direct consequence of the view. So, what is their mental state about? It's about some intentional object, i.e. the set of distinctions, relations, and Q-folds currently transfixed in the system!

I think it should be clearer after the explanation of how qualia according IIT is generated that not only is the theory committed to intentionalism, namely, that all mental states are about something (i.e. the cause-effect structure instantiated at some given time), but that furthermore the nature of such intentional states, when they are conscious, are phenomenal. That is, those intentional states are about something only in virtue of the fact that they have a phenomenology associated with them. Given that the only states that exist intrinsically according to IIT are those that have integrated information, and integrated information is identical to phenomenal experience, those mental states that are conscious

[49] A Q-space is an abstract space in which all the possible configurations of a cause-effect structure might be found. Each quale would have a unique shape in such a Q-space, specifying that individual qualitative state.

are thus *about* something. Notice that this view is stronger than intentionalism as defined earlier in this section, since intentionalists may think that qualitative states are not intentional; this means those that defend IIT fall into the camp that argues that there is a class of intentional states which are constitutively phenomenal. To understand this, we must look at what it would mean for phenomenology to play a role in constituting the intentional nature of our conscious mental states. Such a view is called a phenomenal intentionality view.

> **Phenomenal intentionality (PI)**: There is a kind of intentionality, pervasive in human mental life, that is constitutively determined by phenomenology alone. (Horgan and Tienson, 2002, p. 520)

That is, according to those that hold PI, there is a flavor of intentionality that comes about as a result of phenomenology and not say by a result of solely our beliefs or thoughts about some intentional object. Take for example the experience of a red rose (to borrow from Horgan & Tienson), one has a number of qualitative features associated with this from a number of sense modalities, the sweet smell, the red color, the thorny stem, etc., those phenomenal experiences are intentional in that they are about something, namely the *what-it-is-like* of that experience. Now this is not the same claim that every and all intentional states are phenomenal, only that a certain class of intentional states are constitutively phenomenal; namely those intentional states with an associated phenomenal state that are accompanied by them, as one cannot detach their phenomenal content from their intentional content. We'll discuss this further after we look at Mendelovici's (2018) view, and compare the two to see which is more appropriate to ascribe to IIT given our discussion of how qualia arises.

The second view I wish to discuss is that of Angela Mendelovici (2018) in which she argues that all intentional states are phenomenal intentional states, or strong PIT as she calls it. Strong PIT is the view that "[a]ll (actual) intentional states arise from phenomenal consciousness"(Mendelovici, 2018, p. 86). Mendelovici arrives at this position by rejecting that there is any such thing as derived intentionality, something that is intentional merely by being interpreted as such by a creature with original intentionality. As Mendelovici says:

> "Suppose you are in a phenomenal state with a blue-ish phenomenal character. It might seem that simply by being in this state, you automatically have blueness before

> your mind; you're automatically representing <blue>. Nothing else need be added to your state in order for you to represent <blue>. In the same way, if you have a phenomenal state with a blue-square-ish phenomenal character, you automatically represent <blue square>… The same goes for other kinds of perceptual states. It seems, then, that having phenomenal states suffices for having intentional states and so that phenomenal consciousness is the right kind of ingredient for giving rise to intentionality." (Mendelovici, 2018, p. 91)

According to IIT, having the phenomenal experience of a <blue square> just is the representation of said blue square. Since, according to the theory, the only way to be having a phenomenal experience of <blue square> is to have a cause-effect structure of the form that specifies such an object. The intentional object is identical to the integrated information structure present in the system, it's what one's phenomenal state is about. Both of these views, the PI view of Horgan and Tienson in which only **a certain class** of intentional states are phenomenally constituted and Mendelovici's strong PIT in which **all** intentional states are phenomenally constituted are compatible with such a claim from IIT. The question we are left with is which view should IIT adopt?

Since any conscious state according to IIT is identical to some cause-effect structure any *conscious* mental state (i.e. intentional states) is identical to the same cause-effect structure since, as you will recall, according to the exclusion postulate there can only be one main complex of integrated information present to one's conscious experience. For the same reason, originally representing a content (say the <blue square>) is identical to the phenomenal property associated with that content, that is, the conjunction of the **distinctions**, **relations**, and **Q-folds**. Because the content present to us when we undergo a conscious mental state is the totality of the cause-effect structures in the brain at some given time any contents of originally intentional states will be identical to those integrated information structures.

This, in itself, doesn't commit IIT to the claim that any and all intentional states are constituted by phenomenology. IIT is a theory of consciousness and thus those mental states that fall under the purview of our phenomenal experience are captured in its scope. There is no need for IIT to overcommit itself to a much stronger claim unnecessarily, in going so far as to claim that any and all intentional states are constituted phenomenologically. IIT need only claim that there are a certain class of intentional states, those mental states that are conscious, that are constituted by phenomenology.

What IIT has to say, if anything, about our unconscious mental states the view should and does keep silent on.

Conclusion:

It has been the aim of this essay to give a philosophical treatment about IIT's implications regarding the marks of the mental – phenomenal experience (qualia) and intentionality. I began this essay by giving a short explication of the essential features of IIT to tackle questions concerned with qualia and intentionality. I then moved on to give an overview of IIT's story concerning how qualia arises as the result of a specific information-theoretic processing – integrated information. I then used this to shed light on our second target, intentionality, and tried to motivate that IIT holds and should explicitly adopt a phenomenal intentionality view of how intentionality is constituted. The position I think IIT is most similar to is Horgan & Tienson's (2002) phenomenal intentionality position. My overall hope is that such a treatment opens pathways of discussion between those involved in developing IIT from the natural sciences and those in philosophy interested in topics of mutual interest.

"Intuitively, it is more reasonable to suppose that the basic entities that all this causation relates have some internal nature of their own, some intrinsic properties, so that the world has some substance to it. But physics can at best fix reference to those properties by virtue of their extrinsic relations; it tells us nothing directly about what those properties might be."

- David J. Chalmers (1996, pg. 153)
"The Conscious Mind"

Karma Police
Arrest this man
He talks in maths
He buzzes like a fridge
He's like a detuned radio

-Radiohead (1997)
"Karma Police," OK Computer

I'm a reasonable man
Get off, get off, get off my case
I'm a reasonable man
Get off my case
Get off my case

After years of waiting
After years of waiting nothing came
And you realize you're looking
Looking in the wrong place.

-Radiohead (2001)
"Packed Like Sardines in a Crushed Tin Can," Amnesiac

The austere language of which IIT is explicated may make it seem like one is "talking in maths" and how can maths ever capture the qualitative aspect of experience, the lived and felt quality of our phenomenology? I'm sympathetic to this, but maths is just another language, one we can use to capture certain features when our mother tongues fail us. I can use English to try and capture the felt quality of experience but that too will always feel deficient in some sense, just as the austere maths of IIT might feel deficit. However, one shouldn't let this superficial outcome of the deficiencies of language stand in the way of trying to give an explanation. I think the structure and dynamics argument we encountered in the first essay falls victim to this trap, and so in the following essay I want to motivate that perhaps we've been looking for consciousness "in the wrong place" as a result of the hard problem. My hope is to vindicate structural and dynamical explanations of a certain variety to help overcome the worries about consciousness and its place in nature.

4. Not all Structure & Dynamics are Equal

Introduction:

The Structure and Dynamics (S&D) argument from Chalmers (2003) is taken as a backbone of the hard problem of consciousness (Chalmers, 1996, 1995). The hard problem of consciousness is the problem of why there is any experience associated with certain physical processes. Here we are taking physical processes to be those describable by the tools of the physical sciences, which Chalmers characterizes as structural and dynamical features. Structure here is meant to refer to the spatial and formal features of some phenomenon and dynamics as the temporal and nomic features of some given phenomenon. This for the most part has been uncontroversially accepted as a plausible description of what consists of the "physical" (c.f (Alter, 2016) for a thorough treatment of the structure and dynamics argument). Largely this is due to Russell's description of what physics is in the business of describing, namely the spatial/temporal and mathematico-causal description of the world (Russell, 1927, p. 390). The disagreement comes in the gap it purportedly shows between the physical and the mental. These gaps are centered around a set of closely related arguments aimed at showing that physical facts do not entail phenomenal facts, and thus there is an epistemic gap between the physical and phenomenal (Chalmers, 2003; Jackson, 1982; Kirk, 1974; Levine, 1983). Some then go on to argue that this epistemic gap entails a metaphysical/ontological gap, and if this is so, then physicalism must be false. Herein I am not so much concerned with showing that physicalism as a metaphysical thesis is false or argue that anti-physicalism is preferred; rather I want to examine the structure and dynamics argument itself and show that it may not offer a nuanced enough description of structure and dynamics. This is not to show that physicalism is correct, but rather that the structure and dynamics as Chalmers has described is not the whole story, I'll let others take it as they will what the metaphysical implications are of such a result.

My considerations of the structure and dynamics argument come from the perspective of Integrated Information Theory (IIT) of consciousness (Oizumi et al., 2014; Tononi, 2012, 2008; Tononi and Koch, 2015), which claims that consciousness is integrated information in a system. Recently I have argued (Mindt, 2017) that IIT is susceptible to the structure and dynamics argument, and thus does not solve the hard problem, aside from falling victim to other concerns regarding physicalism. In this essay I will be turning my sights on myself, so to speak, and will be showing why the structure and dynamics argument doesn't apply to IIT, if we develop a more nuanced

understanding of structure and dynamics. This current essay will hopefully better bring into focus what the work in Mindt (2017) was aimed at which was, to present as a conditional, if we accept the hard problem of consciousness and the S&D argument on its terms, what are the implications for IIT? I did this as an argumentative tool to then in the present work show why the S&D argument is incorrect, if one has a certain understanding of structure and dynamics, and that ultimately IIT may be in a position to account for the hard problem as a result. The problem may not be with structure and dynamics per se, but rather with how we conceive of the physical in terms of structure and dynamics. My hope is that if one can give a coherent information-theoretic metaphysics one might make a distinction between external structure & dynamics and internal structure and dynamics. The remainder of this essay will be an attempt to make clear what I mean by that distinction.

I will begin in section one by giving a brief overview of physicalism and the structure and dynamics argument. In section two I will explore recent work done in the complexity sciences on giving a naturalized definition of semantic information, one that is substrate independent[50]. Section three will look at IIT and its relationship to a more nuanced understanding of structure and dynamics as well as its position in approaching the hard problem of consciousness. Section four will summarize what the preceding discussions in §1-3 have revealed and argue that Chalmers (2003) structure and dynamics argument fails once we see that not all structure and dynamics are equal. Section five will examine how the arguments in the essay relate to the explanatory gap argument as a litmus test for how a more nuanced understanding of structure and dynamics affects arguments traditionally presented against structural and dynamical explanations (conceived as physical explanations) of phenomenal experience.

Section 4.1: What is Physicalism? What is Structure & Dynamics?

Physicalism[51], although often considered the dominant metaphysical position, is a notoriously tricky beast to tackle in terms of what is its actual definition. I will however begin this section with a

[50] By substrate independent, I just mean, that the nature of that information is not reliant on it having any particular kind of realization base, whether that be the mushy and pulsing throb of organic matter, the silica of a computer chip, or any other kind of substrate base one might think of.

[51] I will be using 'physicalism' and 'materialism' interchangeably in this essay, there may be a difference historically in these two views though they are often used interchangeably in the literature and I will be adopting this perhaps lazy habit as a way to ease exposition.

short discussion on what we mean when we say some thing or process is *physical* and what we mean when we ascribe to the metaphysical position called *physicalism*.

So, what is physical? And what is physicalism? To begin let's look at what it means for some thing or process to be *physical*. Stoljar (2017) offers an apt definition of what it means for something to be physical (and is the same conception which Chalmers has in mind in his arguments as well). Stoljar calls this the *theory-based conception of physical* and defines it as:

> **The theory-based conception (TBC for short)**: "A property is physical *iff* it either is the sort of property that physical theory tells us about or else is a property which metaphysically (or logically) supervenes on the sort of property that physical theory tells us about." (Stoljar, 2017, sec. 11)

So, we can consider that which is physical to be anything that falls under the purview of those objects/properties/processes posited as part of some physical theory. For example, mass is posited as a property of objects according to physics and would thus qualify as a physical property according to the TBC. We can now move to defining what the metaphysical view of physicalism would thus amount to. Lewis (1983) offers a definition of physicalism (after much back and forth on possible definitions).

> **Physicalism/Materialism**: "Among worlds where no natural properties alien to our world are instantiated, no two differ without differing physically; any two such worlds that are exactly alike physically are duplicates." (Lewis, 1983, p. 364)

Another way to put this is there's nothing over-and-above the physical, there is no difference without physical difference, etc (sorry to ghosts, ectoplasm, the ethereal plane in general, but there just ain't no room for you in physicalism!). That being said, Chalmers (1995 & 1996) worries were that it doesn't seem to be the case that phenomenal experience fits into such a world view, since it's not clear how the felt aspect of experience, or the what-it's-likeness of experience, is entailed by such physical properties. Another way to put it, it's not clear that the existence of certain physical properties entails the existence of identical phenomenal properties. This is sometimes framed in terms of the conceivability argument (Chalmers, 2003; Kirk, 1974), the knowledge argument (Jackson, 1982), and the explanatory gap argument (Levine, 1983).

That being said, there are some interesting processes we might speak of in the context of the physical, namely certain structural and dynamical processes/properties of systems. Chalmers defines structural and dynamical properties as:

> "First: a microphysical description of the world specifies a distribution of particles, fields, and waves in space and time. These basic systems are characterized by their spatiotemporal properties, and properties such as mass, charge, and quantum wavefunction state. These latter properties are ultimately defined in terms of spaces of states that have a certain abstract structure (e.g., the space of continuously varying real quantities, or of Hilbert space states), such that the states play a certain causal role with respect to other states. We can subsume spatiotemporal descriptions and descriptions in terms of properties in these formal spaces under the rubric of structural descriptions. The state of these systems can change over time in accord with dynamic principles defined over the relevant properties. This result is a description of the world in terms of its underlying spatiotemporal and formal structure, and dynamic evolution over this structure." (Chalmers, 2003, p. 258)

Structure is the spatial-temporal relationships between entities/phenomenon. As an example, take my current spatiotemporal location at the moment of the writing of this sentence, Budapest, Hungary in the year 2019. We can provide further specification in further and further detail, the philosophy PhD lab in Central European University on March 19th, 2019 at 2:44 pm. Or down to the excruciating detail of all the components which compose my body, down to the most abstract and inconsequential details of my atomic structure. Mapping any of these features would be a structural description (of a certain variety) of myself. Aside from this, I have interesting dynamical features, not merely the structural features describing the state of all my atoms in my body. I not only occupy a unique position in spacetime, but also carry a number of causal capacities and my states evolve through time dynamically, governed by certain dynamic principles. For instance, my location in spacetime is accompanied by my ability to retrieve a beer from the fridge and to type an essay on my laptop (I have limited dynamical capacity for much else), and these changes in states evolve over time. There is thus, a kind of symphony of structural and dynamical features of myself that describe my body and its interaction with the world throughout my entire existence. Furthermore, these features of my body (the system in question) is amenable to third-person observation. Someone could put me in some futuristic machine that details every component of my body in excruciating detail and give a full description of these structural and dynamical features. One could call all those details and explanation of their processes the *physical*, though I think a more apt description would be the *extrinsic features* of a system, allowing one to both stay neutral on the battle of physical vs. non-physical, and to rather get at the

more interesting questions concerning processes and interaction (see section 3 for a more detailed explanation of this).

The context in which this essay and the S&D argument originally was targeted warrants comment on what I as well as Chalmers mean when we speak of consciousness or experience. By consciousness in this essay I will be adopting the phenomenal sense of the word, as in a system can be said to possess consciousness or undergo phenomenal experience when there is *something it is like* to be that system, some kind of felt experience (Nagel, 1974). The question is, can S&D of the kind Chalmers describes explain the presence of phenomenal experience; that is do such explanations generate a plausible account of why and how certain system have phenomenal experience rather than others? Chalmers concludes they do not, since essentially they only give one relational and dispositional descriptions of the world (as Stoljar (2006) characterizes Chalmers' S&D argument) and as he argues in Chalmers (1996) this is insufficient to capture phenomenology and thus fails to account for the Hard Problem.

There are actually two questions of relevance in this discussion, (1) can S&D of the kind Chalmers describes capture phenomenal experience and explain consciousness? And, much more importantly for the present work, (2) is there only the kind of extrinsic S&D which Chalmers describes in his argument? The focus of this essay is on the second question, note that the veracity of an answer to the first question only concerns S&D as Chalmers describes it (what I'm calling *extrinsic S&D*) and does not apply to anything outside his description. The next section (§2) will be focused on motivating that Chalmers S&D argument paints S&D features of a system with too broad a brush and misses important nuances in certain structural and dynamical features of a system.

Section 4.2: Complexity Sciences, Meaning, and Intrinsic Structure & Dynamics

Recently, Luis Favela (2019) has argued that IIT is best understood under the rubric of complexity sciences more generally, as he phrases it "I claim that IIT is a complexity science approach to consciousness. Even though there is no single 'complexity science', there are typical concepts, methods, and theories that fall under that heading" (Favela, 2019, p. 23). Favela argues that IIT and complexity sciences share an important feature which makes them natural bed-fellows, namely, a focus on *interaction dominance* rather than *component dominance*. Favela takes this notion from Bechtel & Richardson (2010) whose aim was to understand the extent to which mechanistic explanation fit the models with which scientists were concerned, and in what ways they didn't. Mechanistic explanations

are appropriate and in fact the best way to go in explaining some biological phenomenon when that phenomenon entails systems that are component dominant, meaning they are about the individual parts and mechanisms that compose a system and not about the interactions between those parts and/or mechanisms. Interaction dominant systems are concerned with the opposite, when the behavior of a system is not dependent on the parts or mechanisms themselves but their interactions with one another. A common example that illustrates this, used by Favela, is the difference between a toilet and a locust swarm. A toilet functions as a result of the organization of its parts, and this function is decomposable to those components. An explanation of the behavior of such a system is thus analyzable mechanistically. As for the locust swarm the individual components themselves (the individual locusts) are not responsible for the swarm's behavior or movements, but rather it is the interaction between the individual components. A way to think of this is that in the case of the swarm, the global-dynamics of the system are what's important to the system rather than the individual components dynamics, and with the toilet the individual components dynamics are what is important and not the global dynamics of the system.

I think understanding the project of IIT as being concerned with interaction dominance over component dominance is a useful way of characterizing what the theory describes. It also helps to avoid the issue of conflating the notion of 'mechanism' as it's used in IIT with that of the concept of mechanism (Machamer et al., 2000) as it's used in the philosophy of science (although there may be interesting and relevant similarities). Since as we saw in the previous essay (chapter 3) on qualia the individual elements (i.e. the individual neurons) are not what is important in understanding how the qualitative aspect of experience is generated; rather it is the interaction between all these components, how the distinctions, relations, and Q-folds form that are important for understanding how phenomenal experience is generated through a collection of neurons. We are not interested in the locusts (i.e. the individual neurons) in IIT, we want to know about the swarm (the MICS)!

In a recent essay Kolchinksy & Wolpert (2018) explore how to give a naturalized definition and explanation of semantic information drawing from work in non-equilibrium statistical physics and experiments with autonomous agency, all under the framework of complex systems theory. The notion of semantic information they are working with is as follows:

> **Semantic information**: the information that a physical system has about its environment that is causally necessary for the system to maintain its own existence over time. (Kolchinsky and Wolpert, 2018, p. 1)

The question is how does semantic information naturally arise as the result of certain interactions? We want to give a naturalistic explanation about how semantic information can result from what appears to be non-meaningful physical interactions, and thus one must produce a story about how that occurs[52]. This goes back to a debate originally raised by Bar-Hillel & Carnap (1953) where they raise the issue that Shannon's information measure doesn't answer the question we would expect a complete account of information to answer, namely how semantics arises from the syntax of communication channels. Although there has been much discussion over the years (c.f. (Adriaans, 2010; Bar-Hillel and Carnap, 1953; Dretske, 1981; Floridi, 2011, 2005)) on how to give a semantic notion of information, many are either too epistemic (such as floridi's GDI[53]) or restricted to biological organisms. We can phrase the desisderatum as a semantic notion of information which is substrate-independent, so to speak, that is to say does not rely on the system in question being biological or having epistemic states of the kind that humans have.

Having such a semantic notion of information opens up the possibility that not only biological organisms can have semantic information, but also machines, or something totally alien; since we understand meaning as any information which helps a system to maintain its existence over time. For example take R2D2, everyone's favorite beeping robot from Star Wars, there will clearly be information relevant to R2D2 maintaining its existence over time (its having the information where the repair and charging stations are) and that will be different in value from that merely syntactic information R2D2 gains from its environment (say it's having the information that the sky is blue, depending on what planet he is on). It is meaningful in the sense of maintaining R2D2's existence that it has information about its environment concerning where repair stations or charging stations are, as without this relevant information R2D2 would beep no more. That's not to say other information

[52] Someone may object here that the way in which I am using semantic information deviates from the norm we would usually expect. That person would be correct to point this out, and that's precisely the point of why I want to introduce this somewhat non-standard notion of semantics. We've seemed to make little progress in bridging the gap we're concerned with in our understanding of nature by investigating semantics propositionally (such as Bar-Hillel & Carnap (1953)) or epistemically (such as Dretske (1981)). I propose philosophy look towards understanding semantics as the relationship between a system and its environment, looking for those properties that causally maintain its existence over time. While keeping an eye on avoiding anthropomorphizing such a project, that is erroneously putting top-down constraints on the nature of semantic information having to be propositional or epistemic.

[53] For a criticism of Floridi's semantic notion of information c.f. Adriaans (2010) criticism of an overly epistemic notion of semantic information. Adriaans does a good job of highlighting the two general spheres one might occupy in coming up with a semantic notion of information. On the one hand, one might see information theory as a "handmaiden" to more classical epistemology. Adriaans however, and I tend to agree with him, sees information theory as a competitor to classic epistemology. I'm less certain on how developed a position the latter is, than Adriaans himself appears to be, but I nonetheless see the project of naturalizing consciousness and information as going in the direction Adriaans argues.

might not have value (more on this in what follows), but it will not be meaningful in the sense that Kolchinsky & Wolpert are concerned with.

Kolchinsky & Wolpert argue that one can give an explanation of how semantics comes from syntax by understanding the relationship between a system and an environment and considering the viability function; between its ability to maintain a state of non-equilibrium and keeping a high-state of entropy (e.g. for biological organisms, not dying!). Any syntactic information which does not figure into maintaining this state of non-equilibrium is 'meaningless' in this sense, and thus is not semantic information. Information however that does contribute to maintaining this state of non-equilibrium is meaningful, and thus is semantic information. As Kolchinsky & Wolpert explain:

> "The **semantic content** of a particular system state x is defined as the conditional distribution (under the optimal intervention) of the environment's states, given that the system is in state x. The semantic content of x reflects the correlations which are relevant to maintaining the existence of the system, once all other 'meaningless' correlations are scrambled away." (Kolchinsky and Wolpert, 2018, p. 4)

They are saying that to discover those states of affairs that have meaning one needs to perturb the initial states of system, by way of making interventions on well-defined variables[54] and seeing what information, if any, is lost. If such an intervention affects the optimal viability function then that contains meaningful information, since its information that is valuable to the system for maintaining its own existence. If an intervention is made and no meaningful information is lost, then all that was scrambled away was the meaningless syntactical correlations.

I think motivating this with one of the examples that Kolchinsky & Wolpert use will be helpful at this point, consider the following:

> "Consider a distribution over food-catching birds (the system) in the forest (the environment), over a timescale of $\tau = 1$ year. Assume that at $t = 0$ the birds have cached their food and stored the location of the caches in some type of neural memory. If we 'scramble the information' by placing birds in random environments, they will not be able to locate their food and be more likely to die, thus decreasing their viability. Thus,

[54] See chapter 2 of this dissertation for an explanation of interventionist causation and what this means in the context of IIT.

> a food-catching bird exhibits a high value of information." (Kolchinsky and Wolpert, 2018, p. 3)

There is an intuitive level where this makes sense, it's more valuable to a system, or as in the case above an organism, to have information relevant to its continued existence (i.e. knowing the location of the food cache then say knowing that the sky is blue). Each carry with them a unique set of informational relationships, but one merely indicates a state of affairs (the sky being blue) while the other indicates a valuable state of affairs relevant to the systems continued existence (the cache of food).

The question boils down to what syntactic information is relevant for maintaining a system's existence. This has to do with discovering the value of the information and this amounts to determining the difference between the viability of the intervened distribution and the actual distribution. What comes out, according to Kolchinsky & Wolpert, is the syntactic information that contributes to maintaining the systems existence. As they put it "A positive difference means that at least some of the syntactic information between the system and environment plays a causal role in maintaining the system's existence" (Kolchinsky and Wolpert, 2018, p. 3).

There is something important to note, and what is essential in understanding how this discussion relates to the issues in IIT about how meaning is generated and furthermore how this relates to the structure & dynamics argument more generally. I've merely wanted to highlight in this section that it is possible to generate a notion of meaning – a semantic notion of information from purely structural and dynamical properties. But mind you, this case has been understanding meaning from an extrinsic point of view, that is, some observer intervening on a system (i.e. the bird or R2D2) in the context of its environment (i.e. the location of food caches or repair/charging stations). But how is the meaning generated by a system *for itself*? This is a slightly different question because now we are asking how is the information presented to the system independent of some observer intervening on the system and environment to determine the difference between the intervened distribution and the actual distribution, how does a system do this on its own? Another way of putting this, how does the bird or R2D2 interface with the relevant meaning that is generated by means of that system maintaining its existence in an environment?

Section 4.3: IIT, Intrinsic Structure and Dynamics, and Approaching the Hard Problem

The method discussed in the previous section is strikingly similar to that method by which IIT determines the optimal intervention for quantifying the level of φ^{Max} of a mechanism. Recall from chapter 2 and 3 that IIT looks to find those informational relationships that form irreducible sets, that is elements that form an integrated whole, where the loss of some element affects the amount of information the group of elements taken together express (integrated information). The way of determining this is placing a set of elements in a state and perturbing them by performing 'cuts' (or directed interventions on specific variables, by say introducing noise into one of the elements, etc.) and seeing how the whole is affected by the intervened variable. If the amount of information that the elements express is less once the target variable has been intervened on, then that set of elements forms a collection of elements (mechanism) that has integrated information. As it's put in IIT 3.0, "[e]ach concept of a mechanism in a state is thus endowed with a maximally irreducible cause-effect repertoire (MICE), which specifies what the concept is about (its quale "*sensu stricto*"), and its particular φ^{Max} value, which quantifies its amount of integration of irreducibility" (Oizumi et al., 2014). That is to say, that each mechanism that specifies a concept/distinction has a meaning identified by only that mechanism, and recall that it has this because of its unique cause-effect repertoire that it specifies in the system as a whole.

The difference here between the approach from Kolchinsky & Wolpert and that of IIT, is IIT is concerned with not only the meaning that can be extrinsically interpreted from the outside (i.e. through interventions on a target system), but rather what the meaning is that is intrinsically brought to the fore to the system; how do these unique informational relationships constrain the past and future states of a system as it evolves dynamically overtime? There is however an important connection between these two methodologies for providing a semantic notion of information by discriminating between purely syntactic information and information that causally constrains the system to facilitate its continued existence. IIT is just asking the further question, which is, what makes a system that has internal **and** intrinsic meaning different from a system that has purely extrinsic meaning? According to IIT, it's when a system has the right kind of integration of the kind described in chapter 3. The following section will delve more deeply into this distinction and give an argument as to why not all structure and dynamics are equal.

Section 4.4: Not all Structure and Dynamics are Equal

We've seen thus far that there are a number of different kinds of structure and dynamics. There are those extrinsic structural and dynamical properties which merely indicate a system's syntactical features (using the language of information theory) those features which Chalmers calls structural and dynamical in his characterization of what physical explanation tells us. However, as Kolchinsky & Wolpert show not all syntactical information is "meaningless" so to speak, and thus we can distinguish information that is meaningful to a system from that which is merely a byproduct of the connection between a system and its environment. In terms of meaning we can consider all other information not directly related to a system's continued existence as 'noise' at this level of abstraction. We already see here that there is a difference between intrinsic S&D and extrinsic S&D. But there is a further level at which not all intrinsic structural and dynamical features of a system are the same, depending on what perspective one takes with regard to the target system.

I now want to make a further distinction, between External S&D and Internal S&D.

> **External S&D**: those meaningful or meaningless S&D properties of a system, observed from an external perspective, that are necessary for that system to causally maintain its own existence as an observable entity with certain behaviors.

> **Internal S&D**: those meaningful S&D properties of a system that are necessary for that system to causally maintain its own intrinsic existence; that is maintain an existence which has an internal perspective on its own causal processes (i.e. its dynamical evolution over time).

Some may think this a rather odd distinction to make but in what follows I will endeavor to explain why such a distinction is not only useful, but necessary, for understanding that not all structural and dynamical properties of a system are equal and why IIT's account is so informative on the matter. It would be useful to quickly recap the various kinds of structural and dynamical properties that have so far been discussed. Initially (§1) we discussed structure and dynamics as Chalmers conceives of it as a facet of his hard problem of consciousness, this type of S&D is what I'm calling extrinsic S&D (a non-meaningful variant of external S&D). Next (§2) we discussed some recent work in complexity science in developing a semantic notion of information which I argued reveals that there are not only

structural and dynamical features of the kind Chalmers describes, but also intrinsic S&D (as described by Kolchinsky & Wolpert (2018)) of the kind we get when we look at the interaction of elements in the system, rather than the elements themselves (this would be the meaningful variant of S&D, intrinsic but also interpreted from an external perspective). What about the situation in which we are not only concerned with whether a system can be validly *interpreted* to contain meaning, but there's *something it is like for* that system to have such meaning, as is the case in phenomenal experience? This is why I think it's important to understand that not only are extrinsic and intrinsic structure and dynamics not equal, there is a further level of description contained within intrinsic structure and dynamics. That is there is intrinsic structure and dynamics one might characterize about a feature of reality as it relates to externally capturing what processes maintain a systems existence over time (what IIT calls extrinsic existence) those that are external to a system (i.e. an observer intervening on some variable and determining the semantic value of some set of data from comparing the distributions). Then there are those intrinsic structural and dynamical features of a system which relate to capturing what processes maintain a system's existence over time *internally* to said system (i.e. my consciousness of the semantic features of myself taken as the system in question, what it is like for me to experience the world).

External and internal S&D are merely the different perspectives one can take on a system or as a system on its processing (or more accurately has the ability to take), that is, I can be observed as having certain extrinsic S&D features, say the molecules that compose my body or my motion through a room being tracked by cognitive scientists for some experiment or perhaps the subject of Tononi's brain-in-a-vat experiment from the previous chapter (heaven forbid!). These would amount to being treated as an object of inquiry, in the third-person sense, from an external perspective. Some systems examined from such an external perspective may exhibit certain weakly intrinsic S&D of the kind discussed in the last section, revealing that there is meaning associated with such processes, such as the food-catching bird or the beeping R2D2. Others, more commonly, will exhibit merely the kind of extrinsic S&D features which Chalmers rightly points out will fail to account for the qualitative aspects of our brain's physiological processes. Take for example the behavior of a gas in a vacuum, there is no "meaning" to be found in that system reaching a state of equilibrium, there is merely the indication of various syntactical states of affairs, there is no relevant sense in which those processes are differentiable enough to be relevant for that system to maintain its existence in the sense we're concerned with, from the intrinsic perspective. As concerns the latter extrinsic features, Chalmers is most assuredly correct to think such an explanation would never account for the felt aspect of

experience, the what-it-is-likeness, but if our present discussion has been successful, this should be wholly unsurprising and uninteresting. These are both external S&D of a certain variety, the former more interesting for questions about the mind, as they reveal a system that may have semantic properties. The former being more relevant for understanding the causal structure of the world, one which is revealed to us by the typical tools of scientific inquiry. What is interesting for the question of how phenomenal experience arises as the result of certain structural and dynamical properties is not whether I can be observed to have semantic properties, but that the system in question (i.e. myself) can track those meaningful properties – whether there is something it is like for me to undergo the causal processes necessary to maintain my own existence over time.

I am the type of system (and if the reader is similar enough to myself, will also be the type of system) which can observe internally those state transitions, tracking the structure & dynamical changes as they occur within myself. Assuming that the only kind of structure and dynamics are extrinsic S&D is a sure-fire way to miss out on the question worth asking, which is, if we find ourselves in a world governed by a complicated array of structural and dynamical properties what are the ones which have an internal perspective on their own processes as they unfold? If we are looking for a natural explanation of phenomenal experience that should be what our concern is and shedding light on the nuance of what kinds of structural and dynamical properties there are is the right way to advance this effort.

I've argued in a previous essay (chapter one) that if we take the hard problem and the structure and dynamics argument at face value, and accept them conditionally, then IIT fails to overcome the hard problem because it provides a purely structural and dynamical notion (that is as Chalmers conceives of S&D) of information (as was argued in (Mindt, 2017)). What we've seen thus far is that *not all structure and dynamics are equal*, and that there are relevant differences in structural and dynamical features of certain systems that should and need to be investigated further. If Chalmers's structure and dynamics argument is that there couldn't possibly be a S&D explanation of phenomenal experience, given how he defines structure and dynamics, then his argument doesn't apply to structure and dynamics that fall outside the scope of his definition. I've shown that there is nuance in structure and dynamics, even in the external case as was shown in section three, between meaningful and meaningless syntactical information. The further task is to take those lessons and apply them to the question of those systems which can be said to have semantic information, do those same systems have an internal perspective, is there something it is like for those systems to experience the meaningful causal states that maintain that systems survival over time? And if so, what are the relevant

S&D differences between those systems to warrant such an internal perspective? Both of these questions are susceptible to natural investigation, given we have the correct conceptual framework and tools to perform such an investigation. The project should be to develop and discover those, not to preclude the possibility from the outset by a crude characterization of structure and dynamics. Merely showing that not all structure and dynamics are equal is enough to resist the claims of the structure and dynamics argument, and without that argument, there seems little reason to accept the hard problem as the insurmountable obstacle it appears to be.

Furthermore, I argued in Baxendale & Mindt ((2018), chapter two in this dissertation) that IIT is best understood and best off adopting a broadly maipulationist-interventionist account of causation to avoid the damaging circularity of inter-defining information and causation. This has also been recently supported by (Lombardi and López, 2018) though they disagree with the route which we take to argue for the conclusion, they do however agree with our conclusion. In the third chapter of this dissertation I then went on to give an explanation of qualia, that is how and what IIT describes as the qualitative features of a complex of integrated information – how the system gets the phenomenology it has in virtue of its integrated information. I've endeavored to show that not all structure and dynamics are equal, or at the very least, it appears there may be relevant differences in structural and dynamical properties of certain systems that are not captured by Chalmers' description of structure and dynamics. If there are S&D properties that are not captured by Chalmers description and definition of what is structural and dynamical then the structure and dynamics argument is only true for those properties that are captured by his description. Intrinsic (both external and internal) structural and dynamical properties fall outside the scope of Chalmers description and definition, and thus fall outside the scope of his structure and dynamics argument. If this is so, then perhaps there are structural and dynamical explanations of a certain variety that may account for phenomenal experience. I merely want to show in this essay that not all structure and dynamics are equal and if this is so, then Chalmers structure and dynamics argument may not exclude a certain variety of structural and dynamical explanation of consciousness. Rather, what Chalmers's argument would show is that one cannot give, what I've called in this essay, an *extrinsic structural and dynamical* explanation. That leaves open the possibility of exploring a specific class of S&D properties, *intrinsic structural and dynamical properties*, and how those might explain consciousness and phenomenal experience. Specifically, those properties of a system I am calling *internal structural and* dynamical properties. The following section (§5) will look at how IIT is attempting to give an explanation of *internal structural and*

dynamical properties and what implications that has for developing a naturalized account of phenomenal experience.

Section 4.5: Closing the Explanatory Gap

The hope thus far has been that I've at least given reason to think that not all structure and dynamics are equal, what then are the implications for those arguments against physicalism more generally that purport to show an epistemic gap between structural & dynamical features (treated as physical features) and the mental (e.g. phenomenal experience)? I do not wish to vindicate physicalism or argue in favor of anti-physicalism; I merely want to highlight that given our more nuanced picture of structural and dynamical features we now have access to we might be able to close the gap with such a picture of S&D. Whether that means one should be a physicalist or anti-physicalist I'll leave that to the judgement of the reader. In the next essay, I will argue the best metaphysical framework in which to place such an explanatory framework is neither physicalism or some form of anti-physicalism but rather a form of *Informational Realism* – the view that information is a mind-independent feature of reality and not merely an abstraction or consequence of more fundamental physical processes. The metaphysical picture I have in mind will be more systemically offered in the last essay of this dissertation (chapter 5).

I have endeavored in this present work to offer an explanation of how a theory might give a S&D explanation of certain variety about consciousness, which dissolves the S&D argument of Chalmers. What I would like to do in the remaining section is explore one of the gap arguments that is often given to show that physicalism must be false. I will offer modifications of this argument to show that the conclusion is only secured, that there is an epistemic gap between the "physical" and the "mental," when one *only* considers extrinsic S&D features of a system. I will further expand the arguments and show that this leaves open the possibility that an intrinsic S&D might offer an explanation of phenomenal experience. My hope is to vindicate the process of developing a naturalized explanation of consciousness from the hard problem and the structure and dynamics argument, and by extension, those gap arguments that have arisen as a result. I stay agnostic on whether this vindicates physicalism as a result, I merely wish to vindicate structural and dynamical explanation of a certain variety. The reader may well take the arguments in this essay and use them for the purposes of vindicating physicalism if they so desire but that is not my aim, nor one that I think should be of any

particular interest to people concerned with explaining how phenomenal experience arises as a natural and information-theoretic process.

Section 4.5.1: Explanatory Gap Closed

I want to propose a modification of the explanatory gap argument (below) to show why once we take a more nuanced understanding of structure and dynamics there's no reason to posit such a damning gap between the right kind of S&D and consciousness. Recall from the first chapter that the explanatory gap is concerned with whether our a priori knowledge of structural and dynamical facts can reveal an explanation of phenomenal experience (see chapter 1, §1.3.1). I want to now offer a modified version of the explanatory gap in light of the present discussion in this essay.

Modified Explanatory Gap Argument:

1) Physical accounts explain at most extrinsic S&D

2) Explaining extrinsic S&D does not suffice to explain consciousness

3) Conclusion 1: No physical account (conceived as extrinsic S&D) would suffice to explain consciousness

4) There is a difference between extrinsic S&D and intrinsic S&D

5) There is no explanation of meaning from extrinsic S&D

6) There is an explanation of meaning from intrinsic S&D (Kolchinsky & Wolpert, 2018)

7) IIT offers an internal and intrinsic measure of S&D – Integrated Information

8) Integrated Information is qualitatively and quantitatively different than extent notions of information (i.e. Shannon's enthropic notion, Kolmogorov complexity, mutual information, etc.)

9) Conclusion 2: It is in principle possible to give an internal and intrinsic S&D explanation of consciousness

For this argument to hold I merely need it to be possible one could give an internal and intrinsic S&D argument of consciousness. What the details of such an explanation would be is left for future work, I merely want to show that it is not immediately precluded by the explanatory gap argument as a possibility, if we have a more nuanced understanding of different varieties of structural and dynamical properties. The explanatory gap and other related gap arguments (conceivability argument (Chalmers,

2003; Kirk, 1974) and knowledge argument (Jackson, 1982)) rely on a limited conception of structural and dynamical properties to secure their argumentative force. With a more nuanced understanding, grounded in advances in information theory, complexity sciences and the science of consciousness, we can begin cracking open the box on grasping a natural explanation of phenomenal experience. This may result in a rather radical departure from what we take to be the tools of the natural sciences, in so far as it's a departure from the root of modern sciences (Galilei, 2008, pp. 184–85) as putting those secondary qualities we would normally associate with the mind and consciousness to one side and focus on the primary qualities.

I think it should be noted that there is one strong reason to think physicalism is insufficient as a metaphysical framework for consciousness. If physicalism as a thesis entails that only extrinsic structure and dynamics can explain consciousness, then I think we have strong reasons to think at this stage it's false. If however physicalism can expand and accommodate a broadened notion of structure and dynamics then I think it's just vacuous as a metaphysical picture (similar to worries presented by (Hempel, 1980)) and offers us little in terms of explanatory usefulness. It's for this reason that in the last chapter I focus on developing a more useful metaphysical framework to accommodate the work done in this dissertation, that is developing an information-theoretic ontology.

In Mindt (2017) I placed IIT in the context of Chalmers' Hard Problem and his conception of structure and dynamics and argued if we take those arguments as a conditional IIT doesn't solve the hard problem since it falls victim to the structure and dynamics argument. As a consequence, I argued that IIT falls victim to the explanatory gap argument as well given the same conception of structure and dynamics. In the present work I have argued that not all structure and dynamics are equal and there's nuanced to be found in S&D properties of systems. I've argued that Chalmers' conception (as well as Russell (1927)) of structure and dynamics paints with too broad a brush and only characterizes extrinsic structure and dynamical features of systems. Once we appreciate the intrinsic S&D features of a system then IIT no longer faces the worries presented in Mindt (2017).

Turning back to the explanatory gap, we shouldn't have been surprised that given how Chalmers and those that argue against physicalism characterizes all S&D facts as extrinsic meant it couldn't possibly account for the internal perspective of a system (i.e. what it is like for a system to undergo certain causal processes necessary for its continued existence). If physicalism is the view that all there is to nature is captured by extrinsic S&D, then I think we have strong reasons to disfavor physicalism as our metaphysical picture, and Chalmers's S&D argument is a good reason for that. However, if we have a more nuanced understanding of structure and dynamics, then we can set aside

the conclusion that extrinsic S&D explanation wouldn't suffice for an explanation of consciousness but, nonetheless, we may be able to give an intrinsic S&D explanation of consciousness, particularly those internal S&D properties which IIT seems to be concerned with. Now perhaps the form of a complete explanation will not be IIT, but some other information-theoretic explanation, one which recognizes the relevant differences in informational relationships from the external and internal perspective, between the extrinsic and intrinsic features of a system. The goal has merely been to crack open a door many had presumed shut, what further arguments need to be made remains to be seen.

Conclusion:

It has been the aim of this essay to show that not all structure and dynamics are equal, as a way of opening up the possibility of looking for the right kind of S&D explanation of phenomenal experience. I have argued that we should look towards IIT, or some similar view, which might capture what I'm calling *internal S&D* as a fruitful way forward in explaining phenomenal experience. Opening up this possibility I think this represents a significant step forward in what kind of explanations would suffice to explain phenomenal experience and will hopefully show the main flaw behind the hard problem of consciousness. Although I share sympathies with Chalmers on this matter, I think ultimately his characterization of S&D fails to take into account the interesting and illuminating processes and relationships that arise when we focus on intrinsic interactions rather than the extrinsic components of certain systems.

Both Materialism and idealism have been guilty,
unconsciously and in spite of explicit disavowals,
of a confusion in their imaginative picture of matter.

Bertrand Russell (1927)
"Analysis of Matter"

You say you never compromise
With the mystery tramp, but now you realize
He's not selling any alibis
As you stare into the vacuum of his eyes
And say "Do you want to make a deal?"

- Bob Dylan (1965)
"Like a Rolling Stone"
Highway 65 Revisited

If all has gone well up to this point, and I hope to have shown it has, then I've shown a few things. Firstly, that my argument as to why IIT, or by extension other possible novel information-theoretic accounts of consciousness are in a position to account for phenomenal experience. Secondly, that the hard problem is only a problem when we have a restricted notion of structure and dynamics one that doesn't allow for the novel intrinsic structure and dynamical features that we encounter in complex systems. And lastly, that given the right tools and type of explanation we can begin developing a plausible and naturalistic account of how phenomenal experience arises as the result of certain types of information-theoretic relationships. What remains is to make a "deal" of sorts, that is, given what has come so far in the previous four essays what is the broader metaphysics of consciousness that we should adopt? Do we make a deal with the devil and adopt physicalism, has this all been an exercise to ultimately vindicate a problematic metaphysics of the mind after all? Or, rather, is there a different kind of metaphysics of consciousness that would accommodate the results more readily and in a more interesting way? I will be exploring the latter option and will be proposing what Information-Theoretic Neutral-Structuralism (ITNS). My hope is that this will offer new avenues of research into the metaphysics of consciousness, namely, avenues that avoid the devil's crossroad between physicalism, dualism, panpsychism, and idealism.

5. Information-Theoretic Neutral-Structuralism: A Conjunction of Neutral Monism and Information-Theoretic Structural Realism

Introduction

If we take the lessons from the previous four essays, then we are left with a rather difficult question which is the following: if we are to adopt the tools and concepts employed in a certain variety of information-theoretic explanation of consciousness – in this instance IIT – then what is the ontological and metaphysical picture we are left with? That is, should we want to resist falling into old traps of the battle between physicalism and anti-physicalism then we should look for a metaphysical framework that stays neutral in this fight – the consciousness wars' equivalent of Switzerland stuck between the Allies and the Axis. What will follow are what I take to be two plausible candidates for such a project, one's which come with their own set of troubles, but which I think are preferable to having the mind-body/hard problem looming as a consequence of one's metaphysics. Both of these options will be ones that take on board the principle of informational realism, which is as follows:

> **Informational Realism (IR)**: "the view that the world is the totality of informational objects dynamically interacting with each other (Floridi, 2004, p. 1) … instrumentally and predictively successful models (especially, but not only, those propounded by scientific theories) at a given LoA[55] can be, in the best circumstances, increasingly informative about the relations that obtain between the (possibly observable) informational objects that constitute the system under investigation (through the observable phenomena)." (Floridi, 2004, pp. 6–7)

Floridi borrows his notion of 'objects' in this context (when he speaks of 'informational objects') from object-oriented programming (OOP). The shift from thinking of representing data in an algorithmic sense, as something that takes an input, processes it, and gives some output as defined by the rules of the algorithm, to manipulating the objects as clusters of data; the difference being treating the data as the object to be manipulated rather than the logic procedures that govern it. Floridi uses the example

[55] Floridi (2004) defines a level of abstraction (LoA) as "[a] LoA consists of a collection of observables. An observable is an interpreted typed variable, that is, a variable with a well-defined possible set of values together with a statement of the properties of the system under consideration for which it stands. The target of a LoA is called a system. A system may be accessed and described at a range of LoAs and so can have a range of models." A more detailed account of level of abstraction (LoA) can be found in Floridi (2011, chap. 3, 2008).

of a pawn in a chess game to elucidate this shift in thinking. There is a sense in which a pawn has a number of contingent features, shape, size, color, smoothness, etc., but those aren't the properties one is concerned with when they play chess. Rather we care about the pawn's strategic position on the board (what it can do and not do in relation to other pieces on the board), what possible moves it can make (it can move only one space at a time forward, attack diagonally, move two spaces if its first move, etc.). We treat pawns as informational objects in this sense according to Floridi, as he writes how OOP treats informational objects "…data structures (e.g. the pawn's property of being white) and their behaviour (programming code, e.g. the pawn's power to capture pieces only by moving diagonally forward) are packaged together as (informational) objects)" (Floridi, 2004, p. 5). Those informational objects are then hierarchically organized and inherit the characteristics of the level below, which are then deployed following some general rules for modelling at a particular level of abstraction. We can understand an informational object then as a cluster of data, the cluster of data having various well-defined characteristics, which are "differences *de re*, i.e. mind-independent points of lack of uniformity" (Floridi, 2004, p. 5). When I speak of an informational object/element/entity I am using it in the sense which Floridi does, though with one qualification: Floridi takes these objects in an ontological sense, and I take them as one useful way to talk about an informational object/entity/element.

I will however be arguing for a form of informational realism which is a conjunction of neutral monism and information-theoretic structural realism. I will argue in this essay that this gives us the following thesis:

> **Information-Theoretic Neutral-Structuralism**: information is a mind-independent feature of reality, not merely an abstraction from or as a result of physical and/or mental properties/processes/relationships, it is a distinct ontological category one which is prior to those properties, processes, relationships which we use the concepts 'physical' and 'mental' to describe. This nature is revealed through the structural relationships we discover in our natural investigation of the world, which point to patterns of informational structures. Put another way, following Russell, information is the *common ancestor*, neutral but prior.

This view would of course be in conflict with those who take information to be nothing over and above those physical representations (Landauer, 1996) which one might usually take as prior, more fundamental, etc., than information. The informational realism I adopt is also slightly weaker than Floridi's as I am not committed to maintain a notion of information objects in the way Floridi

advocates, avoiding any kind of anthropocentric constraints on our ontology, as Floridi is committed with his notion of an informational object borrowed from OOP. Although I am not committed to informational objects in the way Floridi is, I nonetheless see the notion as useful in conceptualizing an informational object, though I land in favor of the view that such "objects" are merely points of lack of uniformity in the structure of the world.

The reader has most assuredly reached this point in the dissertation and felt that a firm definition of information is not only desired but necessary should one be basing their metaphysics on such a notion. I have, up until this point, resisted landing down firm on what such a notion of information is and how one should define it. This is largely due to the fact that I intended this dissertation to be something of an exploration of discovery into what such a notion of information might look like and how it might help us make sense of the natural world and the structure of reality, as well as, to shed light on consciousness and its place in nature. However, all expeditions must end, and this last essay is something of an expedition report and as a result I feel it best to lend my rough definition of information[56] to help place down a mile marker for further investigation.

> **Information**: A unique *n*-dimensional structure or shape in an abstract space of possibilities which uniformly increases as novelty increases. The more unique or novel the shape of the structure/object/process the more information contained in that structure/object/process.

Here one may well see the heavy influence which IIT has had on this notion of information. This is largely done to accommodate lessons gleaned from IIT, in that when one unfolds the cause-effect structure of a system, one finds that the information contained in a system's structural relations between its parts can in some instances be greater than those same parts taken individually. Also, accommodating Shannon's original insight that information has to do with novelty or surprisal, that is the uncertainty that is reduced by a system being in a particular state or a particular message appearing given a string of letters. One may well see the apparent usefulness in this definition as well, should a system present with a high degree of organizational complexity then the information uniformly increases with such novel structure. With this now done we can set our sights on the main

[56] I'm indebted to my external examiner Kelvin McQueen for his useful comments in needing to come down firmer on a definition of information and for helping to organize my own thoughts on this matter, particularly the idea that information is a shape in a space of possibilities.

goal of this essay and that is the broader metaphysical/ontological implications of what has been discussed thus far.

This essay will be an investigation into what comes, given the context of coming up with an information-theoretic and natural account of phenomenal experience, if we ask the question "what is the nature of reality?" if we take information as our ontological starting point. If it can be meaningfully said that information can have an existence which is ontologically prior to those properties/processes/relationships which we normally speak of as physical or mental then I think we are in a position to present an interesting and novel metaphysical picture of the world as it concerns the problem of consciousness. This work will be a motivation of such an endeavor by examining two possible metaphysical pictures, the conjunction of which I think can accomplish such a task – those being (1) information-theoretic neutral monism (Sayre, 1976) and (2) information-theoretic structural realism (Ladyman et al., 2007). Each I think are interesting metaphysical alternatives to some of the options currently available to the philosopher of mind concerned with these topics and so I hope the conjunction of the two will offer a fresh take on these perpetually vexing issues of phenomenal experience and its place in the natural world.

Section 5.1: A Plea for Neutrality

I want to begin by first arguing that both of these ontological systems are searching for the same thing, in a different way (this will be explained in much more detail in what comes in the following sections). That same thing is a neutral characterization of what the nature of reality's constituting entities or processes are. This has, for the greater part of human intellectual endeavors been between mind and matter, or as it's more commonly put in our current time, the mental and the physical. Accepting this dichotomy as the grounds for discussion is to accept defeat in my eyes, you accept the incommensurable nature of the two ontological categories and then attempt to make sense of whether there is a divide, or how to build bridges across it, or show there isn't actually a divide, or the divide is some kind of user illusion , etc., ad nauseum. There is thus a need to change the rules of the game and break away from this way of thinking, I see the push towards developing neutral monisms in the early 20th century as a marker of this type of reaction and thinking. As such, I will use the motivations from Russell for developing his neutral monism as a way of explaining the motivations of developing a neutral framework in the first place (§2). Following this I will then give an explanation of what structural realism is, to help put into context both the similarities between the projects and

their important differences. I do this to show that both stem from a desire to create an ontologically neutral and useful metaphysical framework under which to carry out naturalistic explanations of the world. I will then move on in the section following (§3) to explaining an information-theoretic neutral monism (Sayre, 1976) that attempts to do what Russell could not, before then moving on (§4 & §5) to explaining an information-theoretic structural realism (Ladyman et al., 2007). The end result (§6) will be a view which I think entails what I'm calling Information-Theoretic Neutral-Structuralism (ITNS).

Section 5.2: What is Neutral Monsim?

Neutral monism, in its simplest expression, is the view that the ultimate nature of reality is neutral with regard to whether the fundamental entities in the world are physical or mental. Now there are different ways by which one may interpret such an expression, I will be adopting the "neither view" which amounts to that the fundamental entities which compose reality are *neither* physical nor mental, but some third category of entities. I will use Russell's characterization of neutral monism, as I think it puts us in the perfect position to ask the correct question about the underlying nature of reality given what has been discussed thus far in this dissertation, as Russell says:

> "The stuff of which the world of our experience is composed is, in my belief, neither mind nor matter, but something more primitive than either. Both mind and matter seem to be composite, and the stuff of which they are compounded lies in a sense between the two, in a sense above them both, **like a common ancestor**." (Russell, 1921, p. 2)

A good way of understanding the current work is an exploration of what this *common ancestor* is between mind and matter. It will come as no surprise to the reader that this common ancestor I will argue is information, and that IIT offers us, in the current time, the best tool set to approach this common ancestor in attempting to explain consciousness. Note, this is not to say that things do not present themselves as physical or mental, the fact such concepts are continually employed lends one to pragmatically adopt those concepts in describing some phenomenon. When I speak of the interaction of planetary objects interacting with one another in space it is perfectly adequate and appropriate to merely talk of such things in physical terms. Conversely when discussing my desire to reach for a beer

from the fridge it is perfectly adequate and appropriate to speak of this in terms of the mental state I occupy, the behavior that manifests as a result of such a desire, the inebriation that follows suit, and so on. I don't think neutral monism is committed to such descriptions being meaningless, only that it is a mistake to take these as ontological categories that describe the nature of reality. Neither of these descriptions are monistic explanatorily either, they are composite to follow the wording of Russell, my understanding the relationship between planetary bodies or my desire for beer is composed of a number of different concepts and descriptions which are an interplay of what would be called mental or physical. The problem arises when one asks, "but which of these is more fundamental?" To which the only answer, must be in my opinion, neither! This essay is an attempt to motivate why neither of these categories should be considered as candidates to explain the underlying nature of the fundamental features of reality.

Given that my motivations for this work follow those of Russell's its worth mentioning why I don't consider the project from the recently resurgent program that has been taking place in the past few years, that is the Russellian Monism research program; especially since there has already been work done on the connection between Russellian Monism[57] and IIT (Grasso, 2019; Mørch, 2019, 2018). Although I'm sympathetic to the project of Russellian monism, I think the ascription of the intrinsic or categorical nature of matter being phenomenal, as opposed to something less mysterious, is a hasty and uninformative move on the part of the Russellian Monist. The move strikes me as replacing one mystery with another. We don't have an explanation of how or why phenomenal character results as a natural process and I find the ascription of phenomenal properties being the ground for all matter a rather uninformative and impractical move. One can take Russell's discussion of the limits of physics ability to explain the intrinsic nature of matter and do as the modern Russellian monist does and posit phenomenal properties as the categorical nature of matter, though I think that runs into the same problems that Russell's own neutral monism ran into in that it had no practical use (more about this in the next section). The limits of Russell's own neutral monism will be discussed in the next section (§3) to motivate the need for a positive account of neutral monism that offers practical and explanatory use not only to the philosophical question of consciousness but the scientific one as well. The upshot I see of using information rather than the phenomenal as the categorical nature of matter, or rather the more fundamental common ancestor of mind and matter, is that we have a well-

[57] Russellian monism is the view that physics stays silent on the intrinsic nature of matter following Russell's (1927) criticism about the limits of physics to explain the nature of mind and matter. Modern day Russellian monists take this as an opportunity to accommodate the phenomenal in our overall picture of matter by positing the intrinsic nature or categorical nature of matter as being the phenomenal. For a thorough defense of this position c.f. (Goff, 2017).

established mathematical framework to tackle information formally (that being communication theory and branches of probability theory) and a plethora of applications in the natural sciences.[58] It's a concept that is already heavily utilized in discussing the mind and brain, both in philosophy and science, and thus has a natural home in our current project. There is, in my opinion, no 'thing' floating out in nature which we might call phenomenal distinct from more fundamental information-theoretic properties/processes/relationships.

Section 5.3: Information-Theoretic Neutral Monism

With the advent of Shannon's formalization of information something of a revolution took place in the mind and brain sciences. There was now a formal framework under which to couch our explanations of the mind, one which seemed perfectly apt in describing the properties of the mind and our mental lives. Kenneth Sayre (1976) in his book *Cybernetics and the Philosophy of Mind* offers an information-theoretic neutral monism, one which attempts to capture the usefulness of information theory and offer a robust neutral monism. Sayre was dissatisfied with the impotence with which Russell's (1917) neutral monism offered as an ontology. His criticism amounts to the usefulness of Russell's characterization of the neutral element, in that any application of the neutral "sensibilia" would have to be translated back into physical or mental concepts to be of any use in explaining particular phenomenon. As Sayre writes,

> "The basic weakness of the Russellian program, which renders it useless for any practical purposes, is that the concept of sensibilia is devoid of explanatory power. Despite their alleged theoretical applicability to both the physical and mental, the concepts of sensibilia and of perspective space do nothing to increase our understanding of phenomena in either domain. No phenomena are explainable otherwise, and this is true in particular of modes of interaction between body and mind." (Sayre, 1976, pg. 13)

[58] Here someone may well object, but isn't the fact there's a mathematical framework play into the Russellian monist's hand? Wasn't this the exact worry that Russell raised in the first place about the limitations of mathematico-causal explanations we get from say physics about the nature of matter? I think conflating the mathematical with the physical is a mistake, mathematics isn't necessarily physical (thank you to Tim Crane for pointing this out), and so just because a theory has a mathematical framework doesn't directly entail it's physical. Just as we might ask in what sense is number theory physical?

And goes on to say…

> "A consequence is that any explanation of either mental or physical phenomena that came to be couched in the neutral framework would have to be translated out of a nonneutral context in which it had been independently achieved. Despite the ontological and methodological priority that might be claimed for the neutral framework, it is sterile for the explanation of actual phenomenon and hence parasitical upon existing theoretical structures. It is no cause for wonder that physicists and psychologists, among others, have not taken neutral monism seriously, despite its express purpose of clarifying the foundations of the sciences in question." (Sayre, 1976, pg. 13)

We might set out here Sayre's dissatisfaction with the Russellian program as a set of principles that must be satisfied by any positive account of neutral monism – let's call these the *principle of practical use* and the *principle of epistemic use*. Moving forward it will be useful to keep these two principles in mind as requirements of the neutral monism I am arguing for in this essay, they are as follows:

> **Principle of Practical Use**: any information-theoretic neutral monism must be practically useful, in so far as, such an ontological framework must provide clarifications, whether conceptual, explanatory, or experimental, for providing a foundation for the natural science that rest on it.

> **Principle of Epistemic Use**: any information-theoretic neutral monism must be epistemically useful, in so far as, such an ontological framework provides a unifying explanatory framework, by which those natural sciences that rest on its foundation might be recast under its conceptual framework.

I will be using these principles at the end of the current work to determine whether I've offered the foundations of an ontological framework that satisfy these features. Any neutral monist framework that fails to meet these requirements fails in the sense of Sayre's criticisms of Russell's neutral monism.

These principles will serve as our litmus test for the fruitfulness of pursuing such an ontological framework past the scope of the current essay.

The project remains then, if one cannot adopt Russell's form of neutral monism, then a new one must be developed. Whatever entity plays the role of the neutral element in our ontology must strictly adhere to the previous principles we defined above. For Sayre, and consequently for myself in the coming sections, this fundamental element is *information*. This neutral element must be both applicable to the domain of the physical and to the domain of the mental.

I think on first reflection the applicability of information to the mental domain is perhaps the one that needs less motivation than explaining how the physical domain arises as a result of certain information-theoretic entities and processes. The analysis of mental states in terms of information has been widely used across the philosophy and sciences of the mind.[59] There are those who take the underlying nature of physical reality to be the result of information-theoretic processes, perhaps the most famous being Wheeler's (1989) digital physics, often referred to as "it from bit." The basic premise being that one can validly understand all fundamental physical interactions as the posing and answering of yes or no questions, entailing that our interactions with the physical world take the form of a communication channel. Ladyman et al. (2007) gives his Information-Theoretic Structural Realism which at its core is an information-theoretic ontology (More on this in §4). Work has already been done as well on understanding IIT from a quantum field theory perspective (Barrett, 2014). However, the scope of this essay is not to give a detailed explanation of how to translate the neutral entity – information – to physical or mental but rather to lay the groundwork for future work on this topic. I rather want to focus on giving a coherent account of the neutral entity using the problem of consciousness as our target for a successful neutral monism, once that is done it then needs to be shown how to understand the physical and mental in such a neutral ontology.

So, what is the relationship between the neutral entity and the physical and the mental? Most often people take this as asking how one can derive the physical or mental from whatever neutral entity one posits. The way one should caste the question is to switch the priority of inquiry, the question would usually be posed as, "can we derive the physical/mental from information?" I think this is the wrong way to ask this question, and one should rather pose the question as, "could we conceive of the physical/mental without information-theoretic properties or relationships?" I think

[59] I say this as there are many relevant senses in which the connection between the mind and information has been established as being relevant (in particular c.f. Chalmers, 1996b; Dennett, 2017; Dretske, 1981, 1997; Koch, 2012; Tononi and Koch, 2015a)

the answer to the first question could easily be yes, as it's a matter of conceptual mapping and carving. The second question though is more telling in terms of the ontological question we are presently concerned with, that is, what is the underlying nature of reality.

I'm at a loss to understand what it would mean for my mental states to be about something, that is the intentional nature of my mental states, if they don't give one information about either the mental state in question or the environment in which I find myself (chapter 3 of this dissertation focuses on this issue).[60] My perception of the world is perhaps an even more salient example, as it seems rather straight forward that the reason my perception is the way it is, is because it gives me information about the world around me, information which then is brought to the fore for my cognitive control and modifies my behavior and actions. Equally, we might understand all physical descriptions in such a way, as any physical process is one governed by the laws of thermodynamics, that is a series of transformations of energy. Sayre (1976, pp. 36–40) rightfully relies on the laws of thermodynamics to explain how the physical can be caste as fundamentally derivative of information. This is in-line with those views of physics which claim that the underlying nature of those properties and entities physics deals with are information-theoretic, specifically Wheeler's (1989) digital physics.

Developing an ontological framework which doesn't only prioritize physical entities/processes (physicalism) or mental entities/processes (idealism) or the combination of the two though distinct (dualism) means that whatever theoretical entity we posit as the common ancestor needs to be neutral and allow for the translation of those previous entities/processes into a new language. Information I think offers us the best option as such a theoretical entity since we already see it deployed so readily when discussing both physical properties and mental properties, the real matter that needs discussion is flipping the priority question, if that discussion can begin, not asking whether the physical or mental are more fundamental than information, but whether information is more fundamental than the physical or mental, interesting progress can be made on developing a more unified picture of reality.

[60] Someone may well object that just because information is conveyed propositionally that there is no relevant sense which that adheres to the mathematical framework of information theory. I find this quite strange as a response in this context, particularly, because I would have taken it that to answer how semantics track the syntactic features of a system (say how a proposition transmitted via some communication channel has the meaning it does and its relationship to the process by which that meaning is transmitted) was part of the project of investigating the conscious mind from an information-theoretic perspective. Perhaps my opponent may find this unsatisfying as a response but given the rather early days of such a project, I hope one might lend time in developing an answer to such a worry.

Section 5.4: What is Structural Realism?

Structuralism Realism (SR), in the context of our discussion, is about the relationship between our investigations of the world (say our understanding of physics based on experimentation and observation) and what ontological claims we might make about said world, given what those observations reveal to us. Structural realism is the view that we should only permit those entities which figure into our best scientific explanations of the world into our ontology. As Worrall (1989) says of the structural realist "He [speaking of structural realists] insists that it is a mistake to think that we can ever "understand" the *nature* of the basic furniture of the universe" (Worrall, 1989, p. 122). As far as structural realism goes, understood in the context of scientific explanation, "understanding" the basic furniture of the universe may be outside our epistemic abilities. However, I don't think this restricts us from making well-reasoned speculations about what that nature might be given what we know of the natural world. Perhaps such speculation will fall victim to evidence to the contrary in the future, but at least the structuralist can accommodate this in the way SR accommodates the tension between the "no miracles" argument and theory change in science.[61] The tension between no miracles and theory change in science is resolved because although the concepts or terms we use to refer to those unobservable features of our scientific investigations may change, the structure found in our mature theories remain (at least according to the structural realist). Given the context of the present discussion, the SR side of the picture one can take as the explanatory framework under which I think claims about an information-theoretic ontology make the most sense. That is, if we take information as our ontological foundation, structural realism offers us the best explanatory system under which to couch our claims about the nature of reality. The ontological corollary is that of an information-theoretic neutral monism, that is that there is only one ultimate ingredient to reality as defined in ITNS. There may seem to be tension initially between advocating for a neutral monism mixed with structural realism, but I hope to motivate in this work why that tension is only superficial (more on this in §5).

---

[61] The no miracles argument comes from Putnam (1975, p. 73) in which he says in speaking of scientific realism "[t]he positive argument for realism is that it is the only philosophy that doesn't make the success of science a miracle." That is the empirical success of science is not some miracle if we are realists about what science reveals but rather a result of the fact that our best scientific theories approximately approach the truth about what those theories tell us about reality. The issue of theory change in science (sometimes referred to as the meta-induction problem) is the problem of being a realist about our scientific theories when we have good evidence that theories in science change and so how can we be realists about the things our scientific theories refer to when those theories will inevitably give way to new ones (Laudan, 1981).

There are two flavors of SR which are relevant for our present discussion, that is the difference between epistemic structural realism (ESR) and ontic structural realism (OSR). When one adopts structural realism there is the issue of whether the claims made are epistemic, concerned with merely our explanations of reality, or whether they constitute a metaphysical picture, describing the furniture of reality. This is reflected in the differences between ESR and OSR. Epistemic Structural Realism (ESR) is the view that science only gives us the structural properties of nature and not the true nature of the furniture of reality. This is reflected in the quote from Worrall in the previous paragraph, although as Ladyman (1998) correctly points out Worrall (1989) seems to oscillate between SR being an epistemic or metaphysical picture. Ontic Structural Realism (OSR) is the metaphysical option, in that, there is nothing more than the structure revealed to us by our best scientific theories. This is a stronger claim, since in essence it shows that ESR makes a mistake in thinking there is something over and above those structures that we find, that there is a nature distinct from the structures revealed to us in our natural investigations of the world. As with Russell (1927) thinking that there must be some nature distinct from what's revealed to us in the structure of theories seems a mistake. We may not get directly at the nature of the basic furniture of reality through our scientific investigations, but we do get clues, clues which allow us to extrapolate and speculate about what the nature of reality is, as I argue this nature is information-theoretic and that's revealed to us by the structural nature revealed to us by our best scientific theories. I'm thus, ultimately, adopting a form of ontic structural realism, as I think structure is what is revealed by our best scientific theories and the nature of that structure by way of inference to the best explanation is information-theoretic.

Section 5.5: Information-Theoretic Structural Realism

Ladyman et al. (2007) argues for a conjunction of OSR and what they call Rainforest Realism (RR)[62], the conjunction of these two theses results in their naturalistic metaphysics – Information-Theoretic Structural Realism (ITSR). Rainforest realism is just meant to capture the "scale relativity of ontology" (Ladyman et al., 2007, p. 252) in that there are a number of scales at which scientific investigations take place across the special sciences, biology at the scale of organisms or populations

[62] Rainforest Realism comes from Ross et al. (2000) as the name they give Dennett's brand of realism, Ladyman et al. (2007) adopt it because as they say "[o]urs is thus a realism of lush and leafy spaces rather than deserts, with science regularly revealing new thickets of canopy. Anyone is welcome to go on sharing Quine's aesthetic appreciation of deserts, but we think the facts now suggest that we must reconcile ourselves to life in the rainforest." (Ladyman et al., 2007, p. 234)

therein, neuroscience at the scale of our nervous system, psychology at the scales of our behavior, etc., and those are no more less or greater, ontologically speaking, than the scale of physics. ISTR is meant to capture that (1) our best scientific theories and explanations reveal the structural features of nature and that (2) the various scales at which our investigations take place are all as ontologically "real" in the sense that they pick out *real patterns* of reality. Real patterns (Dennett, 1989, pp. 38–42) are meant to be patterns that exist out there in the world which we pick up from our observation and investigations of the natural world from which will always involve our intentional stance towards those observables. Dennett writes, "I claim that the intentional stance provides a vantage point for discerning similarly useful patterns. These patterns are objective – they are out there to be detected – but from our point of view they are not out there entirely independent of us, since they are patterns composed partly of our own "subjective" reactions to what is out there; they are the patterns made to order for our own narcissistic concerns" (Dennett, 1989, p. 39). To connect this up to how Ladyman et al. understand real patterns in the context of ITSR, Ladyman et al. say "Special sciences are free to hypothesize any real patterns consistent with the measurements they accumulate as long as these do not contradict what physics agrees on"(Ladyman et al., 2007, p. 252). This falls in line with their view of RR since those real patterns are 'real' in so far as they (1) make predictions that conform to the evidence available and (2) do not conflict with what physics agrees on.

As Ladyman et al. (2007) write of ITSR,

> "According to ITSR, tables and chairs are real patterns – they just do not have fundamental counterparts. At the fundamental level, where all proper talk about entities whose status fueled the twentieth-century scientific realism debate goes on, reliance on the notional-world idea of cohesive things is just completely misleading." (Ladyman et al., 2007, p. 252)

And furthermore,

> "ITSR explains why parochial causal concepts are robust in special sciences, but also why nothing stronger than the thin notion of flow or process unifies them." (Ladyman et al., 2007, p. 279)

Turning back to the central focus of this dissertation, that is the question of phenomenal experience. Our desideratum is an ontological and metaphysical framework that accommodates (1) the lessons we've gleaned from investigating phenomenal experience from an information-theoretic perspective (chapters 1-4), (2) that our causal framework is deflationary about "causes" being a thing out there in the world distinct from our investigations of that world (chapter 2), and (3) that we should have a more nuanced understanding of structural-dynamical properties/processes/relations (chapter 4), ITSR is compatible with all of this. To comment first on the second quote above, the work in chapter two argued for a deflationary view of causation from a broadly manipulationist-interventionist causal framework, that we gain information from our investigations by perturbing well-defined variables in a target system. The important thing isn't that this tells us that there are "causes" out there making things do what they do, but that it gives us a picture of the structures of reality, namely how various things relate to one-another, what the flow and process of the furniture of reality is as described by their information-theoretic relationships. We've also gained a picture that tells us that the most interesting features of systems are not the "cohesive things," as Ladyman et al. puts it, but the push and pull of a systems information-theoretic properties and processes. For instance, for the question of how individual neurons generate the qualitative character of experience, it's not the individual components or things, but how they relate to each other and the information-structure they compose (chapter 3). Chapter four gave us a picture that even the structural and dynamical properties we encounter in our investigations of reality may have interesting differences in their properties, distinct from the 'things' or 'objects' which are involved in those structures.

There is, however, an important difference between ITSR as Ladyman et al. advocates it and the picture I am trying to draw for Information-Theoretic Neutral-Structuralism (ITNS). Namely that ITSR gives us a "weak unification" metaphysically given the empirical evidence we have. As Ladyman et al write,

> "The metaphysics of ITSR is so compatible. It doesn't imply that the universe is asymmetrical in a way that would explain the utility of the heuristics, but it explains why sciences that gather measurements from specific perspectives would use individuals and causal processes as locators. Such weak, but non-trivial, unification is the metaphysic that empirical evidence currently justifies. Nothing stronger has any naturalistically acceptable justification at all." (Ladyman et al., 2007, p. 290)

I, however, want to push the information-theoretic ontology to its bearable limits, by also combining it with an information-theoretic neutral monism, that is a metaphysical position that posits a neutral ontological category as the nature of the furniture of reality. I agree that ITSR has this limitation, in and of itself, and should settle for the "weak unification" it has metaphysically. However, I think as far as a more speculative foray into metaphysics is concerned, I'm justified in abductively making the argument that the basic furniture of reality is information-theoretic, understood as neutral and structural.

With an explanation of the motivations for developing a neutral monism (§1 & §2), specifically an information-theoretic neutral monism, and an explanation of structural realism and ITSR (§3 & §4), its benefits and limitations, we are now in a position to evaluate the thesis proposed at the beginning of this essay – Information-Theoretic Neutral-Structuralism.

Section 5.6: Information-Theoretic Neutral-Structuralism

The aim of the present work has been to propose a metaphysical framework that accommodates the work done in the previous chapters of this dissertation thus far. I think the best way to achieve this is to do what Russell (1921) suggests and discover that *common ancestor* of the physical and mental, a neutral base from which to develop an ontology. I proposed that we should adopt an information-theoretic neutral monism, in much the same way that Sayre (1976) proposes. I then followed this up by what I take to be a flavor of scientific realism that offers the most natural fit for our neutral ontologies epistemic foundations and that is in line with the work proposed in the previous essays, that is structural realism, specifically Ladyman et al. (2007) information-theoretic structural realism.

The outcome of the conjunction of these two theses I think is a strong commitment to (1) our natural investigations revealing the informational-structure of the world (as expressed in ITSR) and (2) that the nature of the underlying features of reality that our investigations are about is information-theoretic. The conjunction of these two I think can be captured under the definition of Information-Theoretic Neutral-Structuralism which I started with in the beginning, which is the following:

> **Information-Theoretic Neutral-Structuralism (ITNS)**: information is a mind-independent feature of reality, not merely an abstraction from or as a result of physical and/or mental properties/processes/relationships, it is a distinct ontological category one which is prior to those

> properties, processes, relationships which we use the concepts 'physical' and 'mental' to describe. This nature is revealed through the structural relationships we discover in our natural investigation of the world, which point to patterns of informational structures. Put another way, following Russell, information is the *common ancestor*, neutral but prior.

The consequence of the view above is that if what I have argued in this essay and in the previous essays (chapters 1-4) in this dissertation hold, then we have strong utilitarian reasons to adopt ITNS on the grounds that a commitment to information as a mind-independent feature of reality that underlies its nature offers the possibility of overcoming the purported gap between the physical and the mental. We can utilize information as the common ancestor that Russell sought in developing a neutral monism, one which we have strong reasons to adopt given what is revealed to us in the information-theoretic structure of our scientific explanations. I say utilitarian because many will find such a move less attractive than sticking to physicalism but I think the usefulness of developing a framework that can accommodate consciousness in our natural picture of the world while adhering closely to what is revealed in the structure of our theories of the world is more useful than clinging to the physical as the underlying nature of reality. I agree with Russell (1927) when he says that "[b]oth materialism and idealism have been guilty, unconsciously and in spite of explicit disavowals, of a confusion in their imaginative picture of matter." I think the confusion which Russell noted and which now with some of the work done in this dissertation thus far, is that we've failed to appreciate the differences in structural and dynamical properties. Although Russell expressed the confusion, I think he ultimately fell victim himself. If we accept the rules of the game of having to ask, "how does the neutral entities/processes give rise to the physical and mental?" then we've already lost the game. But if we switch the priority in our question, as I proposed at the beginning of this essay that we should, and rather ask "how could we possibly have the physical and mental without the neutral entity/process?" then we have the ability to propose something new and explanatorily useful. If we are to move forward in developing a robust metaphysical picture of the properties/processes/relationships which we allow into our metaphysical picture, I think a perfect first attempt is one which takes on the imaginative task set by Russell in how we picture matter and thus picture it from an information-theoretic perspective.

In section two, I gave two principles which were to stand as a test for our information-theoretic metaphysical framework – the principle of practical use and epistemic use. The principle of practical use constrains any metaphysical framework given to be useful in clarifying the concepts and foundations of those natural sciences that rest on it. Following Sayre (1976) I think an information-

theoretic ontology has the best chance of satisfying this constraint as it gives us a neutral entity/process by which to unify the various natural sciences, that is one which can explain the underlying properties/processes/relationships which we find across a variety of the hard and special sciences. The neutral element which acts as the foundation of our neutral framework must not only be applicable to those processes and interactions described at the level of physics but all the way to those other scales such as the mind and brain sciences, an information-theoretic ontology gives us the tool set to make explanations across various domains commensurable. As for the second principle, the principle of epistemic use, the ITSR side of our system accomplishes this job nicely, as it shows that we can take our structural descriptions of what is revealed to us by our scientific investigations and understand them as representing *real patterns* of the world. Our explanations are then a representation of those underlying information-structures revealed to us by our best scientific theories.[63]

I now want to address what might be seen as some initial tension in what I'm proposing as a ontological and metaphysical picture that falls out from the previous four essays in this dissertation. If SR is the view that we cannot understand the ultimate nature of the basic furniture of the universe and an information-theoretic neutral monism would be one that posits what Russell calls the 'common ancestor" being information, then isn't someone who adopts ITNS claiming to understand the nature of the basic furniture?

I think the tension here is only superficial. I'm not claiming to *know* the nature of the basic furniture of what constitutes the various scales of ontology that we find in our natural investigations of the world. I'm merely claiming that if we follow the arguments, we have better reason to posit information as the nature of what we see around us, rather than say the physical (physicalism) or the mental (idealism). I'm taking seriously Russell's criticism about the lack of imaginative picture of the nature of matter and offering a possible alternative. It's a similar move that motivates those that adopt OSR over ESR, in that, ESR is the claim that only structure is revealed to us by our best scientific theories and the ultimate nature of reality will elude us; those that adopt OSR claim that if structure is all that's revealed to us in our investigations than structure is all there is. I'm now claiming that if ITSR is what is revealed to us in our investigations and we have strong enough reasons to think an information-theoretic framework and ontology does the job of dissolving the hard problem or mind-

[63] I also think ITNS is compatible with the recent suggestion from McQueen (2019) for an interpretation-neutral IIT, though of course how McQueen's epistemic interpretations fall in line with the epistemic framework offered in ITNS remains for future work.

body problem (as I attempted to argue in chapters 1-4), then we have good abductive reasons to posit information as the common ancestor, as that which constitutes physical and mental properties.

This is of course only the first step, there remains the task of showing what kind of work ITNS can do in unifying our picture of the natural world but as far as first steps go, I think ITNS offers a novel and worthwhile attempt at developing a neutral monism that accomplishes the task of unifying our picture of reality with the reality of consciousness.

Conclusion:

I have endeavored in this work to open up the possibility for a metaphysical framework which adopts a conjunction of three theses: (1) phenomenal realism – that the phenomenal character of our experience is real and should be accounted for in our natural picture of the world, (2) informational realism – information is a mind-independent feature of reality, (3) structural realism – that our scientific investigations of the world reveal its structural nature. The view I argued is entailed by the conjunction of these three theses is what I call Information-Theoretic Neutral-Structuralism. I have argued that this is the best metaphysics to adopt as the result of the work contained in the first four essays of this dissertation and one which gives us a new metaphysical framework in which to approach the problem of consciousness and its place in the nature.

"…In order better to recognize [juger] these tiny perceptions [petites percetions] that cannot be distinguished in a crowd, I usually make use of the example of the roar or noise of the sea that strikes us when we are at the shore. In order to hear this noise as we do, we must hear the parts that make up the whole, that is, we must hear the noise of each wave, even though each of these small noises is known only in the confused assemblage of all the others, and would not be noticed if the wave making it were the only one. For we must be slightly affected by the motion of this wave, and we must have some perception of each of these noises, however small they may be, otherwise we would not have the noise of a hundred thousand waves, since a hundred thousand nothings cannot make something…

It can even be said that as a result of these tiny perceptions, the present is filled with the future and laden with the past, that everything conspires together, and that eyes as piercing as those of God could read the whole sequence of the universe in the smallest of substances.
The things that are, the things that have been, and the things that will soon be brought in by the future."

- G.W. Leibniz
Preface to the New Essays (1703-1705)

"And then the past recedes
And I won't be involved
The effort to be free
Seems pointless from above
You're looking down on me
I'd rather stay below
…
Time was so long ago
And things come back you see
To where they don't belong
And every drop of the sea is the whole ocean"

- John Frusciante
"The Past Recedes"
Curtains (2005)

General Conclusion:

If the whole of philosophy has been footnotes to Plato, as Whitehead famously remarked, then it's a fair characterization of the present work that this dissertation is a footnote to Leibniz's footnote to Plato. The quote from Leibniz that starts this dissertation I first read when I was an undergraduate, and it has stuck with me since. Leibniz must have been right, there is an impression of the pattern that the natural world takes contained at the largest of scales of complexity to the most minuscule, and there is a sense in which the past, present, and future are contained in those throws of the natural world. Leibniz attempted to give a unified picture of reality by explaining how nature hung together, Leibniz had his monads, but I think we have a much more powerful toolset to accomplish this task, and one which more comfortably fits into our natural picture of the world, information. I've endeavoured to answer at least one small part of this broad metaphysical picture, that is, what the place of consciousness is in nature if we take an information-theoretic perspective.

I've attempted in this dissertation to lay the groundwork for a bigger project, one that is concerned with finding a unifying ontological and metaphysical picture of reality that accommodates consciousness in our natural picture of the world. To borrow from C.D. Whether there can ever be a place in nature for the mind/consciousness remains as an open a question, though I've tried to motivate that the best approach to do this is from an information-theoretic perspective. I think for all those problems which arise for IIT, currently, the theory at least offers a novel and interesting attempt at accomplishing this task.

I think those who are interested in the metaphysics of consciousness are interested because it appears to be that last joint in nature that can't be carved. Consciousness challenges our picture of what the furniture of reality is, and thus presents a challenge to be overcome in understanding the nature of reality. Such a project by its very nature is philosophical, though our scientific investigations of this world can reveal clues that can inform our philosophy. Our metaphysical speculations about reality should be constrained by our best scientific theories but we shouldn't assume that we'll gain all we need from our current best scientific theories. The key is to strike a balance between the two and develop a position which does not stand in opposition to our natural sciences but one that works in unison. I've striven to do this, though I leave it to the reader to decide if this was accomplished.

I have attempted to offer answers to the two guiding questions in the introduction of this dissertation, which were the following:

- Is it possible to give an information-theoretic explanation of consciousness?
- What would the nature of such an explanation be and would it result in a novel metaphysics of consciousness?

I argued in the first four chapters that there is a possibility of giving a certain variety of information-theoretic explanation and what general form such an explanation might take. Following this in the last essay I answered whether such an explanation would entail for a novel metaphysics. In what follows I will explain in more detail how each of the essays answered these two questions in some way.

To summarize what has happened up to this point. In the first essay I argued that IIT, and by extension any equivalent information-theoretic theory of consciousness, is unable to explain phenomenal experience in a way that overcomes the hard problem of consciousness. I committed IIT to having a purely structural and dynamical notion of information, in the sense of structure and dynamics Chalmers (1996, 2003) adopts. I then showed that IIT also fell victim to the explanatory gap argument as a result. I set this out as a conditional, if one accepts the hard problem, the structure and dynamics argument, and the explanatory gap argument on their own terms then IIT is unable to explain phenomenal experience in a satisfactory way. There are then two strategies one can adopt to overcome those arguments, either (1) to show that IIT does not have a purely structural and dynamical notion of information or (2) show that there's a problem with how Chalmers conceives of structure and dynamics. In this dissertation I opted to pursue the latter. Although I feel the pull of Chalmers arguments there appeared to me something deeply mistaken about his characterization of what our structural and dynamical explanations could reveal about the world (the focus of chapter four).

Before I could get to this task, I first wanted to lay some groundwork. In the second essay, my co-author and I set IIT up against the causal exclusion problem. There were a couple of reasons for doing this and why it's relevant for the overall focus of this dissertation. Firstly, I wanted to highlight one of the problematic aspects I saw of IIT and that's its heavy reliance on the notion of 'causation,' something that I think is unnecessary from an information-theoretic perspective. There are two routes which one might take in interpreting IIT as it concerns the matter of causation, that is (1) that there

are real causal 'oomf' out in the world or (2) that causation is a matter of our explanations but ultimately, we should be deflationary as it concerns causation in the metaphysical sense. I have adopted the second of these positions and in the essay my co-author and I tried to show the damaging circularity for IIT of inter-defining causation and information. We then investigated whether IIT is compatible with a broadly manipulationist brand of interventionist causation and whether this could overcome the causal exclusion problem we posed for IIT. We showed that with slight modifications of some of IIT's claims the view can overcome the causal exclusion problem, the modifications of which satisfy the three conditions we laid out as a constraint on the view we offer: (1) it has the resources to avoid the causal exclusion problem, (2) it does not inter-define causation and information; it is not *essentially* an informational account of causation, and (3) it remains compatible with the empirical data, methodology, and conceptual *aims* of IIT. We concluded that the best way forward for IIT is to drop the strong causal language it currently utilizes for a more information-theoretic focused perspective.

In the third essay I explicated IIT's claims about the two candidate marks of the mental – phenomenal experience and intentionality. My aim was to give a reader friendly description of qualia according to IIT and to argue that IIT adopts a phenomenal intentionality view of intentionality. The hope here was to show that IIT has the resources to approach explaining and illuminating the philosophical question about what the nature of our mental states is. I wanted to at least sketch out what the explanation of how phenomenal character arises as the result of integrated information, that ultimately the nature of our phenomenal experience is a complex structure of a specific type of information-structure. As a consequence, I think it's entailed by IIT's claims that our conscious mental states are phenomenally constituted, and therefore, our conscious mental states if they are intentional, are intentional in virtue of their phenomenal character. I argued that the view that most closely aligns with IIT, that doesn't overcommit the theory to a stronger view than necessary, is the view advanced by Horgan & Tienson (2002).

I moved on in the fourth essay to argue that not all structure and dynamics are equal. That is, there are relevant difference in what certain varieties of structural and dynamical explanations might tell us about systems. I pulled from some recent work in complexity sciences that attempts to give an intrinsic structural and dynamical explanation about how we can draw a semantic notion of information by looking at a system and its connection to the environment. The main aim of that essay

was less ambitious than this though, it was merely to show, that there are structural and dynamical properties of certain systems that lie outside the scope of Chalmers's (2003) characterization of structure and dynamics. I argued that Chalmers's notion of S&D is purely extrinsic and so leaves out the kind of structure and dynamics we're interested in, the intrinsic structure and dynamics of a system. I made a distinction between external structure and dynamics and internal structure and dynamics. I argued that if there are structural and dynamical properties that lie outside the scope of Chalmers definition then his argument only applies to those properties that are captured by his definition. This leaves open the possibility that there may well be a structural and dynamical explanation of phenomenal experience by looking at the right type of structure and dynamics. I argued that IIT is an explanation that attempts to do this, by looking at the internal structure and dynamics of a system, and so ultimately, avoids the worries I raised in the first essay of this dissertation. As we saw, that the arguments were caste as a conditional in that essay, on the grounds that we accept Chalmers's characterization of structure and dynamics; but if IIT takes the form of a different kind of structural and dynamical explanation, then it avoids the criticisms offered in that essay.

The last and fifth essay, tackled the question of what metaphysics falls out of the previous four essays. There are two possibilities which are opened up after the first four essays. In one sense one could plausibly read this as a vindication of physicalism, one that accommodates a realism about phenomenal experience (though as I indicated at the end of chapter four, I think this doesn't work). The other option, and the one I argue for in the fifth essay, is that it opens up the possibility for a form of neutral monism which takes information as the neutral element a view which I think avoids the troubles of previous worries concerning the metaphysics of consciousness. The main reason I feel the physicalist route is unsatisfactory is that the physicalist adopts the same misconception about reality that those who adopt idealism, ESR, panpsychism, Russellian monism, etc., fall victim to, and that is, there is some *categorical* nature devoid from the structures that we find in our natural investigations of the world. It's for this reason that I adopt (1) information-theoretic neutral monism and (2) ontic structural realism (OSR), in the form of information-theoretic structural realism (ITSR). The conjunction of these two views I argue is an *information-theoretic neutral-structuralism* (ITNS). I reject the idea of a categorical nature, because all we have evidence for is the structure we discover in our scientific explanations, we however, are led to abductively infer the character of this structure, which I argue is information-theoretic. The view is neutral since it advocates for a 'neither' view of monism, which is that the nature of the one thing is neither physical nor mental. It's structural since the view

adheres to OSR, in that if structure is all that is revealed by our natural investigations of the world, and there is no need to posit a mystery about what may be lurking as the nature of the basic furniture of reality. I argued that the character of these structures is information-theoretic, as we understand the process/relationships/properties as being information-theoretic in nature. If such a metaphysics resolves the age-old issue of the mind-body problem or the recent incarnation of the hard problem, then I think we have strong reasons to take such a metaphysical picture seriously. Ultimately, I think such a view can do the job that neutral monism was always meant to do, offer a complimentary ontological framework that can unify our natural sciences under one roof.

There is much work that remains to be done which I want to comment on briefly in this conclusion. A more thorough survey of information theory and its applications at various scales of complexity is warranted. This has to do with the relationship between information and thermodynamic entropy at the level of physics (only briefly mentioned in the last essay), the relationship between information and biological processes and function, applications of information across the mind and brain sciences, which is perhaps the murkiest to tackle as its so profligate and oft misused. Such a survey fell far outside the scope of the present work, as I wanted to restrict the discussion to concern merely the place in nature of phenomenal experience. How the view I advocate for in the last essay fares when placed in the context of other interesting natural phenomenon is left for future work.

Furthermore, there is a question which I think stands out concerning IIT and the overall metaphysical framework I've tried to sketch in this dissertation. If we are to understand consciousness as being the result of instances of integrated information, and we have a method by which to determine which systems are local maximums of Φ (the mathematical framework of IIT), then how are we to determine the boundaries of a system? Put another way, say that integrated information is a phenomenon that pops up at various scales of complexity, what in principle reason do we have to prefer one scale to another in determining which systems are excluded and why? Let me motivate this with a simple thought experiment.

Let's say Zeno of Elea got bored wondering how Achilles could ever reach the tortoise and turned his attention to the question of how you determine which scale of complexity to measure consciousness to determine which systems have it and which do not, and furthermore at what spatial and temporal scale they do or do not? Zeno sets out one day and finds a candidate system floating out

in space, it's a blob just jetting around the abyss. Zeno remembered reading about IIT and goes about to test it. Let's assume this is some futuristic Zeno where he has both the means and spacecraft capable of testing such a system. He goes out and tests the blob and finds that it's a local maximum of Φ and must enjoy the delightful state of having phenomenal experiences (whatever those states might be like for such a system). Unfortunately, Zeno sees a ping on his radar and there's another blob floating not far away, which he quickly goes to investigate, except this time he notices something interesting, these blobs seem to be working in unison, he goes out and tests it and realizes that the blob also has a local maximum of Φ, but much to his dismay he determines that the two blobs taken together actually have a higher-degree of Φ than the previous blob taken individually. He decides to survey the entire area and soon realizes there are hundreds, and then thousands, and then millions of these blobs! He continues to test each one individually and then calculating the blobs taken as an integrated whole and continues to discover that each of the blobs have lower Φ without their additional cousin. Zeno slumps back into his chair and remembers fondly when all he had to worry about was Achilles and the tortoise racing. How is Zeno to ever know when he's finished finding the system with maximal Φ? How does he know there isn't a secret network of smaller blobs hidden at lower scales of complexity that have higher Φ? Or for that matter, a super-blob, at a higher scale of complexity that excludes those lesser blobs?

Of course, the above concern is a sceptical hypothesis, but it does give something interesting to chew on for the implications of IIT in determining which systems have phenomenal experience and which others do not. In the case of humans, we have the benefit of knowing that in the event of IIT being true, we are local maximums of Φ, since our consciousness doesn't get systematically excluded by any other system as far as we can tell (fortunately for us). But what about candidate systems totally alien to us? What are the reasons for not testing all possible connections to other systems to determine which one is the excluded system, and which one is not? This is the boundary problem for IIT, one which has interesting implications for the boundary issue in the extended mind thesis as well.

Although all this ground couldn't be covered in this dissertation, I have endeavoured to show a few key things in this dissertation. Firstly, that IIT offers a novel information-theoretic perspective from which to pose questions about consciousness as the result of information-theoretic processes. This moves the discussion about the application of information to explaining the mind past traditional

extant notions of information such as Shannon entropy. Showing the uniqueness of integrated information is of course a matter of debate, but nonetheless it's opened the discussion. Secondly, I've attempted to show that Chalmers's hard problem, of which the backbone is the structure and dynamics argument, mischaracterizes structural and dynamical properties by painting with too broad a brush. I hope to have at least cracked that door ajar to give enough room to pry it open even further in future work. There are relevant differences in structural and dynamical properties that need to be appreciated if we are ever going to gain a foothold on accommodating consciousness into our naturalized picture of the world. Thirdly, I've offered a novel metaphysics of consciousness which I think accommodates both a scientific realist position about our best scientific theories while also maintaining an eye towards the problem of consciousness. I think such a metaphysical framework holds a great deal of promise, though much work remains in fully fleshing out such a position. This dissertation is perhaps just one drop in the ocean that composes the question of what the nature of consciousness and reality is, but an ocean after all is just a collection of those small droplets of water. I hope this is the first of many such droplets.

References:

Aaronson, S., 2014. Why I Am Not An Integrated Information Theorist (or, The Unconscious Expander) [WWW Document]. Shtetl-Optim. URL http://www.scottaaronson.com/blog/?p=1799 (accessed 11.1.16).

Adriaans, P., 2010. A Critical Analysis of Floridi's Theory of Semantic Information. Knowl. Technol. Policy 23, 41–56. https://doi.org/10.1007/s12130-010-9097-5

Alter, T., 2016. The Structure and Dynamics Argument against Materialism. Noûs 50, 794–815. https://doi.org/10.1111/nous.12134

Baker, L.R., 2003. Metaphysics and Mental Causation, in: Heil, J., Mele, A.R. (Eds.), Mental Causation. Oxford University Press, Oxford, pp. 75–96.

Balduzzi, D., Tononi, G., 2009. Qualia: The Geometry of Integrated Information. PLoS Comput. Biol. 5, e1000462. https://doi.org/10.1371/journal.pcbi.1000462

Bar-Hillel, Y., Carnap, R., 1953. Semantic information. Br. J. Philos. Sci. 4, 147–157.

Barrett, A.B., 2014. An integration of integrated information theory with fundamental physics. Front. Psychol. 5. https://doi.org/10.3389/fpsyg.2014.00063

Barrett, A.B., Mediano, P.A.M., 2019. The Phi Measure of Integrated Information is not Well-Defined for General Physical Systems 10.

Bateson, G., 1972. Steps to an Ecology of Mind. The University of Chicago Press, Chicago.

Baumgartner, M., 2013. Rendering Interventionism and Non-Reductive Physicalism Compatible. Dialectica 67, 1–27. https://doi.org/10.1111/1746-8361.12008

Baumgartner, M., 2010. Interventionism and Epiphenomenalism. Can. J. Philos. 40, 359–383.

Baumgartner, M., 2009. Interventionist Causal Exclusion and Non-reductive Physicalism. Int. Stud. Philos. Sci. 23, 161–178. https://doi.org/10.1080/02698590903006909

Baxendale, M., Mindt, G., 2018. Intervening on the Causal Exclusion Problem for Integrated Information Theory. Minds Mach. 28, 331–251. https://doi.org/10.1007/s11023-018-9456-7

Bayne, T., 2018. On the axiomatic foundations of the integrated information theory of consciousness. Neurosci. Conscious. 1, 8.

Bayne, T., 2010. The unity of consciousness. Oxford Univ. Press, Oxford.

Bayne, T., Chalmers, D.J., 2003. What is the unity of consciousness?, in: Cleeremans, A. (Ed.), The Unity of Consciousness. Oxford University Press.

Bechtel, W., Richardson, R.C., 2010. Discovering complexity: decomposition and localization as strategies in scientific research, MIT Press ed. ed. MIT Press, Cambridge, Mass.

Bennett, K., 2007. Mental Causation. Philos. Compass 2, 316–337. https://doi.org/10.1111/j.1747-9991.2007.00063.x

Bontly, T.D., 2002. The Supervenience Argument Generalizes. Philos. Stud. 109, 75–96. https://doi.org/10.1023/A:1015786809364

Burge, T., 2003. Mind-Body Causation and Explanatory Practice, in: Heil, J., Mele, A.R. (Eds.), Mental Causation. Oxford University Press, Oxford, pp. 97–120.

Cerullo, M., 2015. The Problem with Phi: A Critique of Integrated Information Theory. PLoS Comput Biol 11, e1004286.

Cerullo, M., 2011. Integrated Information Theory A Promising but Ultimately Incomplete Theory of Consciousness. J. Conscious. Stud. 18, 45–58.

Chalmers, D., 2003. Consciousness and Its Place in Nature, in: Stich, S.P., Warfield, T.A. (Eds.), The Blackwell Guide to Philosophy of Mind, Blackwell Philosophy Guides. Blackwell Publishing, Malden, MA, pp. 102–142.

Chalmers, D.J., 1996. The conscious mind: in search of a fundamental theory, Philosophy of mind series. Oxford University Press, New York.

Chalmers, D.J., 1995. Facing up to the problem of consciousness. J. Conscious. Stud. 2, 200–19.

Collier, J., 1999. Causation is the Transfer of Information, in: Sankey, H. (Ed.), Causation, Natural Laws, and Explanations. Kluwer, Dordrecht, pp. 215–263.

Crane, T., 2003. The Intentional Structure of Consciousness, in: Smith, Q., Jokic, A. (Eds.), Consciousness: New Philosophical Perspectives. Clarendon Press, Oxford, pp. 33–56.

Dainton, B., 2000. Stream of consciousness: unity and continuity in conscious experience, International library of philosophy. Routledge, London ; New York.

Dennett, D.C., 2017. From Bacteria to Bach and Back. W. W. Norton & Company, New York, NY.
Dennett, D.C., 1989. The Intentional Stance. MIT Press.
Dretske, F., 1981. Knowledge and the Flow of Information. MIT Press, Cambridge, Mass.
Dretske, F.I., 1997. Naturalizing the Mind. MIT Press.
Eronen, M.I., 2012. Pluralistic physicalism and the causal exclusion argument. Eur. J. Philos. Sci. 2, 219–232. https://doi.org/10.1007/s13194-011-0041-7
Favela, L.H., 2019. Integrated Information Theory as a Complexity Science Approach to Consciousness 27.
Floridi, L., 2011. The philosophy of information. Oxford University Press, Oxford ; New York.
Floridi, L., 2009. Philosophical Conceptions of Information, in: Sommaruga, G. (Ed.), Formal Theories of Information: From Shannon to Semantic Information Theory and General Concepts of Information, Lecture Notes in Computer Science. Springer, Berlin ; New York, pp. 13–53.
Floridi, L., 2008. The Method of Levels of Abstraction. Minds Mach. 18, 303–329. https://doi.org/10.1007/s11023-008-9113-7
Floridi, L., 2005. Is semantic information meaningful data? Philos. Phenomenol. Res. 70, 351–370.
Floridi, L., 2004. Informational Realism, in: Conferences in Research and Practice in Information Technology. Presented at the Computing and Philosophy Conference, Canberra, Austrailia.
Galilei, G., 2008. The Essential Galileo. Hackett Pub. Co, Indianapolis, Ind.
Gasking, D., 1955. Causation and Recipes. Mind 64, 479–487.
Gebharter, A., 2015. Causal Exclusion and Causal Bayes Nets. Philos. Phenomenol. Res. 1–23. https://doi.org/10.1111/phpr.12247
Gibb, S., 2015. The Causal Closure Principle. Philos. Q. 65, 626–647. https://doi.org/10.1093/pq/pqv030
Gibb, S., 2013. Mental Causation and Double Prevention, in: Gibb, S.C., Lowe, E.J., Ingthorsson, R.D. (Eds.), Mental Causation and Ontology. Oxford University Press, Oxford, pp. 193–214.
Grasso, M., 2019. IIT vs. Russellian Monism. J. Conscious. Stud. 26, 48–75.
Hankins, P., 2014. Not a panpsychist but an emergentist? [WWW Document]. Conscious Entities. URL http://www.consciousentities.com/2014/01/not-a-panpsychist-but-an-emergentist/ (accessed 11.1.16).
Hempel, C.G., 1980. Comments on Goodman's Ways of Worldmaking. Synthese 45, 193–199. https://doi.org/10.1007/BF00413558
Hoel, E.P., Albantakis, L., Marshall, W., Tononi, G., 2016. Can the macro beat the micro? Integrated information across spatiotemporal scales. Neurosci. Conscious. 2016, niw012. https://doi.org/10.1093/nc/niw012
Hoel, E.P., Albantakis, L., Tononi, G., 2013. Quantifying causal emergence shows that macro can beat micro. Proc. Natl. Acad. Sci. 110, 19790–19795. https://doi.org/10.1073/pnas.1314922110
Horgan, T., Tienson, J., 2002. The intentionality of phenomenology and the phenomenology of intentionality, in: Chalmers, D.J. (Ed.), Philosophy of Mind: Classical and Contemporary Readings.
Illari, P., Russo, F., 2014. Causality: Philosophical Theory meets Scientific Practice. Oxford University Press, Oxford.
Jackson, F., 1982. Epiphenomenal qualia. Philos. Q. 32, 127–136.
Kim, J., 2011. Philosophy of mind, 3rd ed. Westview Press, Boulder, CO.
Kim, J., 2005. Physicalism, or something near enough. Princeton University Press, Princeton, N.J.; Woodstock.
Kim, J., 1998. Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation. MIT Press, Cambridge (MA).
Kirk, R., 1974. Zombies Vs Materialists. Proc. Aristot. Soc. 48, 135–52.
Koch, C., 2012. Consciousness: Confessions of a Romantic Reductionist. MIT Press, Cambridge, MA.
Koch, C., Tononi, G., 2013. Can a Photodiode Be Conscious? [WWW Document]. N. Y. Rev. Books. URL http://www.nybooks.com/articles/2013/03/07/can-photodiode-be-conscious/ (accessed 11.28.16).
Kolchinsky, A., Wolpert, D.H., 2018. Semantic information, autonomous agency and non-equilibrium statistical physics. Interface Focus 8, 20180041. https://doi.org/10.1098/rsfs.2018.0041
Ladyman, J., 1998. What is structural realism? Stud. Hist. Philos. Sci. Part A 29, 409–424. https://doi.org/10.1016/S0039-3681(98)80129-5
Ladyman, J., Ross, D., Spurrett, D., Collier, J.G., 2007. Every thing must go: metaphysics naturalized. Oxford University Press, Oxford ; New York.
Landauer, R., 1996. The physical nature of information. Phys. Lett. A 217, 188–193. https://doi.org/10.1016/0375-9601(96)00453-7
Laudan, L., 1981. A Confutation of Convergent Realism. Philos. Sci. 48, 19–49. https://doi.org/10.1086/288975
Levine, J., 1983. Materialism and qualia: The explanatory gap. Pac. Philos. Q. 64, 354–61.

Lewis, D., 1983. New work for a theory of universals. Australas. J. Philos. 61, 343–377. https://doi.org/10.1080/00048408312341131
List, C., Menzies, P., 2009. Nonreductive Physicalism and the Limits of the Exclusion Principle. J. Philos. 106, 475–502.
Lombardi, O., López, C., 2018. What Does 'Information' Mean in Integrated Information Theory? Entropy 20, 894. https://doi.org/10.3390/e20120894
Lycan, W., 2015. Representational Theories of Consciousness. Stanf. Encycl. Philos.
Machamer, P., Darden, L., Craver, C.F., 2000. Thinking about Mechanisms. Philos. Sci. 67, 1–25. https://doi.org/10.1086/392759
Marras, A., 2000. Critical Notice on Kim's Mind in a Physical World. Can. J. Philos. 30, 137–159.
McQueen, K.J., 2019. Interpretation-Neutral Integrated Information Theory. J. Conscious. Stud. 26, 76–106.
Mediano, P., Seth, A., Barrett, A., 2018. Measuring Integrated Information: Comparison of Candidate Measures in Theory and Simulation. Entropy 21, 17. https://doi.org/10.3390/e21010017
Mendelovici, A.A., 2018. The phenomenal basis of intentionality. Oxford University Press, New York, NY.
Menzies, P., Price, H., 1993. Causation as a Secondary Quality. Br. J. Philos. Sci. 44, 187–203.
Mindt, G., 2017. The Problem with the "Information" in Integrated Information Theory. J. Conscious. Stud. 24, 130–154.
Mørch, H.H., 2019. Is Consciousness Intrinsic? Journal of Consciousness Studies 26, 133–162.
Mørch, H.H., 2018. Is the Integrated Information Theory of Consciousness Compatible with Russellian Panpsychism? Erkenntnis. https://doi.org/10.1007/s10670-018-9995-6
Mumford, S., Anjum, R., 2011. Getting causes from powers. Oxford University Press, Oxford; New York.
Nagel, T., 1974. What is it like to be a bat? Philos. Rev. 83, 435–50.
Oizumi, M., Albantakis, L., Tononi, G., 2014. From Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. PloS Comput Biol 10, 1–25.
Pearl, J., 2000. Causality: models, reasoning, and inference. Cambridge University Press, Cambridge.
Putnam, H., 1981. Reason, Truth and History. Cambridge Univ. Press, Cambridge.
Putnam, H., 1975. Mathematics, matter and method, 2. ed., with additional chapter, reprinted. ed, Philosophical papers. Cambridge Univ. Pr, Cambridge.
Raatikainen, P., 2010. Causation, Exclusion, and the Special Sciences. Erkenntnis 73, 349–363. https://doi.org/10.1007/s10670-010-9236-0
Rosenthal, D., 1994. Identity Theories, in: Guttenplan, S. (Ed.), A Companion to the Philosophy of Mind. Blackwell Publishers, Oxford, UK, pp. 348–54.
Ross, D., Brook, A., Thompson, D., 2000. Dennett's Philosophy: A Comprehensive Assessment. MIT Press, Boston.
Russell, B., 1927. The Analysis of Matter. Spokesman, Nottingham.
Russell, B., 1921. The Analysis of Mind. Routledge, London.
Russell, B., 1917. Mysticism and Logic, and other essays. G. Allen & Unwin, London.
Sayre, K.M., 1976. Cybernetics and the Philosophy of Mind. Routledge, New York, NY.
Searle, J., 1984. Minds Brains and Science. Harvard University Press, Cambridge, MA.
Searle, J.R., 2013. Can Information Theory Explain Consciousness? [WWW Document]. N. Y. Rev. Books. URL http://www.nybooks.com/articles/2013/01/10/can-information-theory-explain-consciousness/ (accessed 10.5.16).
Searle, J.R., 1983. Intentionality: an essay in the philosophy of mind. Cambridge Univ. Press, Cambridge.
Shannon, C.E., 1948. A Mathematical Theory of Communication. Bell Syst. Tech. J. 27, 379–423.
Shapiro, L.A., 2010. Lessons from Causal Exclusion. Philos. Phenomenol. Res. 81, 594–604. https://doi.org/10.1111/j.1933-1592.2010.00382.x
Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., Hassabis, D., 2017. Mastering the game of Go without human knowledge. Nature 550, 354–359. https://doi.org/10.1038/nature24270
Spirtes, P., Glymour, C.N., Scheines, R., 2000. Causation, prediction, and search., 2nd ed. MIT Press, Cambridge, Mass.
Stoljar, D., 2017. Physicalism. Stanf. Encycl. Philos.
Stoljar, D., 2006. Ignorance and imagination: on the epistemic origin of the problem of consciousness, Philosophy of mind. Oxford University Press, New York.

Tegmark, M., 2016. Improved Measures of Integrated Information. PLOS Comput. Biol. 12, e1005123. https://doi.org/10.1371/journal.pcbi.1005123

Tononi, G., 2017a. The Integrated Information Theory of Consciousness: An Outline, in: The Blackwell Companion to Consciousness. Wiley Blackwell, West Sussex, UK, pp. 243–256.

Tononi, G., 2017b. Integrated Information Theory of Consciosuness: Some Ontological Considerations, in: The Blackwell Companion to Consciousness. Wiley Blackwell, West Sussex, UK, pp. 621–633.

Tononi, G., 2012. Integrated information theory of consciousness: an updated account. Arch Ital Biol 150, 56–90.

Tononi, G., 2008. Consciousness as integrated information: a provisional manifesto. Biol. Bull. 215, 216–242.

Tononi, G., Boly, M., Massimini, M., Koch, C., 2016. Integrated information theory: from consciousness to its physical substrate. Nat. Rev. Neurosci. 17, 450–461. https://doi.org/10.1038/nrn.2016.44

Tononi, G., Koch, C., 2015. Consciousness: here, there and everywhere? Philos. Trans. R. Soc. Lond. B Biol. Sci. 370, 20140167. https://doi.org/10.1098/rstb.2014.0167

Tye, M., 2005. Consciousness and persons: unity and identity, 1. MIT Press paperback ed. ed, Representation and mind. Cambridge, Mass. [u.a.] MIT.

von Wright, G.H., 1974. Causality and determinism. Columbia University Press, New York.

Wheeler, J.A., 1989. Information, Physics, Quantum: The Search for Links. Proc 3rd Int Symp Found. Quantum Mech. 354–368.

Woodward, J., 2015a. Interventionism and Causal Exclusion. Philos. Phenomenol. Res. 91, 303–347. https://doi.org/10.1111/phpr.12095

Woodward, J., 2015b. Methodology, ontology, and interventionism. Synthese 192, 3577–3599. https://doi.org/10.1007/s11229-014-0479-1

Woodward, J., 2003. Making things happen a theory of causal explanation. Oxford University Press, Oxford; New York.

Worrall, J., 1989. Structural Realism: The Best of Both Worlds? dialectica 43, 99–124. https://doi.org/10.1111/j.1746-8361.1989.tb00933.x

Yang, E., 2013. Eliminativism, interventionism and the Overdetermination Argument. Philos. Stud. 164, 321–340. https://doi.org/10.1007/s11098-012-9856-0