The Causal Structure of Conscious Agency

by

Holly Andersen

Bachelor of Science in Physics, Montana State University, 1999

Master of Science in Philosophy of Science, London School of Economics, 2001

Master of Arts in Philosophy, University of Pittsburgh, 2008

Submitted to the Graduate Faculty of

Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2009

UNIVERSITY OF PITTSBURGH

Faculty of Arts and Sciences

This dissertation was presented

by

Holly Andersen

It was defended on August 25, 2009

and approved by

James Bogen, Affiliated Professor, History and Philosophy of Science

Rick Grush, Professor, Philosophy at University of California-San Diego

Edouard Machery, Associate Professor, History and Philosophy of Science

Kenneth Schaffner, University Professor, History and Philosophy of Science

Dissertation Advisor: Sandra D. Mitchell, Chair and Professor, History and Philosophy of Science

The Causal Structure of Conscious Agency

Holly Andersen, PhD

University of Pittsburgh, 2009

The Causal Structure of Conscious Agency

Holly Andersen, PhD

University of Pittsburgh, 2009

ABSTRACT: This dissertation presents a new approach to modeling the causal structure of conscious agency, with a foundation in the metaphysics of causation and empirical tools for incorporating scientific results into an enriched causal model of agency. I use an interventionist causal analysis and experimental evidence from cognitive science to demonstrate that conscious awareness plays several significant causal roles in action. I then consider metaphysical challenges to this approach, and demonstrate that higher level causes such as awareness are legitimately causal.

I expose the flawed understanding of causation required for inferring the causal inertness of awareness from experimental evidence. This leads to a differentiation between metaphysical causal questions, about the nature of causation itself, from empirical questions, which apply causal analysis to actual systems in the world. I challenge the practice of focusing on the awareness *of* agency in order to address the causal role of awareness *in* agency on the grounds that it inappropriately internalizes conscious agency.

To demonstrate how we ought to incorporate scientific results into philosophical theories of agency, I offer an empirically enriched view of conscious agency. I rely on an interventionist

approach to develop an evidentiary framework to ascertain the extent to which conscious awareness is a causal factor in action. Based on results from automatism research, I demonstrate at least three important ways in which awareness is a major causal contributor to human action: conscious intentions or goals; conscious perceptual information relevant to the goal; and conscious execution.

I then address the problematic assumption that 'higher level' causes are derivative from lower level ones. I introduce the notion of counterfactual robustness to show how, for single tokens of causation, microphysical explanations are often explanatorily inferior to macrophysical ones, and distort the explanandum. I allay concerns about my variable choice by showing that we cannot, even in principle, replace higher level variables such as awareness with lower level variables such as neuronal processing. I introduce the notion of causal articulation in complex systems as the means by which higher level causes have lower level effects, while avoiding problems encountered by other theories of downward causation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

This dissertation benefited enormously from comments and questions from a large group of people. In particular, Sandra Mitchell, Endre Begby, and James Bogen deserve special thanks for many enjoyable and illuminating discussions, and their dedicated reading of and commenting on various drafts of each chapter. I would like to thank Rick Grush, Edouard Machery, and Kenneth Schaffner for their feedback and guidance.

Thanks also to: Robert Batterman, Nancy Cartwright, Carl Craver, Anthony Dardis, Peter Guildenhuys, Jonathan Livengood, Peter Machamer, James Mattingly, Dennis Pozega, Richard Scheines, Sam Thomsen, and Karen Zwier. Audiences at the Philosophy of Science Association meeting, two Society of Philosophy and Psychology meetings, and at Montana State University, Simon Fraser University, Georgetown, Tufts, and the University of California-Santa Cruz asked many insightful questions that improved my arguments. Support from the University of Pittsburgh HPS department and Provost's office was crucial in completion of this dissertation. I am also appreciative of the support offered by the Center for the Study of Mind in Nature at the University in Oslo. I would like to thank Endre Begby for his extensive editorial and bibliographic support.

This dissertation is carbon neutral, thanks to the assistance of Vic Andersen. Sixty ponderosa pine trees were planted to accomplish this. Over fifty years, with a calculated survival rate of 85%, these trees will sequester approximately 60-70 tons of carbon. Using a conservative estimate, this offsets all of the paper usage associated with this dissertation (including printed articles for research), plus at least ten roundtrip airplane trips.

INTRODUCTION

Causation is at the root of human agency. To be an agent is to have the capacity to cause things to happen. Agency involves a rich nexus of causes and effects: the causal influences of the world on our decisions and intentions, the mechanisms by which we interact with our environment and bring about our goals, and the effects of actions in the world. The task of investigating agency necessarily involves causation – it is the task of mapping out the various causal influences that are part of the incredibly complex phenomenon of agency.

The overarching goal of this dissertation to provide a schema by which to construct a scientifically informed philosophical representation of the causal structure of conscious agency. There are several components involved in such a schema. Before we can find out details about the causal structure of agency, we must have a viable methodology for causal analysis. Such a methodology should provide the means to distinguish between causal and noncausal relationships in systems, and be epistemically justified by the way in which the methodological tools of analysis follow from the nature of causation itself. A foundation in the ontology of causation is thus another component. We should know what we are committed to when we attribute causation. And we need the empirical resources to which such a methodology is applied – detailed information about as many facets of conscious agency as possible. When these components are connected up, we have the means to take scientific results, apply a methodology

to them, and develop an ontologically grounded and epistemically justified representation of the causal structure of conscious agency.

This dissertation provides a blueprint for the long-term project of filling out increasingly fine details in our representation of the causal relationships that are part of conscious agency. This project is much larger than can be accomplished here, but I have at least provided a philosophical foundation for this enterprise, along with a schematic demonstration of how to proceed in fleshing out such an empirically enriched view of agency.

Arguably one of the most salient features of our experience is that of acting, and doing so consciously: by forming an intention to do something, moving our bodies in order to bring about that goal in a certain kind of way, and being aware of what we are doing as we do it. We have convincing phenomenological experiences of *doing* things. The veracity of such experiences has come under fire recently from a new angle. Some research on human behavior in psychology and neuroscience problematizes the naïve view we have of ourselves as aware actors, highlighting failures in our awareness of what we are doing or why we are doing it. Some philosophers have responded by advocating theories of agency in which conscious involvement in our own action is either minimal or nonexistent. Given their apparent basis in empirical research, such conclusions are startling, with potentially disturbing consequences for moral responsibility, free will, etc.

On closer inspection, however, claims that we are minimally or not at all consciously causally involved in our own actions are not supported by a clear view of empirical research. Instead, they primarily rely on assumptions about the kinds of relationships that may or may not be allowed to count as causal; some of them involve outright fallacies in causal reasoning. To understand the real implication of empirical research on agency for conscious causal

involvement in action, we must first start with an explicit look at causation itself and the evidentiary standards for making causal claims.

When we consider the issue of causation in preparation for understanding causal relationships in agency, there are at least two different kinds of emphasis we might use. On one hand, there is the issue of what causation is: what is the ontology of causal relationships, what kinds of relata can be causes and effects? Answering these questions allows us to subsequently develop a methodology for finding causal relationships in the world. On the other hand, there is the issue of what actual causal relationships there are in the world: what specific causes and effects are involved in actual systems? Addressing this second kind of questions requires us to have a methodology for finding causal relationships, and to apply that methodology to physical systems.

The trajectory of my dissertation encompasses each of these components and ties them together into a coherent view of conscious awareness as forming an integral part of the causal structure of agency. I defend interventionism as the best available method of causal analysis currently available. Interventionism has several advantages over other theories of causation, including its applicability to empirical systems in order to yield answers about causal structure about which we did not already have information, and its compatibility with a range of metaphysical views about the nature of causation. I apply the tool of response structures from interventionism to current research in psychology. In the process of finding experiments suitable for filling in response structures, I demonstrate a somewhat counter-intuitive fact about the evidential import of the same research previously used to demonstrate the inefficacy of awareness in action. Research on automatism, the wide range of behaviors that can take place with limited to no conscious involvement, are actually crucial to demonstrating the extent to

which conscious awareness is a causal factor in action. Automatism research provides the contrast class needed to demonstrate the way in which conscious awareness is causally involved.

I show that three variables representing different avenues of conscious involvement in action are justified by an interventionist analysis of current psychological research. These are: consciously held goals or intentions, conscious perceptual information relevant to those goals, and conscious execution of action. Empirically, conscious awareness plays several important roles in action. Claims to the contrary generally involve either fallacies in reasoning, or rely not on empirical results but on assumptions about the nature of causation itself. The most common such assumption is that because awareness is 'high level', it is not genuinely causally efficacious; on this assumption, all causal efficacy occurs at lower levels.

This assumption about the lower level nature of causation could take several different forms, and I address them individually. The Causal Exclusion problem arises when we consider that any given instance of awareness is also an instance of various microphysical events. In conjunction with the physicalist assumption that there are no nonphysical causes, microphysical causes appear sufficient to bring about any effect, leaving nothing for awareness to cause. In response to this, I propose a novel solution based on counterfactual robustness, to demonstrate that the causal capacities of higher level causes such as awareness are not exhausted by the causal capacities of the microphysical events on which awareness supervenes.

A related assumption is that the appearance of causal efficacy for conscious awareness is merely a result of the variables I chose to represent it. Once we find other, more appropriate, variables to represent components of the system of an acting human, the thinking goes, those variables will replace the ones I have utilized and prove to be more genuinely causal. The most likely candidates for such replacement variables would be 'lower level' with respect to

awareness, such as ones involving neuronal activity. However, once we get specific about such a replacement attempt, the push for lower level variables hits a dilemma. Such a replacement might be in appearance only, where the supposedly reducing variables are parasitic on the original ones, relying on them to pick out instances of the causes and effects in question. Or, the replacement variables may not reduce the original variables but instead replace them with a somewhat different phenomenon. In this case, we would need to argue on a case-by-case basis that the replacement is a better way to parse the phenomenon that the original; there is no general conclusion that the lower level is always preferable, or even existent. This leads to an exploration of the interdependence of general or type causation and singular or token causation. This interdependence justifies the claim that when awareness is a cause, it acts *qua* awareness, not merely by dint of its microphysical instantiation.

In fact, not only are higher level causes metaphysically possible, I describe the kinds of conditions under which we can find higher level causes with lower level effects. Sufficiently complex systems are internally segmented in such as way as to allow for what I call causal articulation. Causal articulation describes the way in which higher level causes can affect lower level entities and processes that are distinct from the supervenience base of those higher level causes. Downward causation is possible in causally articulated systems even when we are unable to characterize the relation on the basis of which levels are designated higher and lower.

My dissertation thus develops a very solid foundation in causal ontology, based on which causes such as awareness can be attributed genuine causal efficacy. I show how to apply causal methodology derived from this foundation to research from the sciences in order to begin the project of mapping the causal relationships involved in conscious agency. This is a metaphysically deflationary account of agency – it does not rest on problematic assumptions

about causation, nor derive the structure of agency from *a priori* considerations alone. And it is an empirically enriched account: with a philosophically motivated theoretical system of analysis, it is research in cognitive science that fills in the fine structure of the causal representation of conscious agency. The trajectory of the dissertation starts with a distinction between metaphysical and empirical causal questions, spanned by methodology. I first assume a causal methodology and apply it, yielding empirical results about the causal structure of conscious agency. Then, I turn to establish the metaphysical foundation of the methodology used to establish those empirical claims. This ties together both the metaphysical possibility for higher level causation that awareness would have to have, and the empirical actuality of such causal efficacy for awareness. The metaphysical and empirical sections are thus mutually supportive.

Chapter 1 evaluates Benjamin Libet's work on the timing of conscious volition relative to neuronal processes leading to movement. I demonstrate that his results on the late occurrence of conscious volition are valid only when considered within a theory of agency in which volition is a discrete, short, act that is like a mental muscle flexing to influence neuronal activity. If we eschew this view of volition, Libet's results have no interesting implications for the timing or causal efficacy of conscious agency. I argue that we should reject such a theory of volition on the grounds that it involves a causal fallacy, that of double-counting instances of variables, and inappropriately reifies will. This leads me to distinguish between two kinds of causal questions, metaphysical and empirical ones, which focus respectively on the ontology of causation and the actual causal structures of systems in the world. Methodology connects these two: from our ontology of causation we develop a methodology, which can then be applied to empirical questions.

In Chapter 2, I criticize a widespread tendency in scientifically-informed theories of agency to take the sense (also called the experience or phenomenology) of agency as the paradigm for conscious involvement in action. To illustrate this trend, I examine the work of Thomas Metzinger and Shaun Gallagher. Utilizing the sense of agency to represent the general involvement of awareness in agency separates awareness from and then directs it at distinct agentive processes within the subject. This means that, if awareness is to be causally influential in action, it must do so by acting on those distinct processes at which it is directed, even to the extent of influencing nonconscious neuronal processes. I call this the Micromanagement Model of Agency, and levy three criticisms against it. This view distorts the causal role of awareness, making it appear causally inefficacious; it ignores externally directed elements of conscious involvement in action; and it holds awareness to a higher standard to count as a cause than that to which other causes are held. In sum, it inappropriately internalizes conscious agency.

Chapter 3 establishes the main empirical conclusion of the overall argument. I start by defending a generic form of interventionism as the best available method of causal analysis currently available. In order to model the causal structure of agency, I claim, we should start with this methodology, apply it to contemporary research in cognitive science, and use the results to construct our model. I first address a series of methodological issues that arise in applying this analysis to scientific experiments that use different variables and have distinct tasks involved, in order to demonstrate how to construct a system of appropriately mid-generality, neither too specific to generalize nor too broad to be substantive. Using results from automatism research, I find at least three distinguishable causal roles that awareness plays in action. The interventionist analysis applied to these experiments justifies the claim that consciously held goals or intentions, conscious perceptual information, and conscious execution of movement are each individually

causally relevant in action. This is my basis for claiming that, according to the best available methods of causal analysis, awareness plays an integral causal role in action.

Chapters 4 and 5 establish the metaphysical foundation for this empirical claim. I first consider whether individual instances of conscious causal efficacy are dependent on the causal efficacy of their microphysical supervenience bases. I introduce a new account of causal explanation involving counterfactual robustness. Using the framework of phase space, I establish that the causal capacity of a higher level cause like awareness is not exhausted by the causal capacities of its microphysical supervenience base. I then evaluate the claim that the variables used in chapter 3 to represent conscious involvement in agency are only apparently causal, and can be reduced to more fundamental lower level variables. By keeping track of the intensions and extensions of the original and reducing variables, I show that this is not the case: either the replacement variables redescribe the phenomenon without reducing it, or the replacement variables parasitically rely on the original variables to pick out intensions.

This leads to a discussion of the interdependence of token and type causation. Ontologically, variable causation is dependent on individual instances of causation – the primary relata of causation are single events. Epistemically, though, singular causation is dependent on variable causation. Variables pick out instances according to features they have in common, by dint of which they bring about the effects attributable to that variable. This justifies the claim that individual instances of awareness are causally efficacious *qua* instances of awareness. Finally, I demonstrate that in sufficiently complex systems, there is a causal articulation of subsystems in such a way that allows for the possibility of genuine downward causation.

CHAPTER 1: A new approach to conscious agency

The goal of this dissertation is to demonstrate that conscious awareness is a substantive causal factor in human agency, that ample scientific evidence confirms this, and, combined with metaphysically solid causal analysis, enriches our philosophical understanding of the structure of agency. This chapter sets the stage for what follows by providing a critical assessment of the contemporary debate about the causal structure of conscious agency. I have two primary goals in this chapter. The first goal is to address the role of the work of Benjamin Libet in the philosophical debate about conscious will and agency, after which I will not address Libet's work again in any detail. In this chapter, I will situate his experiment and the standard interpretation of its results in the debate about conscious agency, and clarify the philosophical stakes that ride on various interpretations of his experiment and results. His name and the shadow of his work lay quite heavily over this debate, which makes addressing his work a priority before moving on.

I will argue that Libet's experimental design requires one to accept Cartesian dualism in order to find his results significant for conscious volition. The very way in which he divides up his causal variables requires that conscious volition not have physical manifestation as neuronal activity. If it does involve neuronal activity, then we are unable to conclude anything about the causal efficacy of conscious volition from his experiment. I suggest that the message we ought to take from his experiment is that the conception of conscious volition used in Libet's experiment is deeply mistaken.

This discussion segues into a discussion that elaborates the second goal of this chapter. I will demonstrate, in subsequent chapters of the dissertation, that we lack genuinely *empirical* evidence to the effect that conscious awareness is epiphenomenal in action. There are numerous authors who claim that we do in fact have such empirical evidence for conscious causal inefficacy; their arguments, I claim, do not actually rely on empirical evidence. Instead, it is primarily *metaphysical* assumptions about causation, about what is or is not allowed to count as a cause or what kinds of causal structures are *a priori* allowable as possible structures of conscious agency, that are doing the inferential work in reaching these conclusions. This allows me to clarify the relationship between metaphysical causal questions and empirical causal questions. Metaphysical causal questions concern the nature of causation itself, what causes are or are not, and, based on this ontology, how we can distinguish between causes and noncauses in the world; in other words, the metaphysics of causation provide justification for methodologies for causal analysis. Empirical causal questions, on the other hand, *start* with a set of metaphysical, or at the very least methodological, assumptions about causation, and apply such a methodology to actual systems in the world in order to gather information about previously unknown causal structures. This clarification, the second major goal of this chapter, is also embedded in the divisions of this dissertation: the sections on Libet in this chapter criticizes one set of metaphysical assumptions about causation; Chapter 2 criticizes another metaphysical assumption and its consequences for modeling the causal structure of conscious agency; Chapter 3 demonstrates how to reliably answer empirical causal questions about conscious involvement in agency; and chapters 4 and 5 take up the metaphysical assumption that higher level causes cannot be genuinely efficacious, demonstrating that higher level causes such as conscious awareness are legitimately causal, based on a solid ontology of causation. The metaphysical consideration of causation provides a

justification for my use of the causal methodology that I apply to the scientific literature, rendering these topics mutually supportive.

1.1 Libet's volition experiment

Somehow, after decades of insightful and devastating criticism of his work, we are still talking about Libet's experiments. When the topic of conscious action arises, his name follows closely. Libet's background assumptions, methodology, and conclusions have been criticized from almost every conceivable angle, with sometimes concise and extremely lucid arguments. There is general agreement about the fact that his experimental methodology had serious shortcomings. And yet even those of us who disagree with it are still talking about his work. His results have become a general empirical constraint on any philosophical account of the causal role of conscious awareness in action. It is viewed as a data point that must at least be accommodated by empirically informed views of conscious agency. In order to offer a theory where conscious awareness is causally efficacious in action, it seems one must first explain how this is compatible with Libet's results, or justify why one can ignore his results.

Libet sought to bring the tools of neuroscience to bear on a long-standing question: "how does a voluntary act arise in relation to the cerebral processes that mediate it?" He used an overall measure of electrical activity, the readiness potential (RP), which "is a scalp-recorded slow negative shift in electrical potential generated by the brain and beginning up to a second or more before a self-paced, apparently voluntary motor act" (1985, 529). Using this RP to establish a timeline, Libet posed the question of when, in comparison to changes in the RP, "conscious awareness of the voluntary urge to act" occurs. The reasoning is that "if a conscious intention or

decision to act actually initiates a voluntary event, then the subjective experience of this intention should precede or at least coincide with the onset of the specific cerebral processes that mediate the act" (1985, 529).

Subjects were instructed to watch a glowing dot moving in a circle like a clockface. Once the dot had completed one cycle, they were to move their finger or wrist "at any time they felt the 'urge' or desire to do so; timing was to be entirely 'ad lib,' that is, spontaneous and fully endogenous" (1985, 532). Libet emphasizes that these movements are to be "spontaneous acts involving no preplanning." Subjects were supposed "to choose to perform this act at any time the desire, urge, decision, and will should arise in them" (1985, 530). This time was compared to the onset of ramp-up of the RP. Due to measurement constraints, series of 40 trials were averaged together to find the onset of the RP ramp-up.

The results are well known. The ramp-up of the RP began, on average, 345 ms before the reported onset of the urge to move. Subjects reported becoming aware of the urge to move about 200 ms before the muscle activity associated with movement was detected. Furthermore, when subjects were asked to time when they thought they actually began moving, they reported the actual initiation of movement about 80 ms before the movement began. On one end, it seems, conscious awareness is slow; on the other, it jumps the gun. The first result, that the initiation of preparation for movement occurs before becoming aware of the urge to move, has received a great deal more attention than the second, that subjects report moving before actual movement begins.

The results Libet found were, in a very important sense, startling and inconvenient. If we think of conscious volition as a key causal component in voluntary action, one that kick-starts the chain of events leading to our actions, then finding that conscious awareness is always too late

on the scene makes such conscious volition apparently causally extraneous. Of the philosophical view of conscious volition which Libet operationalized (an issue that will be taken up in detail soon), at least one piece of the total theoretical view needed to go, because as it was, the package was empirically inaccurate.

There have been many authors who criticized Libet's work in the intervening years. Some of these criticisms have been quite specific and effective. Joordens, van Duijn, and Spalek (2002) demonstrate how unaccounted-for biases in the measurement techniques used by Libet resulted in subjects reporting conscious events as occurring later than they actually did, meaning that the time gap between RP changes and conscious events is smaller or nonexistent compared to that claimed by Libet. Gilberto Gomes (1998) demonstrates that Libet mis-analyzed his timing results. The use of the clock face for measuring times leads to a distortion; once the lag time associated with reporting the location of moving objects is accounted for, Libet's effect disappears.

Even neuroscientists who explicitly place themselves in the field of research started by Libet (e.g., Haggard, Newman, and Magno 1999; Haggard and Cole 2005; Haggard and Eimer 1999) no longer use the same methodology. Readiness potentials, for instance, turn out to be too broad to sufficiently differentiate between activity associated with preparations for movement and activity associated with the conscious urge for which subjects introspect. Lateralized readiness potentials, which occur in the hemisphere opposite the hand that is about to move, are a more useful tool, although still imperfect. Philosophers have weighed in on the matter, also, pointing out that Libet relies on a Cartesian dualist understanding of awareness and neural processes, that introspecting for an urge to move does not capture conscious intentions or decisions to move, and that there are causal fallacies embedded in his reasoning (Andersen 2006;

van Duijn and Bem 2005). Despite these criticisms, Libet's work is still utilized (Mele 2003, 2008, Wegner 2002).

The ongoing need to address the philosophical implications of Libet's work, in spite of these methodological shortcomings and whether in agreement or criticism, therefore itself requires some explaining. Libet must have done something right, must have hit on some key issue or division, to account for this kind of longevity to his work. Libet has been rightly acclaimed as breaking ground by bringing conscious awareness into frontal contact with neuroscience: he explicitly asked the question of how conscious experience fits into the sequence of events which neuroscience had isolated as important for generating movement. In the face of the daunting prospect of ascertaining whether or not a causal relation exists between conscious awareness and neurological events, Libet developed a clever procedure to utilize a specific, epistemically convenient feature of causes and effects: effects cannot occur prior to their causes. By attempting to fit conscious awareness onto the timeline on which neuroscience had already plotted, Libet managed to challenge our beliefs about the causal efficacy of conscious volition.

My hypothesis concerning the longevity of his work in the philosophical arena utilizes a Duhemian perspective on Libet's background theoretical assumptions and commitments, which are necessary to concretely implement his experiment and which are themselves indirectly tested. I will uncover the distinct theoretical assumptions that are needed to connect the general causal question of whether conscious volition is a cause of action to the details of Libet's actual experimental design; how these theoretical commitments give significance to the experimental results that they would not otherwise have; and what other conclusions could be reached if one conjoined different theoretical assumptions to the same experimental data. This Duhemian perspective will be broadened to outline a spectrum of causal questions which can be asked

about the role of conscious awareness in action. At one end of this spectrum are purely metaphysical questions about causation: what it means to be causal, how we should represent causes and effects, how assumptions about the nature of causation can be utilized to develop methodologies for finding causes in the world. At the other end of the spectrum are empirical causal questions, where we start with a metaphysics and methodology of causation and then apply those tools to isolate and clarify particular causal structures in physical systems. Most issues will span this spectrum, involving both metaphysical and empirical causal questions. But it will turn out to be useful in understanding arguments against conscious causal involvement in action to clarify that these arguments are primarily metaphysical rather than empirical.

Before getting to this more general point, I will provide a breakdown of three major steps made by Libet which, in my view, account for the ongoing attention to his work. First, Libet implemented one key assumption of physicalism: he assumed that conscious awareness, if causally involved in action, should be straightforwardly measurable. Prior to Libet, probably with no small residual influence from behaviorism, psychologists and neuronscientists were much less likely to directly engage with questions of conscious awareness.

Second, in designing his experiment, Libet managed to operationalize certain key features of a standard philosophical view of conscious volition. To design an experiment that tests conscious volition, one has to start with a reasonably specific view of what volition is. What is important is that there is more than one option at this stage: there were already other philosophical conceptions of volition besides the one Libet chose, although he did utilize one common enough to be labeled ‘the standard view’. Choosing a view of volition as part of background theoretical assumptions is necessary; choosing the particular view he did is not

necessary. This choice point is worth noting because it will be returned to, in order to explore the implications of choosing a different theory of volition.

Third, Libet demonstrated that this package of theoretical views, operationalized into an experiment, led to counter-intuitive results. At least one feature, if not more, of the standard understanding of volition would have to go, because it conflicted with empirical evidence. The package of theoretical commitments had to be altered in at least some minimal fashion to eliminate tension with the experimental results. Libet recommended a specific theoretical change as the appropriate response: nothing about the view of volition changed except that the label for one variable went from 'conscious' to 'unconscious' (see diagram in the section below). Much of the subsequent debate involving Libet can be seen as not being so much about Libet's work itself as it is about what change in theoretical framework should be utilized to deal with the tension that emerged from the conjunction of the standard understanding of volition and the results of Libet's experiment.

There are three key moves here – earnestly attempting to measure conscious causal contributions to action; operationalizing conscious volition to incorporate it into the scientific domain; and demonstrating the empirical untenability of an existing understanding of how conscious volition is causally involved in action, thus requiring the reevaluation of the role of conscious awareness in action. I will argue that much confusion has come out of this topic precisely because the role of theoretical commitments has been obscured. Clarifying the distinct roles played by theoretical assumptions, especially those concerning causal structure, and by empirical evidence, will let us pinpoint the location of disagreements. I'll now unpack each of these three steps in much greater detail, sketch where the wrong turn was, and demonstrate how

we should reconceive the project of understanding the causal contribution of conscious awareness to action.

1.2 Bringing agency into neuroscience

Taking physicalism seriously, and all that it entails, is the kind of general claim we can agree to while still disagreeing about how to apply it in all the gritty details. There are (as with naturalism) almost as many varieties of physicalism as there are physicalists, but one consequence of almost every conceivable version is that if we are willing to commit to conscious awareness being a potential cause of things like actions, then we should be able to accommodate this influence within a scientific understanding of the world, and perhaps even be able to measure such influence using scientific methods.

Deciding whether or not a phenomenon is scientifically testable is not a purely empirical matter per se. It is a theoretical one: we test whether or not it is true by seeing if we can actually perform some kind of scientific investigation of claims involving conscious awareness. If conscious experience and especially conscious agency can be scientifically tested, then we already have at our disposal a number of very effective tools for understanding features of the physical world that can be utilized for this phenomenon. In particular, we have neuroscience, psychology, cognitive science, etc.

The way in which Libet decided to test conscious causal contributions was by assuming that conscious volition could be plotted on the same timeline, or included in the same causal chain, as other physical events with which we are already familiar. Neuroscience has worked out a series of events that can be temporally ordered on a single timeline which, while not as precise

nor as fully fleshed out as it will be in the future, still contains significant information about the mechanisms by which we move our bodies. If conscious agency is scientifically tractable, then we ought to be able to figure out how the sequence of events that constitutes agency fits into that timeline. As Libet knew, one feature of this timeline is the ramp-up of the readiness potential up to a second or so before movement occurs. That puts at least two points on the line already – the first discernible phases of RP ramp-up, and the occurrence of the actual movement associated with it. Testing conscious awareness and its role in action then entails ascertaining where specific features of conscious agency fit onto this timeline. This, in some broad and basic way, is what Libet set out to do with his experiments.

But there is more to bringing conscious experience into contact with scientific experimentation than simply lining up sequences of events or episodes on a timeline. While temporal ordering captures something important about a sequence of events, and constrains causal relations in some ways, such an approach still leaves us without an understanding of the kinds of causal structures into which conscious aspects of agency could fit. Timelines of events are not the same as chains of causal processes. Taking awareness seriously as an object of scientific study also means asking how agency – an intrinsically causal notion – fits in with or adds to the other sorts of causes associated with movement which we've worked out. This includes the causal forces we think of as somehow basic to any physical interactions in the world, like that of fundamental forces and particles; it also includes the kinds of causal relations we use to model interactions in the brain. These sets of causal interactions and the relationships between kinds of causes must be eventually integrated into a single coherent causal story about action.

As a first pass, then, this was a significant step that Libet took: to explicitly say that conscious agency is scientifically testable, and that we should be able to start matching up details of the timing, and hopefully thus the causal relations, of agency's conscious aspects with the physical aspects we're already somewhat familiar with. These two things can be explicitly compared – conscious agency must, if it is to have any causal impact in the physical world, be able to hold its own as a cause in the physical theories of neuroscience and physics.

There is a great deal more to be said about the particular issue of timing conscious events, but we'll leave this step for now and move to the second. Libet explicitly brought the notion of conscious volition into contemporary neuroscience; this sort of introduction must proceed in concrete ways in order to be implemented. Additional theoretical considerations were needed besides simply that conscious volition was amenable to scientific investigation. In order to ascertain where conscious volition should be plotted on the timeline of neuronal events associated with movements, Libet needed to operationalize volition into something that could be reliably measured. This meant taking a philosophical notion, conscious volition, and translating it into an experimental design, finding the results, and re-translating those results into philosophical terminology to draw conclusions about conscious agency. This is a key point that is often overlooked: there are numerous ways in which conscious volition could be operationalized, many of which would lead to different conclusions regarding the causal efficacy of conscious agency. Different philosophical views of volition will lead to different experimental designs; even a single view of volition could be operationalized in different ways. Understanding how this operationalization of conscious volition proceeds, what theoretical commitments are available and which subset of these are chosen is thus an important part of situating subsequent conclusions regarding the philosophical notion of volition with which we started.

Libet's W judgments (his shorthand for judgments of willing) are his experimental version of conscious volition, and by looking at how these compare to Ryle's characterization of the "ghost in the machine" myth, I'll demonstrate that Libet did manage to capture a common understanding of the will, one which still continues to pervade philosophical discussions of agency, and which is also clearly described and then rather devastatingly criticized by Ryle (1949). It is this mistaken notion of will, criticized by Ryle but still held by many, which Libet implemented in his experiments and on which his results directly bear.

Some interesting things follow from an approach which uses Ryle's critique to isolate several specific theoretical commitments about volition that were formalized in Libet's experiments. Many philosophers and scientists have taken issue with features of Libet's experimental design. And there are substantive issues with the details of Libet's methodological set-up: passively introspecting for an urge to move is a dubious way to represent active volition; treating neuronal activity and conscious volition as separate variables standing in some kind of causal relation involves treating conscious awareness in a dualistic fashion. These shortcomings, while serious, can be bracketed for now in order to situate the experiment in the philosophical discussion of agency. If we view Libet's experimental design from a broader perspective, many of these seeming methodological shortcomings look instead like shortcomings and ambiguities in the philosophical notion of conscious agency with which he began. Thus specific criticisms of Libet make more sense against the background view of agency he is working from; contradictions and tensions like the passive introspection for an urge as representing conscious decision-making are, in many respects, symptoms of a larger problem.

A cautionary note is in order regarding my use of Ryle's work. He used ordinary language analysis to offer an essentially deflationary solution to problems of mind, offering

suggestive nuggets of thought, rather than a fully developed positive account. The points that he made in this way are still extremely interesting, quite relevant, and, most importantly, testable against or supportable with reference to empirical evidence. My appropriation of criticisms made by Ryle will not rely on the same style or argumentation that Ryle used, but will instead appeal to empirical research. Perhaps most importantly, I am not utilizing Ryle's positive program regarding the nature of the mind and of will, only his critical points.

1.3 Operationalizing volition

I've already glossed Libet's experiments, and will now examine part of it in more detail. A key piece of his methodological set-up is the specific event which represents conscious volition: these are called W judgments, and by examining these W judgments, we can get clear on the view of volition operationalized in the experiment. While Libet's results[1] have been presented in a number of articles and books, I will here reference a representative version of his position, namely the article "Unconscious cerebral initiative and the role of conscious will in voluntary action" (1985). My focus in this section will be on developing a clear understanding of the notion of conscious volition that was operationalized in Libet's experiment, particularly by cashing out precisely what Libet measured as W judgments. These judgments are characterized and referred to in different ways throughout his article, from which we can put together a

[1] Libet actually performed a number of experiments designed specifically to investigate the role of consciousness in action and its relationship to neuronal activity (Libet, 1965; Libet et al 1967; Libet et al 1979; Libet 1989). I will only be addressing one of these, but it is easily the most well known and influential of his experiments, especially regarding the question of conscious volition.

substantial picture of how Libet characterized volition. I will bracket the methodological concerns raised by other authors, and base my argument on a charitable reading of Libet.

Neuroscientists have known for several decades that there is an increase in general neuronal activity in the premotor cortex immediately prior to any movement. This is the readiness potential, or RP, which 'ramps up' or increases in preparation for movement, beginning about a second or so before the actual movement takes place (Gilden et al., 1966; Kornhuber and Deeke, 1965). For any deliberate movement, the earliest known sign of that particular movement is the ramp-up of the readiness potential.[2] Libet compared the beginning of the ramp-up, plus the time at which the movement occurred, to the time at which subjects were first able to report that they were aware of conscious volitional involvement in the same action: "The objective was in fact to compare the time of onset of the conscious intention to act and the time of onset of associated cerebral processes" (Libet 1985, 530).

In order to measure the supposed first moment at which subjects became consciously involved in a given movement, a clock was displayed with a glowing dot circling the face. Subjects were instructed to move their finger "'spontaneously', without deliberately planning or paying attention to the 'prospect' of acting in advance" (ibid., 530), at some point during a one minute period. They were to note where the dot was on the clock face when they first became aware of the urge to move their finger.[3] This report was taken to indicate the act of conscious

---

[2] Recall, though, that the actual ramp-up is only visible when multiple trials are averaged together; this means that it is not possible to utilize such ramping up to tell, on a single occasion, that a subject is preparing to move.

[3] Requiring that subjects make their movements as spontaneously and immediately as possible is a function of the fact that up to 40 rounds of the test need to be averaged together in order to isolate the relevant changes in the RP from background noise. In order to make the trials as consistent as possible, the researchers attempted to reduce extraneous differences such as length of pre-planning that could affect the start of the ramping up of the RP.

volition – this report constituted operationalized conscious agency. Such spontaneous decisions to "move now" are often referred to as proximal intentions (Mele 1992).[4] Reports of when subjects first became conscious of this urge to move their finger, called W judgments (Willing judgments), were then compared to the time of the ramping-up of the RP and of the actual initiation of movement.

The surprising result was that the W judgment, the presumed first moment of conscious involvement in movement, consistently occurred after the ramping-up of the readiness potential. There was a consistent time lag of about 350 milliseconds between the first discernible rise in the readiness potential and the time at which the subjects reported becoming aware of their proximal intentions to move. There was an average of a half-second or so between the W judgments and the actual start of movement. Libet concluded that "cerebral initiation even of a spontaneous voluntary act of the kind studied here can and usually does begin *unconsciously*…Put another way, the brain 'decides' to initiate or, at least, to prepare to initiate the act before there is any reportable subjective awareness that such a decision has taken place" (ibid., 536; italics in original). In response to this result, Libet advances a view of conscious volition that has been sloganized as "free won't." The idea is that conscious will is not involved in initiating movement; rather, conscious volition is restricted to vetoing already-initiated movements within the short window of time between the unconscious initiation of the process and its actual culmination in movement. After the RP has begun ramping up, but before the movement has

[4] There are substantive concerns about how well conscious volition can be operationalized with such instructions: introspecting for an urge to move is at best awkward and at worst irrelevant to the way in which we ordinarily exercise our will. More on the problems associated with an introspective view of conscious volition can be found in chapter 2. I will bracket these concerns as they do not bear on the larger programmatic critique made in this chapter.

actually taken place, we become consciously aware of the pending action and have a chance to allow it to proceed, or to prevent or veto it.

Even if one rejects this alternative view offered by Libet, there appears to be a substantial challenge here to any philosophical theory according to which we as human agents are consciously involved in our own actions: if conscious involvement really is always 'late on the scene', then the causal efficacy of conscious agency is seriously circumscribed. But these results do not problematize all views of conscious volition. Rather, the experiment presupposes a specific understanding of conscious volition, one which so seriously restricts the import of its results as to render the experiment uninformative with respect to other philosophical views of conscious agency. In order to extract and evaluate this background view, we'll take a closer look at Libet's notion of W judgments, and see what they reveal about the notion of conscious volition at play.

Variously referred to as W's or W judgments, these reports by subjects are described as indicating "the time of conscious intention to act," (1985, 529) in the sense of the onset of a "specific conscious intention." By 'specific', Libet wants to isolate a particular stage of a general intention. Presumably when the subjects came in for the study and listened to the instructions, they had the general conscious intention to follow them and perform the required movement. The specific intention to act, under conditions of explicit instructions to avoid preplanning, is also called the time of intention to act, or a proximal intention. This is the moment when "and do it *now*" is added to the general intention to move one's finger. Libet is trying to isolate this very last stage before movement, when some conscious but general intention is made more precise and presumably implemented, and of which the ramp-up of the RP is a signal. Volition is a 'triggering event' of physical processes leading to movement.

By ‘conscious intention to act’, Libet means something slightly idiosyncratic: he is referring to a passive awareness of an urge, an urge to do something whose onset subjects were to identify by (concurrent) introspection (1985, 530-532). After correcting for reaction times, the time of W judgments indicate “the time at which subjects became aware of wanting or deciding to act.” There is something *prima facie* strange about speaking of becoming aware of a decision to act, especially if one is trying to ascertain whether that decision is made consciously or unconsciously. The very phrasing seems to beg the question against the consciously made decisions, presupposing a view in which we become aware that a decision has (past tense) already been made, one that our conscious selves only subsequently become aware of. Since I will be criticizing Libet on grounds for which this point is not germane, I will not address these concerns, merely note them.

The important point here is that the conscious intention Libet is speaking of is really an awareness of some internal state of oneself. This is the reading that makes the clearest sense of the instructions given in the experiment. At some moment, there is a spontaneous ‘click’ of decision-making that transforms the general intention, “I want to follow the instructions and that means moving my finger sometime soon,” to “move my finger *now*.” This is the moment for which subjects are supposed to be on the lookout and to report as soon as they become aware of it. The subjects’ awareness of this switch into readiness mode is the event for which they report the location of the glowing dot on the clock face, namely, the W judgment. The time of this event, minus the standard reaction time associated with noting the time of events on clocks, is compared to the time at which the readiness potential begins ramping-up.

Here we begin to put together a more complete picture of the notion of conscious volition that W judgments operationalize. The view is something like this: there is an ongoing conscious

experience that is in some sense directed towards the task at hand, but does not yet contain any specific component that would cause movement – it is not yet causally sufficient to produce movement. In order for any consciously made movement to occur, a particular series of events must occur. This series is initiated by some discrete event; that trigger event initiates the preparatory neuronal processing measured as the readiness potential ramp-up; and finally, the series of events culminates with a nervous signal sent to muscles, which produces the actual movement. In order for overt movement to occur, then, some event must occur to kick off this chain of events beginning with the ramp-up of the RP and culminating in the lifting of the finger. This triggering event is discrete, in the sense that it starts the process but is not already part of the process of movement itself; it is not the ongoing, conscious but general, intention to move at some time. Importantly, the kick-off event should be temporally indexed to a sufficiently precise degree so as to be compared with a second temporal index, the time of onset of the ramp-up of RP. Since the process at hand is one leading to action, it seems plausible to assume that volition is the event that triggers this process. The question being investigated in Libet's experiment, then, is whether this kick-off volitional event is conscious or unconscious.

Libet accepts that there are difficulties in straightforwardly translating conscious events, such as voluntarily made decisions, into physical correlates, such as identity with a distinctive kind of neuronal processing.[5] His experiment circumvents this problem by relying on timing differences, by essentially taking two distinct sets of processes and attempting to order them temporally into a single causal chain. The physical events just described are on the first of these timelines: the ramping up of neuronal activity, the signal sent to the muscles. The same action described in terms of conscious involvement is on the other timeline: the subjects begin with a

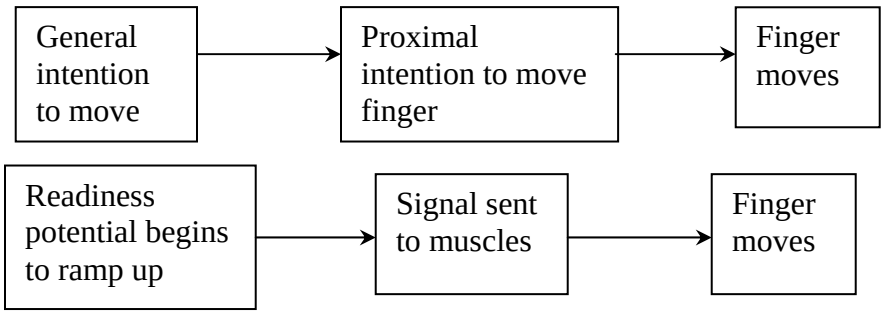
[5] See (Libet 2003).

general intention, they then have a proximal intention, in the form of an urge to move their finger, of which they become aware and only then indulge (i.e. act on), resulting in movement. Given that we can't simply translate the events on either timeline into events of the other timeline, Libet attempts to merge the two to some degree – to see where the conscious events fit relative to the physical ones when placed on a single timeline. We can use the assumption that an effect cannot precede its cause to conclude that at least some events from one timeline cannot be causes of events on the other.

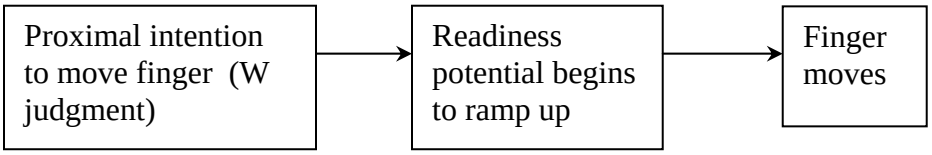
The background theoretical assumptions about volition, those operationalized by W judgments, thus emerge (see diagram 1). Where should we fit the box representing conscious volition among the other boxes representing neuronal and muscular events? Is the time of first conscious involvement the same time as that of the triggering volitional event; or, did the volition-trigger occur earlier in time than the conscious awareness of the intention to act? We assume that effects cannot be temporally prior to their causes: if it had turned out that the consciously reported awareness of the urge to move preceded the RP ramp-up, this would not have proven that the conscious decision caused the ramp-up, but it would certainly be compatible with it. This is what the "classic" understanding of volition would predict about the experiment, according to Libet. Had this been the outcome, there would presumably have been little of interest to gain from the experiment, except perhaps a weak confirmation of an existing view.

Figure 1: Comparison of timelines. Arrows indicate temporal and causal ordering

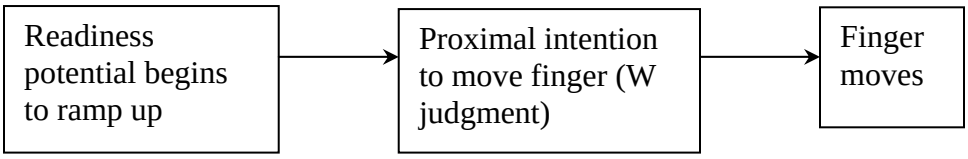
1.a Two distinct timelines, conscious/mental and cerebral/physical



1.b Prediction of the classic view of volition when the timelines are collapsed



1.c Libet's results



If, on the other hand, the chain of physical processes was already in motion before conscious awareness was reported, then that awareness could not be the initiating cause of those processes. "This timing relationship, with the 'physical' (cerebral process) preceding the 'mental' (conscious intention) … strongly supported the view that each individual RP precedes each conscious urge" (1985, 533). This result led Libet to conclude that "Unconscious initiation of the voluntary process appeared to mean that conscious free will could not actually 'tell' the brain to begin its preparation to carry out a voluntary act" (2003, 24).

If this triggering-event view is an accurate picture of volition, and if we are to be genuinely consciously causally efficacious in generating our own actions, then the triggering

event itself must be a conscious event. This conclusion is in direct contradiction to the results obtained in the experiment: the triggering event of the physical processes cannot be the conscious awareness of the urge to move. Some element of the classic picture must be given up to make it cohere with empirical observations. Libet concludes that we must give up the notion of conscious volition as initiating movement: there is an initiating volitional event, but it is unconscious rather than conscious. This conclusion, as we'll now see, only holds if we subscribe to the triggering-event view of volition as the causal precursor of physical processes leading to movement, a view that shall soon be shown to be severely problematic.

To summarize, what Libet has done is to assume that there is a triggering event, an act of volition that starts the gears rolling to generate movement. He then asks if that triggering act of volition is conscious or unconscious. Pointing to the time lag, he concludes that the triggering volition is unconscious, and that we become consciously aware of it only after the process leading to movement has been initiated. W judgments simply mark the moment when we become aware of this act of volition having occurred. This view, it turns out, is strikingly similar to the Myth of Volitions criticized by Ryle, whose criticism of the Myth I will now elaborate and apply to Libet's experiment.

1.4 A Rylean critique

There are several points I will make in this section. First, Gilbert Ryle, writing 30 years before Libet's experiment was first performed, had already executed a dense but devastating critique of the view of volition held by Libet, as well as of typical mistakes associated with this kind of view. This view has returned with some force to philosophical discussions in the form of Libet's

triggering view of volition, and Ryle's criticisms of it are both to the point and under-appreciated. In this section, I highlight how the appropriate philosophical response to Libet's results is underdetermined. Ryle and Libet can be understood as agreeing that conscious volition cannot be thought of as a discrete event initiating physical processes; Libet recommends the conclusion that initiating volition is not conscious, whereas Ryle would recommend—in my view, rightly—dispensing with the view of conscious volition as a discrete event that triggers cerebral processes. The next section argues that this underdetermination is merely *prima facie*, and that we have further reasons to reject the triggering view of volition.

Ryle's overall target in *The Concept of Mind* is the derisively-labeled myth of the Ghost in the Machine.[6] This view is so commonly subscribed to by philosophers and lay people alike as to qualify, he thinks, as the "official doctrine" (Ryle 2000 [1949], 11). It involves a number of problematic features, all of which center around the way in which an agent is cast in terms of "two collateral histories, one consisting of what happens in and to his body, the other of what happens in and to his mind. The first is public, the second private. The events in the first history are events in the physical world, those in the second are events in the mental world" (ibid., 11-12). This way of characterizing ourselves, where an event is either a physical one or a mental one, leads to absurdity, according to Ryle, for a number of different reasons, most of which will not concern us here. The primary problem with this two-history view is that it relies on a family of category mistakes; this will be explored in more detail shortly with respect to volition.

---

[6] Ryle approaches his topic using ordinary language analysis, a style of philosophical argumentation that is, to perhaps put it mildly, no longer in style. His work has been often ignored because of this, which is unfortunate, because his arguments do not rely solely on ordinary language analysis, and can often be recast without it. As such, I will bracket concerns about this aspect of his work since the criticisms I import from his work in this paper do not rely on ordinary language analysis.

In his chapter on the Will, Ryle demonstrates how a standard philosophical understanding of volition is simply another manifestation of the myth of the Ghost in the Machine. In the two-histories view, consciously executed actions pose a particular dilemma because they need to cross between the two series of events, from a conscious decision or intention in the mental history into a movement in the physical history. Neither straightforwardly mental nor straightforwardly physical events seem capable of this, which leads to one of two alternatives, according to this mistaken way of thinking. The first option is that there is no connection between the two series, and physical events run according to determinate laws entirely independently of conscious ones. According to this view, which has been advocated in various forms in recent years, we merely need to explain why the conscious events should follow the physical so closely that we mistakenly think we consciously brought about the physical movement (e.g. Wegner, 2002). The second option is that there is some special kind of mental operation capable of crossing between the two, carrying a command from the mental to the physical to bring about movement.

This second option, positing "processes, or operations, corresponding to what [the view] describes as 'volitions'" (ibid., 63), constitutes the doctrine of volitions that Ryle criticizes.

> Volitions have been postulated as special acts, or operations, 'in the mind', by means of which a mind gets its ideas translated into facts. I think of some state of affairs which I wish to come into the physical world, but, as my thinking and wishing are unexecutive, they require the mediation of a further executive mental process. So I perform a volition which somehow puts my muscles into action. (Ryle 2000 [1949], 63)

In other words, volitions are those acts whereby a conscious intention is specifically and at some definite point in time implemented physically. They are somewhat like the flexing of a mental muscle, done in such a way as to push at something physical. These volition acts are presumably

sufficient to create the physical events needed to bring about movement (given normal bodily functioning) by initiating whatever processes they need to in order to accomplish this.

The problematic view described by Ryle accords remarkably well with the view of volition Libet tested with his experiment. Libet set up his experiment in terms of these two histories or series of events, conscious mental events and cerebral physical events. Libet himself notes that this is a problem, since we don't know how the conscious events, such as making decisions to act, relate to the physical ones, such as the processes leading up to movement. He solves the problem of how to test this view by temporally integrating elements from each of the two histories. For many of the mental events, this is intractable, so he settles for finding out where on this timeline one of the most important mental events should go, namely conscious volition.

In Libet's case, as in the view described by Ryle as part of the myth of the Ghost in the Machine, volitions are those particular mental events which are invested with the power to initiate a sequence of physical events which eventuate in the overt movement of a finger. W judgments are supposed to indicate the time of conscious awareness of these internal acts of volition, an awareness of the mental 'thrusts' that translate conscious decisions into physical movement. The 'classic' picture of volition that Libet set out to test is thus a version of the Myth of Volitions: do these volitions take place early enough to be compatible with the hypothesis that conscious volition is indeed what sets in motion the chain of physical events? Or do these conscious volitions occur too late to initiate the process? Libet's conclusion is that the classic picture of volitions as both conscious and as initiating the physical processes leading to movement is wrong. What is crucial, and this will be reiterated in the next section, is that Libet started with a view identifiable as the Myth of Volitions; he found that there was a problem with

it; and he made minimal modifications to that original view in order to accommodate the experimental results. Libet merely advocated that the initiating volitions be unconscious, rather than conscious, and that the opportunity for the mental to intervene in the physical comes after the chain of events has been initiated rather than before. He otherwise preserved the rest of the classic picture of volitions.

Libet and others still subscribe to the Official Doctrine as Ryle identifies it, or at least the Myth of Volitions. Both Ryle and Libet find some internal tension or contradiction within this view. For Ryle, it is the absurdities which follow from the category mistake of thinking that 'volitions' are discrete thrusts of some mental muscle to translate the mental into the physical. Libet too finds fault with the Official Doctrine: his work demonstrates that although the standard position is to think of volition as triggering events that are conscious, this is incompatible with the results of his experiments. Volition cannot be conscious and also trigger the physical processes leading to movement: Ryle and Libet could both assent to this statement, for different reasons.

Ryle's position on the problems with this view is to jettison triggering volition acts.[7] Libet has a different solution to the problems he finds in the standard view: rather than reconceptualizing conscious volition, he maintains that volitions are triggering acts and declares them to be unconscious instead of conscious. Both find fault with the doctrine of volitions; each resolves the tension in a different way. The implication of Ryle's critical position, getting rid of volition-acts, is entirely consistent with Libet's experimental results. Libet realized that causally

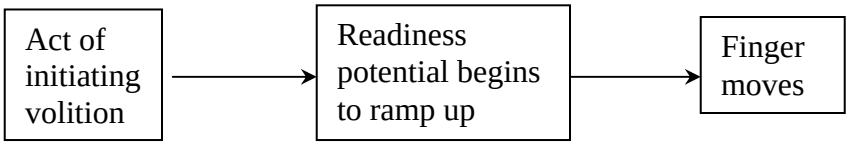
---

[7] It is important to emphasize that I am not advocating Ryle's positive view of conscious agency, when I refer to his position here, it is the critical part of his views. I am contending first, that his criticism of the Myth of Volitions is accurate, and second, that a major problem with the Myth is it's way of conceiving of volitions as little mental act-events that trigger physical events, a view that, regardless of what one subsequently holds, should be eschewed.

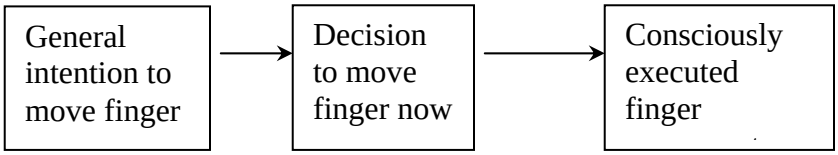
efficacious volition-acts could not be conscious. This by itself underdetermines which part of the standard view's infrastructure should be revised and which parts retained. In point of fact, Libet advocated changing conscious volition to unconscious volition. Yet this was not the only move available as an appropriate response to the results. Libet could have instead concluded that because the conscious volition-triggers he sought using W judgments were not found, this is not an adequate way to operationalize conscious volition (although this would have entailed re-designing subsequent experiments to investigate the new understanding of conscious will). Consider the following comparison of causal graphs, depicting a Libetian and an alternative way of representing the situation.

Figure 2: Two methods of representing volition

1.a Libetian representation[8]



1.b Alternative representation



We need not initially take a stand on which of these representations is a more accurate representation of conscious agency. The problem can be phrased entirely in conditional form. If the alternative representation of causal structure more accurately describes conscious will as

[8] See also figure 1. The two diagrams differ slightly with respect to Libet's representation: in this diagram, the initiating act of volition is what Libet takes to initiate the ramp-up of the readiness potential, which figured in the same position in the earlier diagram.

studied in the experiment, but we instead utilize the Libetian one when we analyze the results, then we will reach the conclusion that no event of conscious volition was found. This leads to one of two responses for the Libetian: either volition is conscious but non-initiatory of the process, or it is initiatory but nonconscious. Both of these conclusions will be inaccurate in this scenario, because they follow solely from an artifact of the causal representation we used, the theoretical commitments with which we started. Libet's results are insufficient to adjudicate between these background theoretical pictures.

This helps explain the fall-out Libet has had for the philosophical discussion on mental causation, personal responsibility, free will, and so forth. As Ryle correctly indicated, many of us fall into versions of the traps of thinking identified as the Official Doctrine. The work of Libet then strikes us as requiring an internal reorganization of that view – we recognize the inconsistency of his empirical results with a widely-held but still somewhat intuitive view of conscious volition. This leads some, like Daniel Wegner (2002), to proclaim all conscious will illusory; it leads others to defend some sense of moral responsibility and free will from the threat of determinism that unconscious volition poses. Instead, I propose that we should move one step further back and re-evaluate the view of will embedded in our initial theoretical commitments. If we did, we would no longer feel philosophically obliged to continue worrying about the implications of these timing results for conscious agency.

1.5 Rejecting the standard view of volitions

In the previous section, I illustrated how the appropriate philosophical response to Libet's scientific data is underdetermined. This section argues that this underdetermination is merely

prima facie, and that we have reasons to prefer the alternative resolution (see diagram 2) over that of Libet, because of several category mistakes in the chain of reasoning that leads from Libet's actual results to the conclusion that conscious will does not initiate action. The problematic step is the reification of will, a necessary assumption to lead from Libet's results to his conclusions. I will argue that this reification misconstrues the relation of volition to cerebral processes, and ignores the relationship between conscious will and cerebral processing.

Among the family of category mistakes that Ryle accuses the doctrine of the Ghost in the Machine of making is one that turns out to be particularly relevant to Libet's experiment. Consider first a general case. In the first chapter of *The Concept of Mind*, Ryle identifies a certain kind of category mistake. Suppose a visitor were to be shown around a campus: she sees the various buildings, the labs, the student accommodations, the lovely greenery. At the end of the tour, she asks, "Well, I have seen buildings – but where is the University?" The University, the guide explains, is not some other further thing besides what she has already seen; it is a different category of thing. A similar example is provided more recently by Nancy Cartwright (1999, 40): she washed the dishes, taught a class, wrote a grant proposal; but when did she work? Work is not another thing one does on top of those she has already done, it is simply a way to draw together all those activities into a common category which they each instantiate. Treating the University as one more building among others, or work as one more concrete activity one does in addition to the others, is a category mistake.

The same applies to the case of will. As I've just elucidated, Libet attempts to collapse events on two timelines into a single timeline. In doing so, he commits (at least) two category mistake, ignoring a key feature of causal relationships; the consequences of doing so are both a dubious reification of conscious will and the commitment of a causal fallacy. This first category

mistake is treating conscious will as a single discrete event which occurs at some given moment, as the flexing of a mental muscle rendering what had been mental into something physical. When we exercise our wills, there are many reasons to think that we do so *by* doing other things: we form intentions, we make plans and decisions, we select actions, and we consciously execute them. Taken together, these things constitute the exercise of our will, just like the buildings etc. taken together are the University. Willing is not a single event which can either be conscious or merely physical. It is a term that draws together a number of different activities and processes under a single more general term.

Consider another case. I realize I would like a cup of coffee; I form the intention to get one; I decide now is a good time and then walk to the coffee shop and purchase it. Yes; but when did I exercise volition? When did the act of volition occur – when I stood up, or when I walked over, or when I placed my order? These are the wrong sorts of questions to ask. All of the activities I just described are part of what it is to 'will' something; there is not an extra stage in the process that is the stage of willing, just like there is not an extra building that is the University and not an extra chore that is Cartwright's work. Treating willing as an extra and separate thing one could do is to inappropriately treat as concrete a concept that is not itself concrete, but brings together other concrete activities and processes. The category mistake Libet makes is to reify will as a separate act among others.

This category mistake carries a consequence that is quite germane to the conclusion about causation drawn by Libet. The consequence is an error in causal analysis that – if we allow for the sake of argument that volition can be treated as a discrete act – assumes it is coherent to simply insert conscious involvement in action (the W judgments) as potential causes and effects capable of standing in direct causal relationships to the other causes and effects such as cerebral

processing. The question in Libet's experiment is whether the box representing W judgments should go before or after the box representing RP ramp-up; in other words, whether the conscious element is a cause or an effect of the cerebral processing. This second category mistake, treating conscious will as being the same kind of potential cause as cerebral processes such as the RP, has ramifications for the connection between conscious will and cerebral processes. Either they cannot stand in a causal relation to one another, or they can stand in such a relationship but only at the cost of treating conscious will as lacking any instantiation as neuronal activity.

I'll spell out this dilemma. The basic problem is that, by ignoring the difference between a variable that represents something essentially conscious, and a variable that represents physical activity, a causal fallacy is committed. It is a widely accepted condition that in order for two things (whether they are singular events or a collection of events) to be so much as potentially capable of standing in a causal relationship, they must be logically distinct. Being logically distinct means that events that count as occurrences of the potential cause cannot also count as occurrences of the potential effect – no single occurrence can be both a cause and its own effect.[9] An example of this can be seen in the following. Suppose we have a glass of water, and we inquire about the molecules in the glass. There are a number of causal relationships we could potentially investigate regarding chemical properties of the water and its macroscopic behavior; but if we were to ask what the causal relationship is between the molecules and the water itself, we would commit this particular category mistake, and thus commit the causal fallacy. The molecules can't *cause* the water, nor can the water cause the molecules. The molecules *constitute*

[9] A good discussion of this requirement and its consequences for individuating causal variables can be found in Dardis (1992) Dardis (2008) also contains a discussion of the consequences of this requirement of distinctness for the debate on mental causation..

the water, and this is precisely the sort of logical relationship that precludes there also being a causal relationship between the same relata. There is no instance of the water separate from the instances of the molecules such that one could cause the other.

In general, causally related variables cannot already stand in some other kind of relationship, such as identity, constitution, supervenience, et cetera. These are precisely the kinds of relationships that we must assume to hold between conscious will and cerebral processing, on pain of embracing Cartesian dualism.[10] Whatever else we want to say about conscious will, and whatever difficulties, surmountable or not, there are in relating features of conscious experience to cerebral processing, we must at least acknowledge that conscious will involves some kind of cerebral processing. We can remain agnostic about the specific logical kind of relationship between cerebral processing and conscious will (e.g. identity, supervenience, etc.), as long as we recognize that we are committed to *some* such relationship. This alone makes it fallacious to treat conscious will and cerebral processing as also standing in a causal relationship at the same time. In this manner, the category mistake thus brings along with it a particular kind of causal error.[11]

Consider diagram 1 again, in which the two timelines appear, comparing the prediction of the 'classic' understanding of volition versus the representation drawn from Libet's results. The

---

[10] Libet himself does subscribe to a form of Cartesian dualism, where conscious elements of agency are entirely distinct from, and supposedly able to causally act on, neuronal processes. However, the overwhelming majority of philosophers who discuss Libet's work, especially those who rely on it to conclude against conscious causal involvement in action, would not want to advocate Cartesian dualism. The charge of such dualism, then, is not a problem per se for Libet. But it is worth emphasizing that the Cartesian dualism is, by dint of this causal mistake, built into the very structure of the experiment in such a way that one cannot endorse the results without also thereby implicitly accepting the dualism that was crucial to the experimental design.

[11] It is worth noting that both of these errors appear to rest on the overly simplified causal representation employed in the reasoning behind the experiment and from the results to Libet's conclusion. This billiard-ball model of causation, where one cause after another act in linear fashion, is ill-suited to the complex and multi-level system that is conscious agency and all its associated physical activity.

conclusion reached in the experiment was that conscious volition could not initiate the processes that are measured as part of the readiness potential since conscious awareness occurred after the readiness potential has begun to ramp up. In order to reach the conclusion that we do not consciously initiate our own actions, it is necessary to reify volition as a separate stage of the process and to incorporate the physical processes on the same time line as the conscious ones. In this diagram can be seen both the category mistake, namely treating conscious will as a discrete event like the flexing of a mental muscle, as well as the causal fallacy that follows from the category mistake, namely treating physical cerebral processes as logically distinct from conscious awareness. Both of these mistakes emerge from the background view of volition that Libet operationalized. The problems are thus endemic to the philosophical view he started with, and are not merely difficulties with the experimental set-up or with a particular line of reasoning.

Most importantly, both of these mistakes are *necessary to reach the conclusion* that we do not consciously initiate our own actions. The conclusion of circumscribed conscious causal involvement is a function of the initial philosophical paradigm of volition. If volition were not treated as a discrete event, or if conscious will were not placed alongside physical processes on the same timeline as if volition were wholly distinct from such processes, then the experiment would have little or no significance for our understanding of conscious agency. And this is, I suggest, precisely the position we should take regarding Libet's experimental results.

To wrap this point up, Libet's experiments certainly do demonstrate the incompatibility of both thinking of will as a discrete act of volition akin to flexing a mental muscle, and attributing causal efficacy to it. And, considered in a conditional form, use of this paradigm of conscious volition will lead us astray if the structure of conscious agency differs in some way from this view. But the point can be made stronger. If conscious agency were differently

structured than Libet's standard view of volitions, then his experiment reveals nothing special about the structure of conscious volition. Moreover, we have good reason to think that conscious agency is structured differently than Libet assumes. There are several foundational problems with Libet's approach, as I've just shown, in particular that it commits a reifying category mistake, as a result of which an error in causal reasoning is incorporated into the structure of the experimental design.

Besides these problems with a Libetian view of conscious agency, a consilience of other reasons speak against it. As mentioned earlier, even neuroscientists working on similar subjects no longer rely on Libet's methodology. Empirically, the experiment is flawed, relying on an overly broad measure of neuronal activity. The introspective and reactive nature of W judgments renders them a poor way to operationalize conscious intentions and decisions. And philosophically, this sort of view has been rejected not only by Ryle, but by others since.[12] Because this problematic view of conscious volition came wrapped in an experimental package, it seemed to have a gleam of scientific credibility that took precedence over 'mere' philosophical criticism, such that scientifically informed philosophical views of conscious agency had at the very least to be compatible with its results. There are sufficient shortcomings to the view and the experiment, however, that it should no longer be considered a constraint on theories of conscious agency.

This brings together the second step made by Libet, operationalizing a specific philosophical view of volition in order to incorporate conscious volition into scientific

[12] For one influential account of why this view of volitions is unsatisfactory even when we are focused on supposed pure events of willing, see Davidson (1980, 84).

investigation, and the third step made by Libet, demonstrating that at least one feature of this philosophical view of volition must be jettisoned. The reason Libet's work still features prominently in philosophical accounts of conscious agency is that it demonstrates the empirical incompatibility of integral features of a traditional philosophical view of conscious volition, namely the view that had already been devastatingly criticized by Ryle. In fact, we can map out much of the contemporary debate on this topic in terms of different responses to Libet's results, the rejection of different sets of premises to accommodate the apparent timing difficulty.

Consider the position of the standard view of volitions after Libet's experiment. It can be formulated as a collection of premises:

1. Agency is implemented in discrete acts of volition.
2. These acts are conscious (and, presumably, intentional).
3. They have physical effects.
4. These physical effects include the initiation of neuronal motor processes leading to movement.

As we've already seen, there an internal contradiction in this view: the triggering event for the specific processes leading to movement can't be all three of physically efficacious, do so by initiating brain processes, and be conscious (premises 2-4). Something has to give. I have argued that we should reject premise 1, the idea that volition is implemented in discrete acts. If we do this, then the rest of the premises, as written, no longer apply since they concerned those acts, and we are in a better position to develop the finer-grained structure of conscious agency from existing cognitive scientific work.

A prominent response by scientists and philosophers to Libet's results, including that of Libet himself, is to label the triggering volitional event unconscious instead of conscious: retaining premise 4 and rejecting premise 2. If we choose this route, we could argue that initiative volition is unconscious, and conscious agency has only a very circumscribed role in

action. There is a small window of time between the subject's reports of awareness of the decision to move and the beginning of movement, during which we could consciously veto the already-initiated movement if we chose. This is a small causal role for conscious agency, but still something. Alternatively, we could respond to the tension by arguing that conscious agency is entirely epiphenomenal, that it has no causal efficacy on physical actions: we could reject premise 3. This is approach taken by Daniel Wegner, for instance, with his Theory of Apparent Mental Causation.

And of course, we could be sufficiently concerned about the impact on issues like moral evaluation and responsibility for actions that we try another route to salvage causal efficacy in the face of apparent physical determinism: one could attempt some kind of reconciliation of 2 and 3, with the understanding that volition must be both intentional in character and causally efficacious in the world in order for agents to be held morally responsible for their actions. Such an attempt at reconciliation of 2 and 3 would overlap with the territory of the older debate of free will versus determinism: what kind of causal wiggle room can we find in a fully determined universe, or of what use is the wiggle room we get via quantum indeterminacy?

Attempts at reconciliation of 2 and 3 for conscious volition have broader implications for the relationship between higher and lower level causes, and in particular between 'mental' and 'physical' causal relations. The Causal Exclusion problem of Davidson and Kim captures one facet of this, as does the debate about the possibility of downward causation and emergent causation. Given that, on a number of understandings of 'level', conscious features of agency are at a higher level relative to physical processes in the brain, and given that causes at the lower level are sufficient to bring about their effects, there appears to be no causal work left to be done by causes at a higher level, including conscious agency (Kim 1993, Horgan 1989, Shoemaker

1999, Walter 2006, Wilson 2009). Compounding the difficulty is the vast number of ways in which a system can be broken down into levels so that comparisons of causation at different levels can take place: there are mereological levels, functional levels, first and second (and even higher) order properties, and so on (Craver 2007). The concept of emergence has been offered as a way to secure genuine causal powers for higher level features, and for conscious awareness, although there is very little consensus about what emergence is. Others have relied on characterizations of strong and weak supervenience to establish either that there could be such high level causal efficacy, or at least that we can't show there is no such efficacy. The goal of philosophical positions that appeal to emergence or higher level causes is often the retention of some kind of genuine causal oomph for conscious agency, preserving premise 3 at the same time as premise 2.

I am not claiming that all these discussions are explicitly pitched in terms of reactions to Libet's work. Nonetheless, much of the literature on this nexus of topics makes sense when viewed as different reactions to the potential consequences of Libet's work for our understanding of conscious volition. In particular, these positions can be understood in terms of which of the above premises describing the standard view of volitions they reject. At this point, we have a new understanding of the context into which Libet's work fits, and why it continues to strike us as relevant to scientifically informed philosophical discussions of conscious agency.

One upshot of the critique of Libet's background view of volition is that we have another avenue by which to respond to the problems generated by Libet's results, namely by jettisoning the problematic view of conscious agency that led to the tension in the first place. By choosing an alternative conception of conscious will, in large part because of the problems associated with the view operationalized by Libet, those experiments no longer have significance for

philosophical views of agency. A great deal of philosophical work has been put into the development of philosophical positions that can accomdate Libet's results as well as maintain the attribution of causal efficacy to conscious agency. But once these efforts are put into perspective via these four premises, we can see that there is a much more straightforward and effective resolution to the problem involving a different way of operationalizing conscious agency. I will provide such an alternative account in chapter 3.

Now that we've elucidated both why Libet has been such an important part of the philosophical discussion on volition for the last several decades, as well as uncovering the category mistakes and causal fallacies committed in his experiment, we can safely leave Libet's experiments behind while developing an alternative approach for the scientific investigation of the causal role of conscious awareness in action.

1.6 Metaphysical and empirical causal questions

It has turned out, as we proceeded through the three important steps made by Libet and the problems associated with his conception of volition, that the work to reach the conclusion he did was not performed by the experimental results themselves. Rather, it was assumptions about the manner in which volition could act, and the dualism implicit in treating conscious decisions as lacking neuronal instantiation, that were instrumental in reaching the conclusion that conscious volition is not a cause of actions like finger liftings. This point can be generalized to apply much more broadly. Arguments to the effect that conscious awareness is not causally involved in action (and many arguments that it is only tangentially or tenuously involved) rest primarily on assumptions about the kinds of causes one should allow, and the kinds of causal structures one

should allow, not on the details of scientific evidence, which cannot alone decide the case one way or another. Metaphysical assumptions about causation itself provide the foundation from which we can interpret the evidence provided by scientific research.

There are two ends to a spectrum of causal questions, as has already been briefly discussed earlier. These could also be thought of as two distinct axes, along which we could plot causal questions according to the degree of empirical or metaphysical considerations needed to answer them. The first dimension regards the more general question of causation itself: what does it mean to attribute a genuine causal capacity at all, and what does it mean to do so in the context of the extraordinarily complex and multiply hierarchical system that is the human organism? Can higher level or compound causes of any kind be considered genuinely causal, or are they merely epiphenomenal, or at most dependently causal, while their lower level instantiations or components do the actual causal work? These metaphysical causal questions apply to any higher-level causes, but especially to conscious awareness, probably the most complex higher-level cause one could argue for. Questions about the nature of causation itself, independent of any particular causal system, are metaphysical causal questions.

In order to answer metaphysical causal questions, we can't simply look to the world to see how things are. We may agree completely about the empirical facts and yet disagree about the metaphysical status of these facts. For instance, we may agree that it seems to us that our conscious decisions are causes of our behavior, and yet still disagree about whether or not conscious decisions are actually the kinds of things that could be causes, or if they merely appear to be because of their consistent connection with other, microphysical, 'genuine' causes. No amount of additional scientific investigation, unaided by metaphysical assumptions about the nature of causality and the corresponding evidentiary standards provided by such assumptions,

could resolve the question of whether it is the micro-components or the conscious decision pulling the causal weight.

Ideally, we should be able to derive a methodology from our metaphysics of causation. A methodology for how to find causes, and how to distinguish causes from noncauses, should arise out of a theory of what causation is. Metaphysical features of causation should provide the framework by which to develop techniques for finding such causes in the world; these techniques need not be entirely fool-proof, but should include justification, with recourse to the metaphysical theory of causation being assumed, of why the methodology will be sufficiently reliable to be at least useful. Causal methodology thus spans and connects the metaphysical and empirical dimensions or axes.

Along the second dimension, that of empirical causal questions, we find causal questions that can be answered by recourse to scientific results. Once we begin with some metaphysical view of causation, or (at the least) a methodology for finding causal structure in the world, we can pose empirical causal questions about the actual structure of conscious agency as we find it, and then proceed to answer those questions using scientific resources. Empirical causal questions, in order to be answered, must presuppose a set of metaphysical assumptions about the nature of causation, and by doing so start with a methodology of how to find causes. While the metaphysics of causation establishes and justifies such methodologies, empirical causal questions require the application of methodologies to actual systems. We could not meaningfully interpret the causal implications of scientific research without at least some kind of methodological view about causation – we need to have at least some sense of what a cause is before we can label some relationship causal.

These dimension form two distinguishable lines of inquiry needed to ascertain the causal structure of any system in the world. This distinction will provide a format for the dissertation: there are two basic and interconnected kinds of questions that can be asked about the role of conscious awareness in agency; two avenues of investigation to answer these questions; and two kinds of argumentation or evidence that can be provided by philosophers. The resources to address metaphysical questions about conscious agency will differ rather markedly from the resources needed to answer empirical causal questions about the same. Even though there may be no absolute line between a metaphysical causal question and its empirical counterpart, a great deal of clarity regarding the kind of answer we should expect can be gained by isolating out one aspect or the other. Experimental results can answer a lot of questions, when deployed in the appropriate context; they can obscure other questions when deployed instead of methodological or metaphysical considerations.

The remainder of this chapter will illustrate this distinction between metaphysical causal claims and empirical causal claims, plus the methodological commitments that bring the two into contact. I'll start with an example from neuroscience, voluntary versus involuntary attention, because it is a case in which metaphysical causal questions are easily confused with empirical causal questions, and clarifying that ambiguity clarifies our understanding of attention. I'll then move on to show how a number of prominent positions in the literature regarding conscious involvement in agency can be understood as focused primarily on empirical work, or on metaphysical issues – this distinction neatly classifies much of the debate. That sets up a framework for the major part of the dissertation: chapter 1 and 2 largely criticize the conflation of metaphysical assumptions with empirical results; in chapter 3, I demonstrate how there is substantive empirical evidence to support the claim that conscious awareness is causally

involved in actions; and in chapters 4 and 5, I take up the metaphysics of causation to show that higher level causes such as conscious awareness are legitimate causes.

1.7 Two kinds of causal questions applied to attention

I've already introduced a distinction between two dimensions of causal questions, and briefly described how theoretical commitments span the distance between these dimensions. My goal in this section is to illustrate the distinction with a concrete example where failing to sufficiently distinguish these two dimensions also leads to an avoidable confusion about the causal relationships involved; after this specific example, I will provide some more general ones, and by the end of the section it should be clear how this distinction plays out, and provides a clearer understanding of how to answer the question of the causal role of conscious awareness in agency.

The case of voluntary versus involuntary control of attention, also referred to as top-down versus bottom-up attention, is interesting because there are (at least) four distinct senses of top-down and bottom-up causal influence involved in the ordinary usage of these terms. Using four distinct means of decomposing the attentional control system in the human brain, there are four corresponding senses in which the 'top' could influence the 'bottom' and vice versa. What leads to confusion is that even though there are four decompositions into levels, and four mechanisms by which causal influence may be propagated, the causal influence propagates in the same direction for all four (i.e., for top-down it goes from the top to the bottom, for all four senses of top and bottom, but with different implications for causal structure for each of the four instances of top-down causal influence). It is thus extremely easy to equivocate between the distinct senses

of top-down or bottom-up causal influence. By teasing out four different ways of decomposing the same hierarchical complex system into levels, and how causal influence propagates in each of these four cases, we can see how these decompositions naturally fall into two categories. Three of the four decompositions involve straightforwardly empirical causal questions about how one level influences the other. The fourth, however, involves metaphysical assumptions about high-level causes and whether or not they are genuinely causal; the answers one provides to this question will determine whether or not one attributes genuine top-down causation to this fourth means of decomposing attention.

Voluntary and involuntary attention are two broad categories of attention as differentiated by the kind of control subjects have over them. Involuntary attention is what the name suggests: the kind of attention that cannot be voluntarily controlled, but which is instead involuntarily captured by salient parts of the environment. A sudden loud noise and bright light flash in our peripheral vision capture our attention, whether or not we decided to pay attention to them. Voluntary attention, on the other hand, can be consciously and deliberately directed towards features in the environment, or away from them. We might decide to attend to a small visual detail that would otherwise make little impact on our notice; or, we might decide to ignore the flashing light after it has startled us. Involuntary attention is also sometimes called exogenous, because the stimulus is external to the subject, and voluntary attention called endogenous, because the impetus is internal to the subject. Neuroscience has a good understanding of the mechanisms involved in each of these two kinds of attention (Banich 2004).

Voluntary attention in particular is characterized by what is referred to as “top-down” control of attention, while involuntary attention is characterized by “bottom-up” control. This notion of control is essentially causal: it indicates the directionality of causal influence between

two levels. There are multiple different senses, however, in which the top could influence the bottom or vice versa, depending on how one specifically defines the levels in questions. There are four senses of top-down and bottom-up causal influence that I'll elucidate here to illustrate the difference between empirical and metaphysical questions about such interlevel causation. Each sense of top and bottom corresponds to a different way of individuating levels into top and bottom, and thus each level breakdown will correspond to a somewhat different mechanism of causal influence connecting the top and bottom levels. The first such level breakdown is according to temporal order in a chain of processing, where the earlier stages in such a chain are lower or 'bottom' with respect to later stages in the chain. The second level breakdown is functional, regarding complexity of signal and receptive field size; lower levels for attention to features of the visual field, for instance, involve simpler signals processed by neurons in V1 with a very small receptive field (such as signals indicating only the presence or absence of something in a very small portion of the entire visual field), while higher levels of processing of visual information involve much more complicated signals, such as edge or face detection, by neurons that have a much larger receptive field. The third level breakdown involves the difference between higher and lower order cognitive abilities: executive control, in the frontal cortex, is considered a higher cognitive ability than those associated with, for instance, homeostatic regulatory mechanisms in the brainstem. The fourth and final such level differentiation involves conscious awareness, as it controls voluntary attention, versus nonconscious neuronal processes – a consciously made decision to attend to some feature of the visual field will alter the way in which specific neurons fire.[13]

[13] One could also differentiate based on anatomical location, but I do not include this as a potential level decomposition for two reasons: it is not itself hierarchical; and it is sufficiently

The first three of these level breakdowns (temporal, functional, and cognitive abilities), I claim, involve straightforwardly empirical causal questions about the way in which causal influence is transmitted. They are also, for this reason, level differentiations for which we can give an account of the nature of the relationship between the higher and lower, or top and bottom, levels. The manner in which lower functional levels transmit causal influence to higher ones, and vice versa, can be expressed in terms of specific mechanisms by which influence is propagated. The fourth sense of levels, conscious control of attention and neuronal processes, is really about how causation figures in the relation between neural processing and the associated conscious awareness. This fourth sense concerns the possibility of downwards causation between levels, and whether the relationship from the bottom to the top could be causal, or if it must be some other relationship instead (such as supervenience, identity, instantiation, etc). Each of these senses will now be elaborated more fully.

Voluntary versus involuntary control of attention, top-down and bottom-up respectively, is most often defined in terms of temporal order of signal processing in the brain, and the functional differences that correspond to these temporal orderings. For ease, I will utilize primarily visual examples, where control of attention is not identical with but closely tracks eye movement or saccades. Visual input to the eyes projects from the retinal area to the primary visual cortex at the back of the brain, an area called V1. V1 projects to areas V2, V3, and onwards towards the prefrontal cortex. In the temporal sense, the earlier parts of this chain of projection, especially, say, V1, are the ‘bottom’, while later stages are the ‘top.’

---

covered by the first three decompositions. These four are not intended to be exhaustive, but merely sufficient to illustrate the distinction between empirical and metaphysical causal questions with respect to voluntary and involuntary attention.

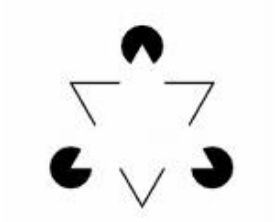
Bottom-up attention occurs when something visually striking causes a saccade to that location: a bright flash of sudden movement in the peripheral vision stimulates a sudden involuntary saccade. By so saccading, the organism's general attention is brought to the startling object (presumably a very useful tactic for avoiding unpleasantness). The eye movements are controlled, in the case of involuntary attention, by the superior colliculus. In contrast, top-down attentional control occurs under a different set of circumstances and relies on a different part of the brain to guide saccades, the frontal eye fields in the prefrontal cortex (PFC). Visual signals project from V1 onwards into the PFC; there are also backwards projecting signals that return from PFC to lower or earlier areas of visual processing. When this occurs, later areas of processing can subsequently affect 'earlier' ones, once the first signal has reached the 'higher' areas and had time to return through backwards projection. Simple location in the order of processing, then, allows areas in the PFC to be considered as at a higher or top level with regards to areas like V1 that are then at a lower or bottom level, in the sense of temporal ordering in a chain of processing. In this temporally order sense, the top affects the bottom when signals propagate form the PFC back towards V1, and the bottom affects the top when signals propagate from V1 onwards towards the PFC.

The second sense of top-down and bottom-up results in the same areas of the brain being labeled 'top' and bottom', but now these areas are ordered by complexity of signal processed rather than temporal order of processing. V1 and other earlier areas are also, functionally speaking, lower with regards to later areas. Each new area is capable of culling very specific information from the signal it receives, and of culling it from a different receptive field. Regions that are more distal from the retina, further along the temporal chain of projection, receive a processed signal carrying different information than what reaches V1. Cells earlier in the chain

have relatively small receptive fields (the areas of the visual field to which they react), and simple stimuli to which they react, such as edges and contrast. Each projection, which forms a column, changes both of these features: the receptive field of neurons gets larger, and the complexity of their preferred stimulus increases. Because higher-order features of vision are not processed until further stages, and the full information from the visual stream is utilized in prefrontal areas, the order in which areas receive visual information from the retina corresponds to the degree of complexity of objects – later stages of processing involve more complexity in the signal and in the objects to which the neurons are capable of responding. Thus, the 'top' areas are functionally dependent on the lower, and functionally more complex.

Ordinarily, the order in which subsequent areas process bottom-up projections is what gives rise to or allows the increase in complexity of signal processing. Each subsequent cell receives information from multiple earlier cells, combining the information from their receptive fields into a larger receptive field capable of discriminating more complex and larger objects. Sometimes, however, the backwards projections from PFC back to V1 also have a top-down influence on the behavior of the 'lower level' cells. A classic example of this is the Kanisza triangle. Lines and pie-pieces are arranged so that there appears to be one triangle occluded by another otherwise invisible triangle:

Figure 3: Kanisza triangle



We easily see the invisible or white triangle – the lines are automatically filled in by our visual cortices, even though they do not exist. The seeing of a triangle, a large-scale abstract

object that fills a large chunk of the visual field, has a top-down effect on neurons in V1. Recall that the size of the receptive field of neurons increase further along the chain. A given V1 cell may have a receptive field that is where the line would be, if there were a line; since there is no such line the receptive field is actually empty. This cell initially does not fire, because it has nothing in its receptive field. However, after the signal has reached PFC and is projected backwards again, that cell will fire as if there actually were a line in its receptive field, even though the receptive field is still blank. The functionally higher levels have 'told' the lower levels, as it were, that there is a triangle that should be noticed. The second sense of top and bottom, then, concerns complexity of processing, and is characterizable as a functionally hierarchical parsing of top and bottom.

Importantly, the 'top' level in the temporal sense corresponds to the later parts of the chain of projection, further towards the PFC, which is also the 'top' in the functional sense. The 'bottom' level in the first sense, especially V1, is also the 'bottom' in the functional sense. These two senses of top-down and bottom-up are closely related: the increase in functional complexity occurs because the ordering of the chain of projection. The same is true of the third sense of top-down and bottom-up attentional control, namely higher and lower cognitive abilities. The 'top' area, the prefrontal cortex, is also the area associated with higher cognitive abilities such as executive control, reasoning, abstract thought, and so on. The earlier, 'bottom', portions of the chain are associated with lower cognitive abilities. The areas designated by the higher and lower levels in this third sense correspond to the higher and lower levels in the first two senses, even though the levels are divided according to different criteria. Thus, the causal mechanisms by which influence propagates from lower to higher and higher are connected for each of these three level differentiations: the temporally ordered projection of signals from V1 to PFC allows for an

increase in the complexity of signal processing, which also allows for the higher cognitive capacities associated with PFC as opposed to V1.

Notice that so far, all the causal influence can be described in a straightforwardly mechanistic way. We have a good empirical understanding of the influence of neurons on other neurons, one brain area on another brain area. As we've seen, voluntary control of attention proceeds through a different part of the brain than involuntary control – the frontal eye fields and superior colliculus respectively. The means of causal influence of this control is reasonably well-understood. The same is true of the top-down and bottom-up influences along the chain of projections from V1 to PFC and back to V1 again. Nothing in the first three senses of top and bottom levels has been causally problematic – insofar as one is willing to believe there is causation, we can account for the causal relationships between the higher and lower levels.

The fourth sense of top-down and bottom-up, however, involves a different sort of question about the nature of the interlevel causal influence. In this fourth sense, the top level is conscious awareness, and the conscious control that can be exerted over attention, as well as the conscious perceptual awareness of what is attended to. The bottom level is that of neuronal processing, potentially including both the neuronal processes which subserve voluntary attention, as well as the processes that subserve involuntary attention (one could consdier all neuronal processing as lower level with respect to conscious awareness, or one could single out the neuronal processes associated with involuntary attention as lower level with respect to awareness). Bottom-up control in this fourth sense would refer to the influence that the nonconscious control of eye saccades has on conscious awareness: we become aware of what we automatically saccade to, even though we did not consciously decide to do so. It could also refer to the relationship between the neuronal processes associated with voluntary attention, such as

the connections between V1 and the frontal eye fields, and with the conscious awareness of the voluntarily-attended-to stimuli. Top-down causal influence could refer to influence that conscious awareness would have on neuronal processes when we decide to deliberately turn our attention to something. This might be either the influence of conscious control of attention on the pathways going from the frontal eye fields to V1; it could also indicate the influence of conscious awareness of attended-to stimuli on the neuronal processes associated with involuntary attention. These are markedly different kinds of causal relationships (if, indeed, they are all causal – the relationship between conscious awareness and its control of attention and the neuronal processes associated with voluntary attention is not, I've argued and will argue further elsewhere, a *causal* relationship) than we found in the first three level differentiations. One could hold that there really is causation in the world, and that the first three senses of levels involve meaningful causal relationships, while still being skeptical that the relationship between conscious awareness and neuronal processes is genuinely causal.

Thus, this fourth differentiation of top and bottom turns on metaphysical considerations about the nature, or even existence, of the causal influence between top and bottom, considerations that were not present in the first three senses of top/bottom. This applies to both the possibility of top-down as well as bottom-up control. Let's start with the easier case of bottom-up attention. Automatic neuronal processes cause eye saccades to salient features of the environment, and we become aware of this feature that captured attention. But is this a *causal* influence, or is there some other kind of relationship between conscious awareness and neuronal processing? If awareness supervenes on any of these same neuronal processes, then it would be inaccurate to say that one caused the other. Rather, we might say that the bottom gave rise to, or instantiated, or is a correlate of, the top. All of these locutions have been used to describe this

relationship, and it is far from clear that, according to this sense of top and bottom, the bottom could causally influence the top directly.[14]

The other direction fares no better. If there is controversy about whether or not the bottom could causally influence the top, it is overshadowed by the controversy surrounding the question or whether or not the top level could affect the bottom – if conscious awareness is the sort of thing that could cause anything neuronal to occur. The problem of causal exclusion, introduced by Davidson and Kim and widely debated in the last two decades, addresses just this issue. These controversies highlight the deeply metaphysical nature of the question about causal influence between awareness and neuronal processes. In particular, the discussions on this topic is a strong indicator of the fact that we cannot simply go and test conscious awareness and neuronal processing to ascertain what the causal relationships between those two levels are, in the way that we can simply go test the causal relationships between V1 and the PFC. We must, in order to answer this last question, also address an explicitly metaphysical dimension. Is conscious awareness capable of causing anything to occur, or does its apparent causal efficacy depend completely on the causal efficacy of the neuronal processes on which it supervenes?

This example of voluntary and involuntary attention, and the contrast between ways of parsing higher and lower levels, are intended to highlight the difference between causal questions that are primarily empirical – the first three senses of top and bottom – and causal questions which are heavily metaphysical – the fourth sense of top and bottom. The ways in which we answer these questions differ, and the tools at our disposal for answering them differ. The

[14] For a more detailed presentation of this argument, see the earlier section in this chapter regarding the implicit dualism Libet is committed to by dint of his choice of variables; there is also an extensive discussion of this point in chapter 4.

importance of being clear about primarily metaphysical versus primarily empirical causal questions extends beyond just the case of attention.

When we answer empirical causal questions, we utilize a set of methods, looking for specific causes and effects, specific relations in the world, taking for granted that there are such things. We have already committed ourselves to some view on the nature of causation, even if it is a vague or broad one (many ontological views of causation may be compatible with a single causal methodology). Even though it may not be a specific view, it must at least justify these tools as the appropriate epistemic approach to the system under consideration. Further, when we utilize some methodology for sussing out causal structure in systems like that of visual attention, we are not asking deeper ontological questions about the nature of those causes. We assume it means something to label a relation causal, that there is some difference in the world being picked out by the label; we are *using* the notion of causation rather than defining or investigating it. While most methodologies implicitly constrain the sorts of relations or variables that can count as causes or effects, and thereby have something that can count as an ontology of causation (at least, an extension for an ontology), one can remain deliberately agnostic about what precisely it means to call something causal, while at the same time looking for actual instances of causation.

Conversely, when we are addressing metaphysical questions about causation or building the foundations of a theoretical framework, we don't do so by looking at specific systems like visual attention and trying to ascertain their causal structure. The way in which we utilize specific instances of causation in answering metaphysical causal questions is by choosing examples where we already have an intuition about whether or not something should be considered causal. We test accounts of what causation is by seeing if these accounts correctly

label the causal and noncausal relations in accordance with intuition. The investigative aims of the methodology and metaphysics of causation differ; the roles of specific systems in the world for answering these questions also differ.

I will eventually show that there is no genuinely empirical evidence to the effect that conscious awareness is causally epiphenomenal in action. Arguments to the contrary primarily rely, I claim, on metaphysical assumptions about what is or is not allowed to count as a cause, and these assumptions are, once examined, unfounded. The empirical evidence offered in support of such claims does not do the argumentative work of reaching the conclusion, which rather rests on metaphysical assumptions about what kinds of causes are permissible. From an empirical perspective, there already exists a wealth of evidence to the effect that specific features of conscious agency are involved in specific ways in bringing about conscious action, if we rely on a causal methodology that is independently justified and involves better assumptions about the metaphysics of causation. This is true even though there is also evidence that conscious awareness is sometimes less involved in action than we may have naively assumed. Cognitive science and psychology simply don't support the *empirical* claim that conscious features of agency are, in any broad or significant sense, causally inefficacious.

Untangling these two kinds of causal questions has a number of philosophical benefits. First, I've argued that broad empirical claims of conscious causal inefficacy involve causal mistakes, contentious metaphysical assumptions about causation, or are in conflict with empirical evidence. This means, though, that we can avoid committing mistakes in our causal analyses of conscious agency by avoiding broad claims of inefficacy. We should instead focus on claims with smaller scopes of generalizability that are richer in specifics, claims that flesh out particular causal roles played by conscious features of agency in a variety of circumstances, be

those roles greater or smaller than we had anticipated. Second, we already have a good sense of the methodologies that are useful for distilling causal structure from data, and there are already such techniques being used in the sciences. The interventionist approach to causation, as developed by philosophers such as Woodward, Spirtes, Glymour, Scheines, and Pearl (see chapter 3) is independently well-justified as a methodology for finding causal structure in the world. We can co-opt the analytical tools of interventionism as a means of answering empirical causal questions about the causal structure of conscious agency.

The point of the second half of this chapter can be summarized as follows: if we conflate empirical causal questions with questions about metaphysical causal commitments, we will make confused and inaccurate claims about the role of conscious awareness in agency. In order to interpret the significance of experimental results for agency, we need a set of theoretical assumptions about causal structure and causal relations, and we need to be explicit about the role those assumptions about causation play in reaching conclusions about causal structures in specific systems. We saw this in the case of Libet's view of volition: in order to gauge the impact of his results for conscious volition, he needed to assume a specific view of volition, and he assumed it was a reified extra event in the causal chain. In turn, we saw that a different assumption about how volition related to that causal chain led to different conclusions from Libet's same evidence. If we respect the distinction between different kinds of causal questions about conscious awareness' involvement in action, then we can make interesting claims about empirically ascertained causal structure, and how that empirical structure fits into a metaphysics of causation for complex hierarchical systems, and how our metaphysical understanding of complexity and causation then informs our understanding of the causal claims made in the sciences regarding agency.

If the best available methods of causal investigation demonstrate that some variable involving conscious awareness has a causal effect on some other variable involving overt behavior, then in one important regard there is nothing more to showing that the former is causally involved. If our best methods of causal analysis say that something is a cause, then empirically speaking, it is. Asking for some further justification is akin to stomping one's foot, or hitting the table with a fist: "Yes, but is it *really* causal?!" In this kind of question, the 'really' is intended to get at something more ontologically fundamental than what science can provide (whether or not there is such a thing). The 'really' emerges from the worry that it looks as if conscious agency were causally efficacious, but perhaps we should not call what conscious agency does causal, because (for instance) what looks like conscious awareness causing something is actually some other physical cause, and we failed to incorporate that other physical cause into our research. A 'really' worry is that we haven't yet conditionalized on the right variable, but if we do the apparent causal role of conscious awareness will evaporate. Or, perhaps because conscious awareness has such an intimate link with the physical processes associated with movement, we should attribute the causal oomph to the physical processes and think of awareness as just along for the ride.

Once we agree to a causal methodology, there is nothing more, empirically speaking, to cashing out the causal efficacy of conscious awareness than what we find in the sciences that investigate awareness and agency. If we are asking for more than that, we aren't asking a question the scientists can answer anymore. We've already gotten their answer; further questions about where the 'real' causal oomph of conscious agency need to turn to the metaphysics of causation, not empirical applications of it.

CHAPTER 2: The internalization of conscious agency

Research in cognitive science in the last two decades has led a number of philosophers to conclude that as agents, we have little to no conscious causal influence on our own actions. These views range from the mild, that conscious awareness has a constrained role to play in action, to the very strong, that awareness has no causal role whatsoever and that our strong impressions to the contrary are simply persistent illusions. In my view, these conclusions are symptoms of a problematic view about the possible roles that awareness could play in conscious action, rather than the inevitable result of the research itself. My focus in this chapter is on the model of conscious agency that is often used when philosophers incorporate scientific results into our philosophical notions of agency, and I will show how holding such a view of agency will almost always lead one to conclude that awareness is causally inefficacious in action, regardless of what the research demonstrates. This problematic view of agency is predicated on the assumption that the *awareness of* agency (alternatively, an experience, sense, or phenomenology of agency) is the appropriate experimental measure for understanding the contribution of awareness to agency.

My argument is not that the awareness of agency holds nothing of interest for the philosopher of action; indeed, it should be part of the research to which we look for an empirically informed view of agency. Rather, the underlying problem I address here is a tendency in the literature to focus almost completely on the awareness of agency, to the

exclusion of other kinds of research involving awareness in action. This trend of concentrated focus on the awareness of agency leads to systematic distortion in the resulting views of agency.

A number of authors utilize the sense or awareness of agency as a method for investigating the extent of the conscious causal contributions to action (Metzinger 2006; Velmans 2004; O'Shaughnessy 2003; Carruthers 2007; Nahmias 2005; Choudhury and Blakemore 2006; Wegner and Wheatley 1999; Haggard 2003). This trend includes both the assumption that conscious awareness is involved in action as awareness *of* acting, and, as a consequence, the assumption that introspective reports by agents should reveal awareness of the mechanisms involved in the exercise of agency, if the agent was consciously involved in that exercise (Mossel 2005; Bayne and Levy 2006; Horgan, Tienson, and Graham 2003).[15]

The focus on the awareness of agency and assumption that introspective reports should reveal the mechanisms of action together constitute a flawed model of agency that I will call the Micromanagement Model (henceforth MM). On this model, awareness is inwardly directed at distinct agentive processes leading to action, and in order to exert causal influence on action, awareness must somehow "fiddle with the knobs" of those other agentive processes. Holding awareness to such a standard in order to count as causally efficacious in action makes it appear appropriate to conclude that our conscious selves play little to no causal role in producing our actions, even though the empirical evidence considered more broadly does not support such a conclusion.

---

[15] There are philosophers (*inter alia*, Chalmers 1996, Newell 1992) who take there to be an important difference between awareness and conscious awareness, whereas I do not distinguish between these (and specifically avoid use of the term 'consciousness'). For reasons of space, I will not make a philosophical defense of my usage of these terms, beyond noting that, in the scientific literature on which I rely, awareness and conscious awareness are used interchangeably. As my goal is provide an empirically grounded account of conscious agency, the extant science should guide usage of these terms.

The focus on the sense of agency has three specific shortcomings to be explored in this chapter. First and most importantly, it directs the causal influence of awareness inward, towards our own internal processes leading to movement. This internal re-direction separates awareness from agency and allows awareness to be causally efficacious only insofar as it affects other internal processes, requiring micromanagement of internal processes. Second, we have reason to think that the neuronal processes leading to movement cause both the movement and the sense of agency itself. This means that the awareness of agency may usefully serve as an indicator of the exercise of agency, since it is a secondary effect of agentive processes, but that it does not contribute to our understanding of the causal role of awareness in agency. Finally, the expectation that subjects be able to provide introspective reports on how they acted, or regarding all other causal influences on their behavior, is a causal double standard. It holds awareness to a higher standard of causal efficacy than other kinds of causes, because we ordinarily do not balk at attributing causal efficacy to one factor even though there may also be other factors that were causally influential on the same effect.

My claim is neither that there is nothing worthwhile to be found by looking at the sense of agency, nor that the authors who discuss it would explicitly endorse what I call the Micromanagement Model. Rather, my argument is that the level to which the discussion of conscious agency has focused on the introspectively-directed sense of agency obscures the causal role of world-directed awareness in action. This focus has thus excluded some otherwise relevant experimental literature from being properly considered for the light it can shed on the causal structure of conscious agency.[16]

---

[16] See chapter 3 for an elaboration of the causal contributions of these externally directed avenues of awareness.

After examining the way in which the assumptions that constitute the MM emerge in scientifically oriented views of conscious agency, I'll take a brief look at a view of conscious agency that is motivated by purely philosophical considerations. I demonstrate how Jennifer Hornsby's view of tryings (1980) internalizes the causal role of conscious agency in a way strikingly similar to that of the Micromanagement Model. While Hornsby herself goes on in subsequent work to develop a view of agency that does not involve such internalization, there are a number of other philosophers who have advocated similar views of tryings (Pietroski 2000, O'Shaughnessy 1973, Zhu 2004). This part of my discussion demonstrates how a single assumption can be found running through distinct views of agency that were generated in response to very different motivations, leading to similar problems for representing the causal structure of conscious agency.

My overall conclusion is meta-methodological: while there are interesting things to be learned by looking at how internally directed elements of awareness figure in action, the extent to which attention has been focused on the awareness of agency has obscured the other avenues of conscious causal involvement in action, and consequently distorted our view of the causal structure of conscious agency.

2.1 The awareness of agency: Metzinger and Gallagher

I will first characterize two assumptions that go hand in hand regarding the use of the awareness of agency, and then demonstrate how they figure in the work of two prominent contemporary philosophers, Thomas Metzinger and Shaun Gallagher, as examples of the larger trend discussed in the previous section. Metzinger and Gallagher espouse different views of agency that are

motivated by distinct considerations, yet both utilize the awareness of agency as the general measure for the involvement of awareness in agency and consequently both end up with similarly internalized views of agency.

The first assumption concerns the way in which conscious agency is implemented in our actions, and how the sciences operationalize it: experimental results concerning the awareness of agency (alternatively, the sense or phenomenology of agency) have direct implications for the causal contributions of awareness to agency. The basic idea is that we can learn about the causal structure of conscious agency – the kinds of causal contributions made by awareness to agency, and the contexts in which awareness is or is not causally efficacious in action – by examining several kinds of cognitive scientific research. Such research includes the contexts in which we experience, or fail to experience, the sense of agency (Haggard, Clark, and Kalogeras 2002). It also includes the contexts in which the awareness of agency comes apart from action, namely, where intentional action is present but the sense of agency is not, or where the sense of agency can be induced even though we aren't acting (Sata and Yasuda 2005). This body of research can also include the relationships between the neuronal processes involved in the production of action and the neuronal processes that are associated with the sense of agency (Cunnington, Windischberger, Deecke, and Moser 2002; Farrer and Frith 2002). These factors are elements of the awareness of agency. The first assumption is that these factors also help us understand the specifically causal contributions of awareness to action.

The second assumption is that subjects should have introspective access to the mechanisms by which they act, or to some (or even all) other causal influences on their actions, if they are to be credited with conscious action. The idea is that insofar as we are consciously involved in our own actions, we ought to be aware of various features of those actions, and thus

be able to provide introspective reports on those features. For instance, we might be expected to produce introspective reports on details about the way in which we exercised our volition (Frith 2002). We may also be expected to be aware of other causal influences on our action, such as features of the environment that influence our action (Nisbett and Wilson 1977; Greenwald and Banaji 1995). This second assumption makes explicit an implication of the first assumption for the study of agency, namely that we ought to attribute a lack of conscious causal involvement in actions in cases where we can demonstrate that the subject was unaware of some such feature of her own action or of some causal influence on her action. Unless one can report on other causal influences on action, the thinking goes, one wasn't consciously acting. If one accepts the view that the relevant kind of awareness is awareness of one's own agency, then the intentional content of that awareness should accurately reflect the distinct agentive processes at which it is directed. Insofar as subjects are unable to provide evidence of awareness of a given agentive process or causal influence on action, subjects were not consciously involved in that action.

I'll now take a closer look at how each of these assumptions play out in the work of two representative authors, Thomas Metzinger and Shaun Gallagher. In the next section, I will demonstrate how these two assumptions, taken together, lead to what I'll call the Micromanagement Model of agency. Metzinger and Gallagher are concerned with developing our understanding of conscious agency or conscious volition, and they both offer the sense or phenomenology of agency as the means for doing so. These two authors are useful examples because they do make substantial progress on understanding some features of conscious volition, and offer otherwise interesting views of volition. While neither would subscribe, I think, to the Micromanagement Model when it is explicitly described, their work leads in that direction, precisely because of their focus on the sense or phenomenology of agency.

Thomas Metzinger (2002, 2006) develops a scientifically informed, phenomenological and representationalist model where conscious volition is fundamentally internal. He identifies two central projects for theories of volition. The first is "*describing the phenomenology of will* more precisely" (2006), and the second is connecting the phenomenological structure of will to its neural correlates in order to explain that structure.

He notes two features of experienced will that render this task difficult. This first is that the phenomenology of will is *thin*: it is not particularly crisp, vivid, or distinct, in the way that the phenomenology of color, for instance, is crisp, vivid and distinct. The second is that it is *evasive*: the act of introspecting for our own experience of will sometimes seems to make the object of introspection fade or dissolve or change. These features of the phenomenology of volition render the task of developing a full theory of the causal structure of conscious agency much more difficult, according to Metzinger.

> What we can do is to conceptually describe the contents of experience in a way that, at least in principle, allows us to tie specific phenomenal properties to specific causal roles realized by certain parts and dynamical subcomponents of the brain, thereby generating testable hypotheses. (2006, 21)

The problem as he sees it is that we need first to have a sufficiently full understanding of the phenomenology of agency before we can turn to explaining that phenomenology. This constitutes the task of modeling the structure of conscious agency. He describes the task at hand more specifically:

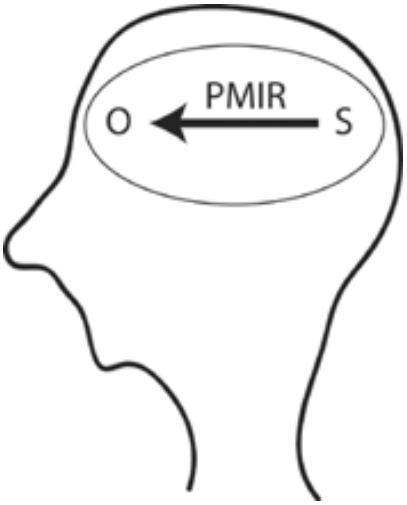
> Here, I am only concerned with the structure of phenomenal volition. The two central issues are these: How can a system represent itself *to* itself as 'having an intention,' and how can we understand **the special causal linkage between an active goal representation and the overt motor output**? I will propose that both are simultaneously mediated through a single and specific representational structure, namely, the 'phenomenal model of the intentionality relation' (PMIR). (2006, 22; emphasis added)

The PMIR is a phenomenological representation of the self as engaged in a volitional act which is intentionally directed at some object, always some kind of goal component. "Phenomenologically, a PMIR typically creates the experience of a self in the act of knowing, of a self in the act of perceiving—or of a *willing* self in the act of intending and acting" (ibid., 23). To properly understand the volitional PMIR, Metzinger contrasts it with an ordinary intentional relationship. What we ordinarily think of as agency *simpliciter* is an intentional relationship between the subject and something out in the world: something on or towards which we are acting. He calls this "good old fashioned intentionality", with which we are all familiar. The PMIR is not such an ordinary intentional relationship. Instead, it is a representation in phenomenological awareness of such an intentional relation between the subject and the world. The PMIR never reaches out into the world; it is an entirely internal representation by the self to the self of what would be an outwardly directed intentional relation. The intentional relationship itself is the content represented by the phenomenal model, and thus Metzinger holds that conscious volition is a representing of ourselves to ourselves, as agents.

Metzinger emphasizes the difference between an ordinary intentional relationship, directed into the world, and the representation of an intentional relationship which constitutes conscious volition, in the graphical depiction of the PMIR below. There is an 'old fashioned' intentional relationship, that between the self and some goal component in the world. This relationship, however, is not itself volitional, in the sense of conscious volition. The self represents itself as having an intentional relationship of this form, and it is this self-representation (not the intentional relationship itself) that is the explanandum Metzinger thinks we should be seeking; it is this character of conscious volition that our theories should accommodate. "My central point is that we do not only represent but that we also corepresent the

representational relation itself—and that this fact is relevant for understanding the architecture of volition" (ibid., 24).

Figure 4: Phenomenal Model of the Intentionality Relation



Note how the arrow representing the intentional relationship does not actually reach into the world (diagram is Metzinger's)

Recall the earlier passage where Metzinger said that "the special causal linkage between an active goal representation and the overt motor output [is]… mediated through a single and specific representational structure, namely, the 'phenomenal model of the intentionality relation' (PMIR)" (2006, 22). The primary point to take away is that conscious volition has been rendered internal, even introspective. Metzinger's PMIR analysis of conscious volition is essentially internally directed, a representation not of something in the world but of the self as engaged in a certain kind of relationship. Conscious volition is modeled as taking what many would call volition proper as its content, rather than volition being constituted by what this content depicts.

A second important thing to note is that the outwardly directed elements of volition have either been folded into the internally directed ones or else rendered non-conscious in some fashion. There are two ways of understanding Metzinger on this point, both of which lead to a similar result for conscious volition. On the first understanding of his view, awareness is

involved via a self-representation-to-self, and so any outwardly directed elements of awareness that could potentially be involved in agency are simply incorporated into the internal representation. “Whenever [the brain] catches itself in the act, whenever it corepresents the representational relation with the help of a PMIR, then it also generates the consciously experienced ‘arrow of intentionality,’ paradigmatically experienced in having the feeling of ‘projecting’ visual attention outward, as it were, or in attentionally ‘tracking’ objects in the environment” (ibid., 26). Insofar as awareness might actually have some avenue of action involvement that is straightforwardly world-directed, it must be brought inside the self-representation-to-self, with the effect that it is no longer a genuinely externally directed arrow of intentionality. On the second way of understanding Metzinger, the PMIR seems to imply that insofar as there are ordinary intentional relationships involved in agency, these are not conscious. Ordinary intentional relationships are the objects of conscious volition but, because they are not themselves allowed to be part of conscious agency, must be unconscious themselves.

The upshot is that, however one understands Metzinger with respect to outwardly directed elements of agency, volition is never both directed into the world and conscious; it can be at most one or the other. Either way, he is committed to the first assumption, that conscious volition is internally directed at other agentive processes.

Shaun Gallagher is another example of an empirically-informed philosopher who develops a view of conscious agency that ends up internalizing the contributions of awareness. Gallagher is a good second example because he develops his view to accommodate a different set of scientific results than does Metzinger, and takes a much less reductionistic view than Metzinger, but still relies on the awareness of agency as the concrete way to measure the causal role of awareness in action (Gallagher 2007). His aim “is to investigate both the phenomenology

and science of agency. In its proper sense, I understand agency to depend on the agent's consciousness of agency" (2007). On Gallagher's view, agency depends on the consciousness of agency in a strong way: it is a precondition of attributing conscious agency to a subject that the subject have a particular kind of sense of agency associated with the action in question.

Gallagher emphasizes the role of the sense of agency in action by developing a distinction between the sense of ownership of action, that it is oneself which moves or thinks, and the sense of agency, that one is the originator or source of this movement or thinking. The sense of agency involves the feeling that one is the causal source of an action. Gallagher reaches this distinction by defending the immunity principle, the principle that we are immune to making identification errors that involve self-attributions using the first person pronoun (Gallagher 2000), for attributions of thought as well as attributions of action. For instance, the immunity principle would preclude someone from both thinking, "I think it is going to rain," while also thinking, "But maybe it isn't me who thinks it is going to rain." A large part of Gallagher's motivation in this article is saving this immunity principle from apparent refutation by schizophrenic patients, who do *prima facie* appear able to think such thoughts: to have some thought, the content of which they report from a first-person perspective, while not necessarily thinking that this thought is appropriately attributable to themselves – some outside person or force might be thinking the thought instead.

Schizophrenic patients appear to make such mistakes in action attribution as well as thought attribution. They able to both act as agents and mistakenly attribute agency for their own and others' actions. They may overattribute agency to themselves, thinking that they are able to control others' actions, or underattribute, thinking that someone else is controlling their own actions. This threatens the immunity principle, which many have thought to be tautologically

true. Gallagher proposes the distinction between ownership of action and sense of agency to restore the tautological nature of the immunity principle by showing how schizophrenic subjects don't actually violate it. According to this distinction, patients in such cases have a sense of ownership of action – it is their body, and not someone else's, that moves. But they fail to have the sense of agency – that they are the initiators of the movement. Gallagher's reason for asserting this last point is that schizophrenic subjects lack the right kind of experiential basis to legitimately claim that they experience the sense of agency for another's movements that they believe they control. He thinks that schizophrenic subjects could not have such a sense for another's movements because the sense of agency has been shown (in healthy subjects, one should note) to be a product of a number of processes, including afferent feedback from one's body and copies of efferent signals sent to muscles. Schizophrenic patients wouldn't have the sense of agency for another's movement based on such processes.[17]

Thus, if we require that the sense of agency, not merely ownership of action, be present in order to count someone as consciously causally contributing to some action, we can maintain the immunity principle for action attribution without schizophrenic patients posing a counterexample. Gallagher's position on this seems to involve a double conditional: one was

---

[17] I think this is a substantive and probably false empirical claim. The very problem in overattribution of agency is that subjects cannot, from their first-person experience, tell the difference between the feeling they get from generating their own movements and the feeling they get from another's movements. Whether or not these feelings of control are based on the same underlying neuronal processes is less relevant, if they are phenomenologically identical. Similarly, it is possible that part of what is wrong in such cases is that watching another's actions does invoke similar neuronal processes in the schizophrenic subject in a way that does not occur for normal subjects. As such, they would have the same experiential basis for thinking they controlled another's actions as for their own. However, this is not the focus of my argument here; whether or not one thinks Gallagher is correct in his assertion that schizophrenics lack the right experiential basis, he does use this assertion as a step in arguing for the need to tie the causal contributions of awareness to agency to the sense of agency specifically.

consciously causally involved if and only if one had the sense of agency; if one failed to have the sense of agency, then one was not consciously causally involved.

This connection between the sense of agency and the immunity principle draws the exercise of agency into extremely close contact with the ability to provide introspective reports of the experience of acting. Gallagher is using the sense of agency as a measure of or indicator for conscious agency generally. Gallagher (2007) states that it is the sense of agency that gives us insight into the causal contribution of awareness to agency. Even though there are elements of conscious involvement in action that are not introspectively directed, the fact that these other elements also contribute to the sense of agency provides Gallagher with the means to assimilate them into this single, internal, measure:

> …clearly our sense of agency for the action will be tied to that intentional aspect, and that aspect is where our attention is directed – *in the* world, in the project or task that we are engaged in. So clearly a form of *intentional feedback*, which is not afferent feedback about our bodily movements, but some perceptual sense that my action is having an effect, must contribute to the sense of agency.
>
> I suggest, then, that the sense of agency, at the first-order level of experience, is complex because it is the product of several contributory elements: efferent signals, sensory (afferent) feedback, and intentional (perceptual) feedback. (Gallagher 2007, 8)

On this view, to gauge whether or not we are agents, we do not simply look out into the world to see what we have or haven't done (the intentional feedback mentioned in the first paragraph); we gauge conscious agency by introspecting to see if we have the appropriately-based sense of acting for the act committed.

Gallagher mentions distinct ways in which conscious awareness could be involved in action; in particular, he mentions attention that is directed out into the world, not merely at our own bodily movements or sensory feedback. This externally directed element of awareness is distinct from the awareness of agency, but receives consideration here only for the causal contributions it may make to the *sense* of agency, not to agency itself. Gallagher brings up a

number of distinct, externally directed elements of awareness in this passage, but then considers them only in terms of how they relate to the sense of agency. He thereby ends up ignoring the possibility that awareness may have a causal contribution to action distinct from its contribution to the sense of agency.

As I've shown, both Gallagher and Metzinger develop theories of agency that internalize the role of conscious awareness, directing it inwards towards other agentive processes. This illustrates the first assumption discussed above, that of a focus on the awareness of agency as the relevant measure of conscious causal involvement in action. This focus on the awareness of agency corresponds with an associated neglect of other possible avenues of conscious involvement in action. Other, especially externally-directed, forms of awareness feed into the conscious sense of agency by providing mechanisms by which we monitor our actions and compare them against intentions and goals. Both authors ultimately ignore the possible causal roles played by these other aspects of conscious awareness insofar as they contribute to action in ways other than giving rise to the sense or phenomenology of agency.

Similarly, both Metzinger and Gallagher end up with views that are committed to the second assumption listed above, that subjects' conscious involvement with their own actions is to be gauged in terms of introspective reports on how one feels as an agent. While this section has looked briefly at the views of only two authors, the trend is, I claim, more general: a number of contemporary authors looking to the sciences to enrich our philosophical views of agency are looking primarily or even solely at the awareness of agency, and looking to such awareness in order to gain insight into the specifically *causal* structure of agency. I turn now to exploring how this focus on the awareness of agency distorts the causal roles that could be played by awareness in agency.

2.2 The Micromanagement Model of Agency

This section generalizes the two assumptions from the previous section into what I call the Micromanagement Model (MM) of agency. There are three important problems with this view that will be discussed here: the constricted range of possible causal influence that results from directing awareness internally; the fact that the awareness of agency is an effect of the processes that lead to movement, and is thus precluded from having any kind of substantive causal role; and the close tie with introspective reports, which enforces a double-standard on awareness as a cause to which we do not hold other kinds of causes.

A fundamental feature of the approaches discussed in the last section is that they render awareness internally or introspectively directed – the object of awareness is the variety of distinct and internal agentive processes at which such awareness is directed. It is straightforward to see how, if we focus on the awareness of agency and make the fairly basic assumption that the causal influence of awareness is directed at the same object as is the awareness itself,[18] then the only possible avenue of causal involvement for awareness is to affect the other, internal, processes associated with agency. This is what I call the Micromanagement Model of conscious agency: if

[18] I do not discuss this particular assumption, that the causal influence of awareness is directed at the same object as awareness itself, in this dissertation, as it deserves a much fuller treatment than there is space for here. It is worth keeping this assumption in mind, however, because it is one of the key differences between the Micromanagement Model as I present it here, and the view of tryings by Hornsby that will be discussed shortly. Hornsby is commited to the first assumption described in the previous section, that of the directionality of the causal influence of awareness, but then divides the direction of intentionality of awareness from the direction of its causal influence. This means she does not advocate the second assumption, regarding introspective access to mechanisms or other causal influences of action.

awareness is to be causally efficacious in action, it must somehow reach into the self, perhaps even down into neuronal processes, and fiddle with the knobs. In order to have a causal effect on action, awareness must be directed at *and then change* some otherwise ongoing, separate internal process, perhaps even a completely neuronal process. Our awareness is directed at and acts on our own agentive processes. Thus the moniker of 'Micromanagement' – awareness is expected to micromanage other processes within the agent in order to have a causal influence on action.

This is, in some sense, a new version of an old problem. MM treats the conscious agent as directing their own actions like a puppeteer with a marionette: by pulling the right strings or pushing the right buttons, by triggering the right neuronal processes, movement is effected. Even Descartes warned in *Meditations* VI against treating the relationship between mind and body as one of a captain and a ship. Internalist views of agency have been problematized on these grounds by a number of authors (Smith 1983, Gjelsvik 1990). The Micromanagement Model is a new manifestation of this marionette or boat-captain view of the relation between the agent and his or her body. The new twist arises from the reliance on scientific research, and the added expectation of some authors (Libet 1985, for instance) that the micromanaging influence of awareness should be directly on neuronal processes. The three criticisms I'll now elucidate are thus fairly specific to the kind of internalist view of agency that relies on scientific evidence, in particular introspective reports about the experience of agency.

The first and most serious problem with the Micromanagement Model of agency is that it leaves open few if any viable avenues for meaningful conscious causal influence, and, most importantly, it does so *before* turning to empirical evidence. In other words, MM views begin the investigation of the causal structure of agency by first constraining the kinds of structures that we could possibly find in the world to just those where awareness is internally directed, and only

then turns to science to determine which of these few causal structures are empirically supported. When awareness is internalized, it can only act on other internal processes. This dramatically constrains the possibility of finding awareness to be causally efficacious, in two distinct ways. First, there are good reasons (as we'll see shortly, and as should be intuitively apparent) to think that awareness will be systematically unable to causally influence other internal processes, especially nonconscious ones. If we think, as physicalists ought, that awareness is itself associated with at least some kinds of neuronal activity, or supervenes in some fashion on neuronal activity, then it is incoherent to also expect awareness to alter such associated neuronal activity. It is a basic requirement on causal relata that the cause and effect not also bear a logical relation of some kind (such as identity, constitution, or any other kind of supervening relation). As such, there are already constraints on what internally directed awareness could causally influence.

There are further constraints on what awareness could causally influence when we take into account the requirement that the causal influence of awareness be directed toward the same object as awareness is itself directed. It is incoherent to require awareness to causally influence the very neuronal activity with which it bears some kind of logical relation; and yet it is, almost by definition, impossible for awareness to be directed at neuronal processes that are nonconscious. When we 'look inside ourselves' introspectively, we certainly don't see the machinery, as it were, ticking away. Thus, on the internalist picture, it is entirely unclear what awareness could be directed towards such that it would even be capable of having a causal influence on it. Turning the causal influence of awareness inwards means the only responsible conclusion is that awareness is causally inefficacious, and we don't need to look to the sciences to tell us that. Accepting awareness of agency as the relevant kind of evidence regarding causal

structure prejudges the conclusions that can be reached towards the causal inefficacy of awareness. Perhaps one thinks this is the correct judgment to reach; even so, it is question-begging to reach this conclusion by assuming that awareness can only act on other internal states or processes.

This segues into the second way in which the internalization of awareness in agency constrains the possibilities for conscious causal involvement. The focus on the awareness of agency has, in the literature, been accompanied by a corresponding neglect of other potential avenues of conscious causal involvement. Recall how both Gallagher and Metzinger explicitly mentioned other potential avenues of conscious involvement, through world-directed elements of awareness, and then proceeded to consider these avenues only in terms of what they contribute to the awareness of agency. Focusing on the awareness of agency, and thus implicitly buying into a version of MM, need not but in practice has tended to induce a blind spot for other avenues of conscious causal involvement. Thus, the internalization of conscious agency both constrains the possible causal roles for awareness by forcing it to act on other internal processes or states, and also obscures the possible causal roles of other, world-directed, elements of awareness.

The separation of agency from awareness of that agency is a crucial condition for this criticism of the Micromanagement Model to apply. It renders awareness external to the flow of an otherwise self-sufficient process leading to action. For some authors, this self-sufficient process at which awareness should be directed is straightforwardly neuronal, such as activity in the motor cortex. Libet provides a classic example of this. An idea underlying his work (see, e.g., Libet 1985) is that there are neuronal processes in the premotor cortex that lead up to movement, and so in order for conscious awareness to be causally involved in producing movement, it must be causally influential on those premotor cortex processes. This leaves the possible effects of

awareness as directed at and changing features of these basic brain processes. This is why I call it Micromanagement: neuronal processes are in gear leading up to action, and if awareness is to be causally involved, its influence must be on precisely those processes leading to action. Awareness must reach down into the premotor cortex, as it were, and poke at parts of it.

The second major problem with MM also illustrates why focusing on the awareness of agency distorts the picture we get of the causal contribution of awareness to action. As we saw in the discussion of Gallagher's view, the sense of agency likely arises from, or is strongly influenced by, the same neural processes that also lead to motor action. A probable candidate for what gives rise to the awareness of agency is a copy of the efferent signal that goes out to muscles in order to move. But this connection between efferent signals and the sense of agency has significant consequences for the question of causal involvement of awareness. If we take the sense of agency to be the primary way in which awareness is involved in action, then we find that it must always follows motor processes, never causing them. This is no surprise. We have already eliminated any possible causal role for awareness by making it a common *effect*: one process leads both to movement and to the sense of agency. But we are not led to this conclusion by the scientific results themselves; we are led to it by the scientific results conjoined with a philosophical model that focuses exclusively on the sense of agency, and a failure to sufficiently consider other possible aspects of conscious involvement in action.

If research on the sense of agency is taken to illuminate the causal role of conscious agency, the game is already up before the empirical research begins. The premotor processes leading to action will always be a necessary condition for awareness of agency, which means that by the time we are aware of agency, the neuronal preparations for movement are already in full swing. This precludes awareness of agency from occurring before these processes are engaged.

Once again, we see that if this is the only potential causal role for conscious awareness, then the reasonable conclusion to draw is that there is no conscious causal efficacy. The Micromanagement Model pushes us towards conclusions of conscious epiphenomenality.

The third major problem that arises with the MM concerns its epistemic consequences for how we should investigate conscious causal influence, the second assumption discussed in the previous section. If awareness is directed at the processes leading to action, then we should be aware of those processes; this may mean being aware of stages in the process, of the elements that combine over the course of it, or of the causal influences on that process. The contrapositive also follows: insofar as we fail to be aware of some causal factor involved in action-producing processes, then awareness must not have been directed at those processes. That awareness failed to be directed at some element of the relevant processes means, under the MM model, that awareness was not causally involved in the action that resulted from those processes. Thus, it becomes a necessary condition for awareness having been causally involved that we be aware of the separate processes leading to action.[19] This use of introspection to get at the mechanisms of action follows naturally from a view in which conscious agency is considered to be introspective. On this view, we cannot be causally efficacious conscious agents unless we are *also* able to provide introspective reports of how we wield that influence. This requirement for accurate introspective reports on the processes of agency in order for conscious awareness to count as causally efficacious in action can be found in Gazzaniga (1998, 2000), Wegner (2002), and in most of Benjamin Libet's work, as well as that of authors who use Libet's work to establish limitations on the causal efficacy of awareness in action (Haggard and Eimer 1999). In the

[19] The Micromanagement Model allows for the possibility of awareness directed at agency without thereby also being causally involved in agency; awareness of agency is not a sufficient condition for causal involvement.

classic Nisbett and Wilson (1977), instead of requiring accurate introspective reports on the mechanisms of action, the expectation is that subjects should have introspective access to all of the other causal influences on one's action. This means the variety of ways in which the physical and social environment in which we act influences our behavior should be transparent to us: introspection should allow us to see everything that changes our behavior.

Requiring accurate introspective reports is problematic because it is a higher standard than that to which we hold other causes. In essence, it requires awareness to be *more* causal, to be a kind of supercause, in order to be counted as causally efficacious in action. Consider a scenario where variables A, B, and C each have a causal influence on some effect D. This might be icy roads, bad brakes, and a drunk driver for A, B, and C respectively, and a car accident for the effect D. The accident D can be simultaneously causally influenced by each of A, B, and C individually. Perhaps the accident would not have occurred at all had the driver been sober and able to react faster to the bad road conditions and compromised brakes. Perhaps the accident would have been less severe had the road not been icy. These are ordinary causal considerations that we make all the time about events that have more than one causal influence.

The point I want to emphasize is that we are justified in saying, for instance, that C has a causal effect on D. There is nothing problematic about asserting that drunk driving was a causal factor involved in bringing about the car accident, in spite of the fact that there were also other factors, namely the icy roads and the bad brakes, which were also involved. When we change the scenario to one specifically involving conscious awareness, however, intuitions differ on the legitimacy of making such claims. There is a tendency to think that the existence of other causal influences *in addition to* that of awareness must mean that awareness was not in fact influential; some authors seem to think that awareness must be able to trump or override other causal factors

in order to be genuinely causal. Insofar as we do not hold other causes to such a high standard, this is an inappropriate standard to hold simply for awareness.

The Micromanagement Model and its epistemic consequences manifest in a variety of philosophical approaches to conscious agency. Under MM, the possible stances on the causal involvement of conscious awareness in agency are quite limited: leaving awareness out of the loop entirely, so that we are unaware of what is happening and not causally involved in it (see Gazzaniga 1998); construing awareness as able to report on these processes without affecting them, (see Wegner 2002); or treating awareness as reporting on processes while also interfering with them (thus micromanaging the processes already leading to action). This latter view is not advocated as such by any author, and for a very good reason: it is extremely improbable that conscious awareness somehow reaches down into the brain and fiddles with lower level processes in order to exert causal influence. The point, however, is that once we start thinking of conscious agency in terms of awareness of agency, the last option is the only one which leaves awareness with any genuinely efficacious role in action. But it is only because of implicit acceptance of something like the Micromanagement Model that this strange option is the only one left where conscious awareness has causal efficacy in action. If we assume something like the MM, the sensible thing to do is to deny a causal role to conscious awareness before we even investigate the relevant empirical research. If, however, we wish to have a genuinely empirically informed view of conscious agency, we need to eschew the MM and the two assumptions that give rise to it.

2.3 Hornsby's internalist view of agency

Thus far, I have focused on the philosophical views of agency that are scientifically oriented, attempting to use a variety of cognitive science research in order to enrich our philosophical views. My criticisms have centered on the use of a particular area of experimental results, namely those concerned with the sense or awareness of agency, as a means of investigating the general causal structure of conscious agency, and the way in which this approach problematically internalizes awareness. In this section, I want to turn away from the scientifically oriented approaches to agency and look at an example of a philosophical view of conscious agency that comes from an entirely different approach, one with distinct motivations and methods of analysis, to demonstrate how such a view can commit a strikingly similar internalization of agency, and have the same kinds of drawbacks as those I've discussed for the MM. The view of agency I'll examine here is that of 'tryings' as developed by Jennifer Hornsby in *Actions* (1980).

This section is not intended as a criticism of Hornsby's position *per se*. In her later work, she makes little further mention of the notion of tryings as it is set out in this earlier book, and instead advocates a different kind of view of conscious agency. What I want to emphasize is how she can make a similar mistake regarding the internalization of conscious involvement in agency while responding to motivations that are entirely unrelated to those of Metzinger or Gallagher, and while using a method involving the analysis of language and linguistic usage, rather than scientific experiment.

In chapter 2 of *Actions*, Hornsby deploys a distinction she has just made with respect to bodily movements and actions in order to solve a problem discussed by Taylor (1966) and von

Wright (1971). The distinction is between transitive and intransitive forms of verbs and verb phrases. As she sets up the question of the relationship between bodily movements and actions, she notes that bodily movements may all be described intransitively, without the presumption of agency on the part of the subject of such movements, while some may also be described transitively, with attribution of agency to the subject. “His arm moved” is intransitive, while “He moved his arm” can describe the same bodily movement but does so transitively. Actions will often (although, she later argues, not always) include bodily movements properly described with transitive verb phrases.

The problem set out by Taylor and von Wright involves the possibility of generating causal loops between actions, their causes, and their effects. The example is that of a subject who, in trying to please an experimenter, wishes to cause certain muscles in her arm to contract. The subject can only cause these muscles to contract by clenching her fist – she has no direct control over the muscles themselves, but is aware that a fist clenching will bring about the correct muscle contraction. On this portrayal, the action of fist clenching causes the muscles to contract. However, it is the contracting of the muscles that causes the fist to clench – one can’t have a fist clenching without there already having been the right kind of muscle contraction. On this portrayal, the muscle contraction causes the fist clenching. Thus a causal loop appears: the subject causes the muscles to contract by clenching her fist, which fist clenching was caused by the muscle contraction. Taylor and von Wright offer solutions to this dilemma that involve backwards causation in the near-past: they maintain that we can causally affect the immediate past in such a way that it is appropriate to say that the fist clenching caused the muscle contraction, even though the muscle contraction occurred earlier in time..

Hornsby is dissatisfied with this solution. She instead offers a solution with two components. The first component is her distinction between transitive and intransitive descriptions of movements. If we distinguish the transitive causal chain from the intransitive causal chain, there is no longer a loop. Transitive fist clenchings cause transitive muscle contractions; intransitive muscle contractions cause intransitive fist clenchings. We cannot simply identify the transitive and intransitive descriptions of muscle contractions and fist clenchings, and so there is no need for backwards causation. Rather, "when a man clenches his fist, this has (at least) two effects: neither of them goes beyond the body, and the first of them, the muscle contraction [intransitive], causes the second, the clenching of the fist [intransitive]" (1980, 22).

The second component, most interesting for our purposes here, is a view of tryings that emerges from her first component. The causal loop problem has to do with a splitting of the causal path at or just before the muscle contraction occurs. If we locate the core of the action before such a split is possible, then there is no longer a causal loop – the action would cause both the muscle contraction and the subsequent bodily movement of fist clenching in the right kind of way. This is what she means when she claims that the action of a man clenching his fist has two different effects, one of which is the bodily movement of a fist clenching. Whatever constitutes the action, it is earlier than and further 'inside the body' than the muscle contraction. "Every action is an event of *trying* or attempting to act, and every attempt that is an action precedes and causes a contraction of muscles [intransitive] and a bodily movement [intransitive]" (ibid., 33, italics in original). Each successful action, each action that results in the intended bodily movement, was a trying that was also successful. Sometimes we engage in tryings without being successful; in these cases, there is still, Hornsby maintains, something that we *did*, but for

reasons beyond our own control, we were unsuccessful. Tryings, and actions themselves, are thus inside the body, in a manner of speaking: they are that in which we directly engage (successful action being the result of the world, including our bodies, complying with the intention we had in the trying), which occur prior to and cause any actual bodily movements.

To see what is problematic about Hornsby's move to relocate tryings inside the body so as to have them precede muscle contractions, we may change the scenario somewhat and consider a different example. Let us suppose we have a subject hooked up to some kind of brain scanning equipment, following instructions from an experimenter. The experimenter wishes to have the subject's brain light up in a particular area, and both experimenter and subject know that this is the area which is activated when the subject clenches his fist. It is activity in this area that stimulates the muscles in the arm to contract so that the fist clenches (in other words, we are simply moving one step further back along the causal chain). The subject cannot directly activate those portions of his brain, but can do so only by clenching his fist.

We again have a scenario where there is a possible causal loop: the fist clenching causing the relevant brain activity, versus the relevant brain activity causing the fist clenching. Again we can use Hornsby's distinction to discern a transitive and intransitive causal path: transitively, the fist clenching causes the brain activity; intransitively, the brain activity causes the fist clenching. If we wish to use Hornsby's solution, however, we run into serious difficulty when we try to push the action, the trying, further back along the causal chain so as to occur earlier than, and cause, the relevant brain activity. In order to precede and preclude the causal loop, the trying must occur before the relevant brain activity, and be in a position to cause it.

This is problematic on two fronts. First, as something of a side point, Hornsby is going to have difficulty accounting for what a trying is, if it does not at least involve some neuronal

activity. We might want to accept the basic physicalist position that, regardless of what tryings are or how conscious agency is enacted, it is at least going to be associated with some kinds of neuronal activity (we can remain agnostic as to what that association entails). If so, we are going to run out of options for what tryings are such that they can precede any arbitrary step in the intransitive causal chain leading up to a fist clenching. This problem cannot be resolved by stepping yet further back in the causal chain, because we can simply reformulate the scenario as I did with the brain activity instead of muscle contraction, and the same problem will emerge. No matter how many steps back we take in order to insert tryings prior to any intransitive bodily happenings, we will still encounter the problem of locating tryings such that they can enter into a causal relationship with the next step in the causal chain. If that next step is neuronal activity, in the relevant area for our experimenter or in whatever area gets activated just before that area, then either some of that activity is associated with the trying and we run into the same problem of generating a causal loop, or we must reject the physicalist's assumption and accept a strangely Cartesian notion of trying as disembodied yet able to affect physical goings-on in the brain. Thus, the first problem can be put in terms of a dilemma. On one horn, we accept a basic kind of physicalism, and we simply run out of options for locating the trying – there isn't any further we can meaningfully go back in time in order to locate the trying prior to possible causal loops. On the other horn, we reject such an assumption and embrace a kind of Cartesian dualism where tryings are literally a poke from the mental at the physical, a marionette body controlled by specially pulled strings.

Second, and most important for the purposes of this dissertation, this moving back of tryings puts us in the position of having to think of agency as directed at our own internal processes – in this case, even our neuronal processes. Whether or not we are aware that this is

what we are doing when we try to perform some action,[20] our trying is causally directed at the next step in a physical chain that begins with brain processes, proceeds through muscle contractions, and concludes in fist clenchings. We act by acting on these other processes, somewhere along the line, rather than acting directly on something in the world. Even something as simple as a handwave is only an effect of our actual action of trying, not an action in and of itself. All actions are entirely internal to the body, directed at the sorts of processes that can affect the body itself.

In order to solve the problem of the causal loop generated by such scenarios of muscle contractions, and fist clenchings, Hornsby internalized our agency in such a way as to require it to be directed at other processes also internal to ourselves. Her motivation for making this philosophical move is starkly different than that of philosophers examining scientific results for ways to refine our views of agency. Yet the consequences are quite similar: the causal role for actions becomes limited to that of poking at other internal processes, hopefully but not necessarily resulting in the desired bodily movement. Her view of tryings instantiates the first assumption underlying the Micromanagement Model.

[20] This connects to the MM section: on the Micromanagement Model, the direction of awareness' intentionality is the same as the direction of its causal influence; for Hornsby, it appears that awareness can be directed at something – the fist clenching, for instance – while the causal influence of that same awareness is directed elsewhere, presumably at the processes necessary to instigate the muscle contraction leading to the fist clenching.

2.4 Conclusion

Gallagher and Metzinger are only two examples of a broad trend in contemporary philosophy of psychology that addresses conscious agency by focusing on the sense or awareness *of* agency. This has the effect of separating awareness from agency so that awareness can then be turned inwards and directed at our own agency. This focus on the sense of agency is sometimes combined with the recognition that awareness is involved in action in other ways than as sense of acting, but too often these other avenues of involvement are considered only in terms of what they contribute to the sense of agency.

I have argued that the focus on awareness of agency as the relevant target for investigating the causal efficacy of conscious agency is misplaced. This dominant focus can be generalized as the Micromanagement Model, in which conscious awareness is assumed to be introspectively or internally directed, to reach down into the brain and alter ongoing neuronal processes, and, as an epistemic consequence, to be able to provide a report of those processes in order to be counted as genuinely causal. When we fail to find these features of conscious influence in experimental results, many authors have claimed that this failure provides empirical justification for the position that conscious agency is illusory, inefficacious, or generally confabulatory. I have shown that these conclusions do not follow from the evidence itself, but instead from an assumption about how conscious awareness could be involved, namely as awareness of agency.

The view of tryings as developed by Hornsby illustrates how one can fall into this mistake even while considering the issue of conscious agency from a very different perspective.

Internalizing agency, requiring that it be directed at other internal processes, is a recipe for depriving conscious agency of meaningful causal efficacy in action. The Micromanagement Model makes sense of both the tendency to focus on the awareness or sense of agency, and the importance placed on instances of failures of introspective awareness regarding factors that are known to be influential on behavior, both of which are commonly seen in current discussions on conscious agency. It also clearly illustrates why we should take care to avoid this way of thinking about the causal contributions of awareness to agency.

CHAPTER 3: Applying interventionism to conscious agency

I have thus far identified and criticized a trend in scientifically-informed philosophical views of agency to conclude that neuroscience, psychology, and cognitive science offer experiments which undermine the idea that we are consciously involved in our own actions. I want to demonstrate here that these views are not supported by the available evidence, and that in fact, current scientific research, in conjunction with our best available methods of analyzing that research to uncover causal relationships, demonstrate that conscious awareness makes a substantive contribution to agency. Other authors, and my first two chapters, have criticized the conclusion that we have little to no conscious involvement in our own actions. I want to take the next step and demonstrate how we should reason with the available evidence: we have independent evidentiary standards for making causal claims, and these can be applied to this case.

A significant part of this task involves presenting a method by which to apply inferential techniques derived from interventionist causal theory to a set of existing experiments in order to derive conclusions regarding general causal relationships supported by those experiments. I model how to use this method by applying interventionism to a batch of experimental results. The process of incorporating disparate experimental evidence about conscious action into a single framework suitable for causal analysis is philosophically quite interesting. This chapter thus has a dual role. First, it will demonstrate that, when we rely on an independently justified

method of causal analysis, there are substantive empirical reasons to think that we are consciously causally involved in our own actions, in spite of individual experiments that may seem to demonstrate otherwise. This underscores how apparent problems with conscious causal involvement in action are primarily of a metaphysical, rather than empirical, nature. Second, it develops general evidential standards, applicable to a wide range of cases, regarding the type and amount of information needed to bring together individual experiments with the interventionist analytical framework so as to draw causal conclusions.

With respect to the question of the causal efficacy of conscious agency, I will argue that three variables representing conscious involvement in action are each separately demonstrably efficacious on behavior, and that these variables also causally interact in the production of action: consciously held goals or intentions; conscious perceptual information relevant to those goals and the means to achieve them; and conscious execution of action. Relying primarily on automatism research, I will show how these three variables emerge as mid-level generalizations that are specific enough to make substantive claims about agency, yet broad enough to span a significant range of experimental research in a unified way. I also clarify the evidential standards that must be met in order to demonstrate that something is or is not a cause. These standards are not symmetrical: it is much easier to demonstrate that something like conscious awareness is causally efficacious than it would be to prove that it is never a cause. I will show why this is so, and explore the consequences of this asymmetry for empirically oriented views of conscious agency.

In section 3.1, I justify why interventionism is the appropriate causal methodology to use for this particular problem. In section 3.2 I provide an overview of interventionism and the specific tools that it provides for analyzing evidence to reach conclusions about causal

relationships. Section 3.3 explores some major epistemic considerations in finding experimental research that is suitable for use with the interventionist tools. I turn in section 3.4 to automaticity research, describing some of the relevant areas of research within it and how they relate to this task. Section 3.5 distills three different variables from the automaticity research, to represent distinct facets or avenues of causal involvement of conscious awareness in agency. The tools from interventionism are then applied to these variables, with the conclusion that each is a cause in human action. Section 3.6 concludes: I offer this approach as preferable to arguments involving monolithic claims about conscious awareness being simply either involved or not involved. This approach adds empirical structure and complexity to our understanding of the role of conscious awareness in action.

3.1 Choosing a causal methodology

In Chapter 1, I introduced a distinction between metaphysical causal questions and empirical ones. Causal methodology spans the two axes of metaphysical and empirical causal issues. A causal methodology must be grounded in a metaphysical view of what causes are and how we can best find out about them; this methodology can then be empirically applied to systems of which we don't know the causal structure, and the methodology will function in the empirical setting to reveal features of it.

I will use a divide and conquer strategy, by starting with empirical questions about conscious agency in this chapter and then working backwards to metaphysical questions in the next. In this way, we'll see that, once we get clear about the evidentiary standards for making claims about causation, there is no genuinely empirical evidence to the effect that we are not

consciously causally involved in our own actions. Views to the contrary are predicated primarily on metaphysical causal claims: even if many authors provide empirical evidence to support their claims to the effect that awareness is inefficacious, it is primarily the underlying metaphysical assumptions that do the work. My strategy divides such causal claims into empirical and metaphysical components. I demonstrate first that there is ample empirical evidence for the causal efficacy of conscious awareness, and then address the metaphysical problems one might raise with this in the next chapter. In order to this, I need a methodology by which to ascertain the causal role of conscious awareness in agency. This methodology is, like all other causal methodologies, at least partially grounded in a metaphysics of causation, and in this sense may seem slightly question-begging. However, this is not so, for two reasons. If the methodology I use were chosen because it yields the kind of empirical answer I am looking for, and did so because of the assumptions about metaphysics that are required to use it, then the situation would indeed be question-begging. But there are good reasons independent of my project to think that interventionist causal analysis is the best available means of causal analysis currently available. Second, the methodology I use in this chapter is compatible with a range of metaphysical views about the nature of causation; it does not confine us to a single metaphysical account (although at least some of the philosophers advocating interventionist theories, such as Woodward, do subscribe to a particular metaphysical view about causation).

There are a number of advantages to utilizing a broadly interventionist methodology, rather than a methodology based on an alternative metaphysics of causation. The first reason is that interventionism has developed highly useful tools of causal analysis. By contrast, many theories of causation come with no useful tools of analysis. Theories that have been constructed to account for all of our intuitions generally must, in order to accomplish this, sacrifice explicit

rules for finding causal relationships, since one can generate an almost endless stream of strange scenarios as counterexamples to any rule that is sufficiently specific to apply. The case of agency-based or manipulationist theories are an example of this.[21] The view that we derive our notions of causation from the exercise of our own agency does not provide the means to ascertain, for situations where we are unsure of the causal structure and have no particular agentive experience with the variables in question, what causal relations there are in the world. This is not to say that this failure to develop inferential rules is a problem *per se* with these theories. Most if not all of them were created for other purposes than investigating new causal structures in the world. Still, this does imply that such theories are poorly suited for the task at hand here.

Even theories with rules for sorting out causal versus noncausal relationships do not always provide *applicable* rules. Consider the mark/conserved quantity transmission theory of causation due to Salmon and Dowe.[22] There are clear rules for what counts as a causal relationship. And yet these are not rules that we could even hope to use to ascertain causal relationships in cognitive science. Applying such a theory would require us to keep track of a combinatorial explosion of information about transferred quantities in order to trace even a single signal through the brain, not to mention in order to track how it might influence subsequent movement. Before even reaching the problem of information overload, however, we would encounter the very basic problem that these transfers are not ones to which we have any epistemic access. Our access to brain activity is of a much coarser kind, sometimes even simply of an input-output sort, with no information about the detailed microphysical processes in any

[21] Menzies and Price (1993).
[22] Salmon (1998), Dowe (2000).

given instance. Physical transmission theories of causation are simply not usable for the kind of information we get from the cognitive scientific and psychological experiments relevant to conscious agency.[23]

Possible world theories for evaluating counterfactuals fare no better in this kind of task. In order to apply such theories to find out whether or not X is a cause of Y, we must look to the nearby possible worlds in which X occurs and does not occur, and see whether or not Y also occurs. But, to be perhaps overly literal, we have no actual means of checking those worlds. We must already have the information to provide the answers about whether or not Y occurs in the nearby possible worlds. We cannot use this kind of approach to find out causal information that we don't already have – we can use it to represent or redescribe causal information we already have, or to compare conflicting intuitions. It is not an actual tool of discovery.

In contrast, interventionism has a well-developed theoretical structure that can be applied to diverse situations about which we may not already have firm intuitions, and can be applied without already knowing the physical mechanisms[24] which give rise to observed causal relationships. I include the work of Judea Pearl (2000), James Woodward (2003), Clark Glymour, Peter Spirtes, and Richard Scheines (2001), and others under the rubric of interventionist frameworks. There are important differences between some of these approaches,

---

[23] Sections 4.2 and 4.3 of chapter 4 contain more detailed argument about why the conserved quantity transfer theory of causation is poorly suited to any higher-level cause, including but not limited to conscious awareness.

[24] Note that this usage of the term 'mechanism' is not same as one finds in, for instance, Woodward (2003). There, a mechanism can refer to a single line in a set of equations describing the functional relationships among a system of variables, so long as that line can be intervened on independently of the other lines. Here, I intend to refer more to the usage of (Machamer, Darden, and Craver 2000), where the mechanism underlying a causal relationship refers to the series of entities and their activities which concretely instantiate and give rise to the causal relationship.

but those differences manifest only at a finer level of detail than will be necessary to make my case.

A further reason in support of interventionism is that it requires only a minimal ontological commitment. The fundamental notion in interventionism is that of a manipulation or intervention in a system. The basic notion of a causal relationship where X causes Y can be translated as saying that if we intervene to change X, Y will also change, whereas if we intervene to change Y, X will not change. The asymmetries of interventions on causes and effects provide the fundamental metaphysical assumptions of the interventionist program.[25] Interventionism is thus compatible with a number of distinct metaphysical views about causation, so long as these views have the common consequence that causes can make a particular kind of difference to their effects (a minimal and common assumption about the metaphysics of causation[26]).

One of the metaphysical commitments that cannot be avoided with interventionism, and to which I will be committed throughout this dissertation, is a basic realism about causation: namely, that there are genuine causal relationships in the world (no matter what they are), and that we have some, albeit limited, epistemic access to them. Causation is not merely a projection of our own thoughts onto an otherwise acausal universe, as some understand Hume to have claimed.[27] Nor is it merely a naïve misunderstanding of the products of science.[28] I take this realist commitment to the existence of causation, however it may be specificallyconstrued, to be commonly shared among philosophers who work on the issue of whether or not conscious awareness is a cause. If there is no such thing as causation, then the question they claim to

---

[25] The fact that this is a nonreductive characterization of causation does not mean that it cannot also be nontrivial and illuminating, as Woodward (2001, 2003) has compellingly argued.
[26] See, for instance, Mill (1862), Reichenbach (1958), Cartwright (1979).
[27] Beebee (2006).
[28] Norton (2003).

explore is moot. Many, however, take it as a given that there do exist causes, and that the debate is instead about what should count as belonging to the class of causes.

The limited metaphysical commitment of interventionism is valuable in making my case. It emphasizes the extent to which the conclusions I draw here are genuinely empirical ones, reached using the fewest and most innocuous metaphysical causal assumptions I could make while still be talking about causal relationships. These conclusions do not depend on a problematic causal theory, and even for critics of interventionism, the aspects of interventionism which might lead to difficulties (for instance, requiring the system to meet the Causal Markov condition, see Cartwright 2002) are not required to make my point. I emphasize the empirical nature of these conclusions, and the minimal metaphysical assumptions about causation needed to reach these conclusions, to provide a starting point for the next chapter. There are numerous positions on conscious agency that both purport to be empirically supported, and that conflict with the thesis defended in this chapter. Highlighting the causal assumptions needed to make my point, and the empirical nature of the results I explicate here, provides the means to pinpoint the precise disagreement between my own position and conflicting ones. The conflict between my view and views that advocate little to no role for conscious awareness in agency will turn out to be *not* empirical – that is, they are not about the actual causal structures we find in the world, but instead reflect disagreements about the nature of causation and what should be allowed to count as a cause.

3.2 Using response structures

In this section I'll briefly introduce some key tools from interventionism that will be put to use in the later sections. The core notion is that of manipulation or intervention. This is not an anthropocentric view of causation: there is nothing in the notion of manipulation or intervention such that a human agent must actually do it or must be capable of doing it. The idea is simply that when a cause is changed, while holding other potential causes fixed, the effect will also change, and that in changing an effect, its cause will not change.[29] These changes need not be deterministic or regular – a cause can increase or decrease the probability of its effect but not vice versa. Variables are the relata of causation. This is a type-level theory of causation: tokens of causation count as instances of some variable, which is a type-level causal entity about which we make claims of causal relationships.

Each variable includes a specification of the values it can take. This is an essential part of what Woodward has called the "contrastive focus" of interventionism (Woodward 2008). Sets of distinct values that can be taken by a given variable provide the relevant contrasts for each variable, which are specified as part of the variable itself. A switch between one state and another state in the cause results in a switch between one state and another state in the effect. The property of being red, for instance, does not as such and without further specification *cause* anything. However, a stoplight having the property of being red can certainly cause a driver at an

[29] This should come with the caveat that these changes in the effect may not always be actualized: there may be a balancing effect from another source, so that there is no apparent change in the effect. These cases are complications of but not counterexamples to the simply stated form of the criterion given here.

intersection to stop rather than go. We explain what it is about the situation that is causally relevant: being red *rather than* green. Red has other contrasts, such as being purple, but this is not a contrast that is causally relevant to stoplights. The variable values give the contrast class across which the cause is efficacious: having this property *as opposed to that one*, leads to an effect having one property instead of another. The contrast class is given, so that the causally relevant features of the single instance are already picked out by the variable of which it is an instance (see Woodward 2008, 218-220).

The primary tool on which I will rely is that of a response structure (see below for an example).[30] Response structures are tables of values: each column represents a variable, potential causes to the left and the effect in question to the right. Each row is numbered, and the values along a row are the values taken by each variable in some situation in the world. Every combination of values for each potential cause in the system has a row. This means that the number of rows equals the permutations of values for all the potential causes being considered. A row stands for some actual situation in the world, where each variable has the specific value listed in the response structure row. This means we must "set" the values for these variables by manipulation, even if we never actually reach in and physically intervene on the system. We exhaustively catalog the values of potential causes and then use this information in the table format of the response structure to find test pairs. A test pair for a variable X and effect Y is a pair of rows where all other potential causes have the same values in both rows but where the

[30] For a thorough introduction to response structures, see http://www.cmu.edu/oli/courses/csr/.

values for X differ, and where the value for the effect Y then differs. By finding such a test pair, we can demonstrate that "wiggling" X, holding fixed other causes, leads to a wiggling of Y.[31]

The smoking example is represented below in a response structure. The variables are in the top row, along with the values they can take (true or false in this simplified example). The last column is the effect, and the other two variables are potential causes of that effect. All combinations of values are represented for the two potential causes.[32]

Table 1: Sample Response Structure

Assignment #	Smokes [T,F]	Chews gum [T,F]		Gets lung cancer [T,F]
1	T	T		T
2	T	F		T
3	F	T		F
4	F	F		F

What we find in this response structure is that smoking is a cause of lung cancer, whereas chewing gum is not. Rows 1 and 3 are a test pair for smoking, as are 2 and 4. There are no test pairs for chewing gum: no rows in which smoking remains fixed but the value for gum chewing and the effect of getting lung cancer changes. This gives us a useful way to predict the outcomes of interventions: if one is concerned about lung cancer, stopping gum chewing is not good protection, but avoiding smoking is.[33]

---

[31] Variations of this criterion have a long history prior to its usage by Woodward, Pearl, Spirtes, Glymour, and Scheines: see especially Mill (1862), Reichenbach (1958), Suppes (1970) and Cartwright (1979).

[32] I have simplified the case from probabilistic to deterministic simply to make the format of the structure clearer. Instead of T or F only in the effect column, we could use T=A% and F=(1-A)% in each effect outcome box to accommodate probabilistic causes.

[33] There are interesting issues to be explored regarding the homogeneity or heterogeneity of the populations represented in a response structure. Ideally, one would like to construct a response structure using a homogenous population: one where the individual members of the population

A common requirement is that only one test pair of rows need be found to claim that X is a cause of Y, even for response structures with many rows.[34] This requirement means that if a variable ever makes a difference to an effect, it is a cause, even if it only does so under very specific conditions (where the values of the other variables are held at a single set of values). I am going to utilize this weak requirement here, where a single test pair is sufficient to demonstrate that a variable is a cause. For complex systems, with many causes and background conditions that interact with one another, we are more likely to encounter causes that, no matter how strongly causal they are, do not exercise their influence in all circumstances. These are

---

all evince the same reactions under the same conditions, or where the causal structure underlying the population is uniform. There are occasions when we might mistakenly think a population is homogenous, when in fact there are two or more distinct subpopulations, each of which responds to the same causes in a different manner. Constructing a single response structure for such a population may result in apparent changes in outcome for a cause variable which is in fact not a cause in either population, if the populations respond differentially to other variables that are included in the structure. Or, the response structure may appear indeterministic, with probabilities figuring in the outcome column, even if the responses of each subpopulation are deterministic. I will not address these issues, but it will be relevant in a later section of this chapter. What matters is that if we know two populations to have different causal structures, i.e., to respond differentially to the same causes, then we can and ought to utilize two distinct response structures, one for each population, rather than combining them together into one.

[34] The requirements on test pairs in order for a variable to count as a cause can vary in strength, from one test pair being required to count as a cause, to all possible test pairs being required to count as a cause, see Cartwright (1979). This means that one such pair failing to be a test pair would indicate that the variable was not a cause, even it influenced the effect outcome in all other circumstances. No other authors endorse such a stringent requirement on causation, and even Cartwright (personal discussion) has changed her view. Enforcing such a requirement would mean that only those causes which, when they operate, trump all other causes, would be counted as genuine causes. I cannot think of a single nontheological cause that would act in all circumstances, regardless of the presence or absence of other causes, which makes me suspicious of there being any such causes anywhere.

precisely the kinds of systems that will be investigated in this chapter, and using a criterion that allows such causes to be labeled as such is appropriate.[35]

One implication of this standard (and of stronger standards also) is that there is an asymmetry in what is required to show that something is or is not a cause. In order to demonstrate that X is a cause of Y relative to a system of variables, it suffices to demonstrate that there is a single test pair, a single set of conditions under which X causes Y. To show that X is *not* a cause of Y relative to a system of variables, we must instead demonstrate that none of the potential test pairs are in fact test pairs, that X does not cause Y under any circumstances. While this might initially seem like an unfair advantage, this asymmetry is not new. Consider what it takes to establish the claim that all ravens are black: one must demonstrate that all ravens are black. In contrast, to demonstrate that not all ravens are black, we need only find a single nonblack raven. To claim that X is not a cause, there are simply more cases that must be checked in order to verify that it does not act in any of those cases. But to establish that X is a cause, we need only verify an existence claim: there exist circumstances under which X causes Y.[36]

---

[35] We could use test pairs for variables as a means of comparison of relative strength, if we are so inclined. A variable with more test pairs will *prima facie* be a stronger cause, or a more widely distributed one, than another variable with fewer test pairs.

[36] A concern one might have about this method of establishing causation is completeness, that there is always the possibility that some important variable has been left out of the structure, as a result of which a variable appears to be causal when in fact it is not. There are several different angles to this concern. First, one might be concerned about causal sufficiency, which holds when there are no unmeasured common causes of two or more variables in the system. If there is a common cause of two variables, X and Y, and this common cause is not included in the system, then X and Y will be dependent on each other no matter what we condition on, even though neither is the cause of the other. The problem of causal sufficiency, however, is endemic to any application of interventionism, and has by no means proved insurmountable. Further actions, such as intervening on X and Y separately, will indicate that neither is a cause of the other, and thus that some unmeasured common cause has been left out. Second, the results of an interventionist analysis of causation will necessarily be fallible. There is always the chance, even if it is remote, that we are wrong. This possibility alone should not prevent us from using the

These considerations point to a feature of interventionism that differentiates it from a number of other causal theories, namely its emphasis on the pragmatic nature of causal judgment. Claims that a variable is a cause are always made relative to a particular system of variables. If only X and Y are considered, then perhaps we simply find that X causes Y. If we instead consider X, Y, and Z, we could find that X causes Z which causes Y. Or we could consider A, B, and Y, where A and B together are the equivalent of X, and find that A causes Y but B does not. There is no single ultimate right answer regarding causation, no definitive set of variables that is *the* best to use. There are better and worse representations, but no one "correct" representation independent of our investigational goals. This accommodates the impact of investigational interests, but does not thereby introduce an ineliminable subjectivity to the realm of causation.[37] The reason for this is that *given* a set of variables (regardless of the reason for choosing this set), the question of whether or not there is a causal relation between any two nodes has no element of subjectivity to it. Even though the question of causation is always relative to a set of variables, *if* a variable is causal, it means that we can actually intervene in the world on that variable and bring about a change in the effect. [38] [39]

---

analysis anyway: it is still the best available method, and we are using it in conjunction with the best available science on the topic. There is nothing one can do to guarantee infallibility, and to expect otherwise is unreasonable.

A second angle on the concern about potential missing variables is that the included variables may be compounded of other, more fine-grained variables, or that the included variables are in some fashion high level and that we should instead include the low level variables on which these supervene. I will bracket this concern for now, as it is answered in great detail in chapter 4.

[37] See Mitchell (2000).

[38] In fact, being able to intervene and control is a well-known criterion for realism, e.g. Hacking 1982 and Cartwright 1983. The system-relativity of causal relationships does not mean that we need to abandon realism about these causal relationships due to a concern that there is *really* some other variable that is more causal that the apparently causal one that we have already found. This is true for the same reasons that the simple possibility that our current conception of

quantum mechanics may be found to be incomplete or inaccurate shouldn't make us nonrealists about the very small size scale of the universe. That a further variable or new discovery may change our opinions about what the causal relations are, or about the best theory for the physical micro-scale, doesn't mean we actually have reason now to doubt these theories in favor of something else or to think that there are no 'real' causal relationships in the world.

[39] A word is in order regarding the relationship between the presentational structures of interventionism, such as directed acyclic diagrams (DAGS) and the systems they represent. An intervention in a system is represented with an exogenous variable added to the DAG, with an arrow directed into the variable being intervened on. If this intervention is an "ideal" one, then all other arrows into the intervened-on variable are broken – they can no longer affect the variable since the variable is set to a specific value, and thus cannot respond to other causal inputs. In actual systems represented by the DAG, intervention is not always so clear and easily accomplished. It may be that we do not have, pragmatically speaking, anything like an ideal intervention. Instead, we may be limited to interventions that influence the desired variable, but in an incomplete way, allowing other causes to continue influencing it as well. This is referred to with the distinction between hard and soft interventions: hard interventions break the arrows between all causes of an intervened-on variable and the variable itself, whereas soft interventions do not: the latter influence the intervened-on variable without eliminating the pre-existing influence of other causes of that variable. In spite of being nonideal, soft interventions can actually be extremely informative about the causal structure of a system (Eberhardt and Scheines 2006). This is especially relevant for systems where a single component serves multiple distinct functions in multiple processes. Modularity is the requirement that we be able to intervene on each cause independently, without affecting any other variables directly. Intervening on a component as part of an intervention in one process will necessarily also have an effect on the other process, when the component figures in multiple processes at the same time. There are two questions here: must a cause act in accordance with modularity in order to count as a cause in the first place; and is modularity required for inferring causal relationships? (thanks to Edouard Machery for pointing this out)

Serious criticisms of modularity as a requirement on being a cause have been levied elsewhere (see Cartwright 2001, Mitchell 2008a). I will not adjudicate this debate here, since there is the more immediate issue that, regardless of whether or not all causal systems are in-principle modular, the systems to which we actually have access fail modularity. We needn't decide now if this is merely an epistemic failure, or if the systems are genuinely nonmodular. We do not have ideal intervening capabilities, and as such, we may be in practice unable to intervene on each variable alone. Whether or not all causes are ultimately modular is an ontic question about causation itself, while the second question is methodological. Fortunately, we do have an answer to this question: sometimes a system may fail to be modular and yet we can still make inferences about causal relationships in that system. This is fortunate because the situation of single components figuring in multiple processes is ubiquitous in the central nervous system – a single neuron can fire as part of multiple different populations of neurons, while also causally interacting with nearby neurons chemically without firing. An intervention on such a system will be likely to affect more variables than just the one being intervened on. But, even though many interventions will necessarily be "fat-handed" in this respect, this is not always problematic. Our interventions need not be ideal to nevertheless be informative.

I expect many readers are already familiar with this terminology to some extent. Further details can be found in Pearl (2000) and Woodward (2003). For now, the next task will be to apply some interventionist tools to the issue of conscious agency.

3.3 Filling response structures with scientific experiments

In this section I will explain some of the epistemic considerations involved in taking the abstract analytical tools from the last section and fitting them to the specific details of research in cognitive science. This is a two-way fitting process: we need to ascertain what kinds of research are most appropriate to utilize given the tools we have; and we need to choose how to deploy these analytical tools based on the research that is available. The epistemic considerations that I will bring out in connection with the question of the causal efficacy of conscious agency are generalizable to other kinds of systems. This exploration of the epistemic considerations is a two-way fitting process between the abstract analytical tools at our disposal, and the exigencies that come as part and parcel of considering some actual research that was not designed for the sake of fitting neatly into these tools.

In order to establish the claim that there is a causal relationship between two variables, it is not necessary to use a response structure. Causal claims can be established in other ways, and we may often lack the resources to fill out a response structure while still having legitimate evidence for the existence of a causal relationship. But being able to fill out and utilize a response structure for ascertaining the existence of a causal relationship is a strong way to establish such a claim. I will thus aim to meet a higher evidentiary standard than is strictly necessary.

Our aim, then, is to find individual experiments or sets of experiments in the scientific literature that are suitable to serve as rows in a response structure. Each row, recall, stands for some particular situation in the world, where the potential cause variables take specific values, which then determine value of the effect variable. The potential cause variables are considered as 'set' to these values; since we exhaust all combinations of variable values, this is equivalent to setting the variables to each of their respective values. This also ensures that we are talking about real interventions in real systems in the world, as messy and complex as these can be, not merely ideal interventions in idealized systems. A pair of experiments that would fit the bill would be quite similar but not identical: the same sets of variables would be held fixed in both experiments, but in one experiment the test variable takes one value, and in the other the test variable takes a different value.

Since different experiments utilize somewhat different ways of operationalizing the relevant variables, we need to be aware that some degree of generalization over and away from the details of experiment set-ups will be necessary. In order to fill out a response structure, we must generalize at least somewhat to commonalities between these experiments – we are looking for variables at a middle level of generality. This will become clearer as we proceed through this section. Part of this task is finding the appropriate level of generalization for the variables used to represent different facets of agency. This level must not be so specific as to render obscure the connections between sets of experiments, yet also not so general as to cover over causally relevant differences.

We are also looking for pairs of experiments such that the action outcome (the effect variable) for the pair is closely related or the same action. It would be quite difficult to find such groups of experiments with consistent specific action outcomes, or effect variable, for all eight

(or sixteen, etc.) rows in the table. Matching in twos is not so hard; matching them all would be extremely difficult. Fortunately, we don't have to make the action outcome, the effect variable, match exactly for all the rows of the response structure. We can invoke the standard for claiming that one variable causes another: there is a test pair for it. This means we are essentially looking for a *change between* two rows, not an absolute or even quantitative measure, and we can ascertain whether this is the case without having to make the action outcome in every row completely consistent with every other outcome in other rows. We can instead make do with a looser kind of pairwise coherence between the outcomes: for each potential test pair, the action outcomes must match in order to be sufficiently comparable.

Let me spell this out in more detail, as it is rather neat feature of using these response structures and is key for the incorporation of a multiplicity of experiments. Let's start with a system of binary variables A, B, and C, to ascertain the effect on E, where E is measured in a variety of distinct ways in different experiments, many of which have no obvious means of comparison for ascertaining whether the values E takes in the experiments are the same or different. To see if row 1 figures in a test pair for variable A in this system, we do not have to simultaneously compare row 1 to all the other rows; that would involve finding experiments, the outcomes of which are instances of E and which are sufficiently similar so as to allow direct comparison across the outcomes. We proceed pairwise: first we hold B and C fixed at the given values they have in this row, and then locate a row where A takes a different value. Then we simply have to find two experiments that represent E in a sufficiently similar way so as to allow us to check whether E has a different value in each of these two rows. That means we can get away with matching the experiments up in pairs of closely related outcomes, rather than in sets of 8 (for this example) related outcomes.

Starting again with row 1 and looking for test pairs for B and C, the rows that would serve in potential test pairs will differ, and we can find a different experiment that serves as an instance of those variable values than the one we used for row 1 when looking at test pairs for A. This works because each row will have multiple experiments that instantiate that combination of values for those variables. There may be multiple different experiments that each represents a situation in the world corresponding to the values of those variables. (Indeed, it turns out that if we have done an adequate job individuating the variables we began with, and the field of research involving these variables is at all mature, then we should not have difficulty finding a plethora of suitable experiments.) These variables were presumably chosen precisely because they are capable of generalizing over multiple experiments. The potential cause variables must allow us to clearly group a number of experiments according to the values that these variables take, but we can be much more lax about the effect variable, since we are, at this stage, only looking for changes between rows, and not a more quantitative or absolute measure of the effect. This will be clearer in subsequent sections as I demonstrate the technique.

This demonstrates how features of the response structures can guide our selection of the appropriate kinds of experiments by which to fill out such a structure in order to ascertain if there are test pairs for given variables. Now that this background is in place, we can turn to the compilation of experiments to complete this task. Since the question is that of the causal efficacy of conscious agency, it should be obvious that we need to look at fields of research where conscious agency is operationalized. We shouldn't expect it to be operationalized in a monolithic or singular way, and there will most likely be at least several distinct components of conscious agency that will be addressed in the research. In addition to simply finding experiments that

involve conscious agency, we need experiments that explicitly *don't* involve conscious agency – experiments that involve the same variables but with different values, ones that hold fixed the same variables at the same values while wiggling another.[40]

Automatism research may seem as if it counts against conscious causal involvement, insofar as it examines the range of sometimes extremely complex behaviors that can take place automatically, with little to no conscious involvement. But, when considered in terms of how these experiments should be represented causally, automatism provides the right contrast class to the range of conscious actions, making this a fruitful area of research for experiments contrasting with conscious agency.[41]

---

[40] This point is crucial for seeing why so many authors have misunderstood the evidentiary import of single experiments regarding the causal contributions of conscious awareness to agency. While it will be explored elsewhere, I want to note that a single experiment in which conscious awareness does not appear to be involved, yet where there is nonetheless some kind of action that occurs, does not by itself demonstrate anything about the causal involvement of conscious awareness in general. The smallest relevant unit by which we can gauge whether or not conscious awareness is making a contribution to the action is a pair of experiments, in which a variable representing the relevant avenue of potential conscious agency is wiggled between two values. That comparison between the outcomes of two relevantly matched experiments is the smallest unit that supports claims for or against causation.

[41] What automatism research also provides us with is a clearer way to understand conscious involvement in action that is not self-directed or introspective, per the criticisms in chapter 2. Consider the example of conscious execution of action. In conscious execution, we consciously, deliberately, attentively, or intentionally, perform an action. The role of conscious awareness is, on a first reading, to negotiate specific details about the performance of an act, including the goals into which the act fits, the environmental information that bears on how the act should be performed, and the actual careful performance of the act. Automatic behavior demonstrates how integral automaticity is for overall conscious execution of action: without the ability to develop away from conscious execution towards automatic execution, our attentive resources would be completely occupied in the small details of mundane tasks, and we would be unable to perform more complex ones.

3.4 Automaticity and conscious agency

Automaticity, as it bears evidentially on the causal role of conscious awareness in action, has been little considered except insofar as it bears specifically on the *awareness* or *sense* of agency, as criticized in chapter 2. Indeed, automaticity has typically been considered evidence *against* the causal efficacy of conscious awareness in action. Daniel Wegner, for instance, devotes several chapters of his 2002 book to what he calls 'automatisms'. These are actions which are performed by a subject but which the subject is unaware of performing, or which are performed without conscious involvement. An example is the movement of a Ouija board: the subject is responsible for the movement of the piece over the board, but does not experience a sense of agency as doing so. But this does not accurately represent the larger range of automatic behaviors. Whether or not one has a conscious sense of agency is not what makes automatic actions automatic. Automatism and sense of agency vary independently. Because of the tendency to focus on the sense of agency, as criticized in chapter 2, as well as a tendency to consider single experiments at a time, rather than pairs of contrasting experiments, the relevance of automatism research for other aspects of conscious involvement in action has not been sufficiently appreciated.

Automatism refers to those behaviors in which humans engage but not in a fully conscious fashion – actions that are not consciously executed but which are nevertheless voluntary in some sense, often goal-directed, and sometimes extremely complex. The classic example of automatic behavior is driving. When learning to drive a car, we are often quite overwhelmed when first behind the wheel, trying to keep track of all the relevant factors such as

speed, location in the lane, surrounding cars, rear-view mirrors, watching the gauges and controls, and perhaps a nervous parent in the passenger seat. All these tasks must be consciously remembered and performed. Soon, however, we master the behaviors that go into good driving and find it easy to drive without paying much attention to what we are doing. While we are still able to consciously drive, we often do not need to. This recession from awareness of the complex behaviors associated with driving is part of the process of automation.

There are three broad classes of automatism, which can be differentiated according to the level of potential conscious involvement in the action; examples of each will be given shortly. The first class includes automatic behaviors that are not consciously executable: they never involve conscious execution, and could not be consciously interrupted with any amount of effort. The second class includes automatic behaviors that were not formerly consciously executed; they were initially unconscious, but can be interrupted with conscious effort. The third class includes behaviors that initially required conscious involvement, but with practice recede from awareness and are performed automatically. Automatisms belonging to the first class are not consciously interruptible, while those belonging to the second and third class are.

Automatisms that were not initially consciously executed generally involve priming, a form of pre-conscious processing of environmental stimuli. These affect our behaviors without our being aware that we have been primed. The physical and social environments have a much greater effect on what we do than we realize. Some of these primed behaviors can, if we are aware of them, be consciously interrupted; others cannot be derailed even if we are aware of the fact that they are operating.

An example of interruptible automaticity that results from pre-conscious processing comes from Bargh and Chartrand (1999). Subjects were told they were to work on a task while

another participant did the same, the other participant actually being a confederate of the experiment. Confederates in test conditions engaged in some single behavior, either touching their face or shaking their foot often. Compared to control sessions, subjects in test conditions also engaged in an increased number of these same behaviors, without realizing that they were doing so: they touched their faces much more often if the confederate was doing so, or shook their foot if the confederate was. Perceiving a fellow human's behavior primes us to engage in that same behavior. It is by no means a deterministic connection, but it still strongly predisposes us to do something without our being aware either that the other person is engaging in this particular behavior or that it is influencing our own behavior.

This kind of automaticity is based on pre-conscious processing of environmental information, but it can be consciously interrupted. Another class of automatism, also often a result of pre-conscious processing, can't be interfered with by conscious effort. Semantic priming decreases subjects' reaction times for recognition of words which match the semantic content of the prime, and increases the reaction time for words unconnected to the prime. A masked priming word is flashed, so that the word is visible for too short a time to reach conscious awareness before a mask appears over it. Subjects then respond to certain words, either semantically congruous to the prime or not congruous: a prime of 'red' and recognized word 'color' might be congruous, while 'red' and 'justice' are not. Response times in recognizing congruous words are faster than response times for non-congruous words when the interval between prime and word is short (Joordens and Becker, 1997, Forster 1981). Conscious effort can neither allow us to become directly conscious of this effect nor to interfere with it. It is an entirely automatic process. Even if we concentrate and try to consciously speed up our

reactions for mis-matched primes, or slow down our reactions to matched primes, we are unable to compensate for the significant but extremely short time differences in performances.

This brings us to the last class of automatic behavior. Behaviors in this category are ones which used to be consciously executed, and could again be consciously executed, but because of overlearning as a result of practice, can be performed without conscious involvement. This class of automaticity is further broken down according to degree of automaticity per Bargh's (1994) categorization: intentionality, awareness, and control can each be separately automatized, or jointly so. We can automatize goals which we act on regularly in common circumstances. We can automatize awareness as we learn to drive without having to consciously check the rearview mirrors and surrounding traffic, but do so automatically. And we can automatize control, as most of us learn to do with typing on a computer keyboard. The examples that fill out this class of automaticity are drawn from familiar circumstances in everyday life.

Bargh and Chartrand (1999) describe a primary way by which automatic goal-oriented behavior occurs, through the operation of two different processes in sequence. These can be neatly understood in terms of the un/interruptible distinction. The first leg of this route is of the sort of automaticity that not only doesn't but couldn't involve conscious awareness. This is from "environment to perception." It includes the earlier examples of priming, where a prime of which subjects are unaware nevertheless influences reaction time (Draine and Greenwald 1998). Environmental cues, such as behavior displayed by other humans around us, automatically affect our perception so that even though we aren't aware of seeing others touch their face often, this information is unconsciously present. This leg is not consciously interruptible.

The second leg of the path is from "perception to behavior." In this leg, as Bargh and Chartrand claim, the cognitive activity associated with the perception of environmental cues is

sufficient to at least predispose us to the behavior being perceived, without our ever having consciously executed the behavior. We can, though, if aware of what is happening, consciously interrupt these behaviors. This leg also includes behaviors where the perception to behavior path was formerly a consciously executed one, and became automated over time.

Breaking down automaticity into these two legs is useful because it allows us to more clearly explicate what is automatized in behaviors that cannot be interrupted and those that can. It connects well with interventionism: this potential for interruption is essentially the potential for an intervention, namely an intervention on the value for some variable representing conscious awareness. It also allows us to understand some of the causal connections between these three elements of conscious agency – how conscious perceptual information, for example, influences the manner or degree of conscious execution. For instance, when looking at the length of reaction times to matched or non-matched primes, the "environment to perception" leg was never a conscious influence on reaction and could not be made so, and the "perception to behavior" leg cannot be altered under conscious control. For the face-touching example, the environment to perception leg was not initially a conscious one: subjects were unaware of the prime and were consequently unaware of their own response to it. But subjects can be informed about it, and thus the environment to perception leg can become one of which we are consciously aware. And, when this occurs, subjects can consciously interrupt the behavior, i.e. refrain from touching their faces so often.

The point which emerges from this way of dividing up the complex phenomenon of automaticity is that a key difference between automatic and consciously executed behavioris whether or not the environment to perception leg is one of which we are (or could be) aware. The ability (whether or not it is exercised) to consciously execute or inhibit an action rests on the

capacity to be aware of the relevant perceptual information. If we are capable of being aware of the environmental information in question, then we are capable of performing the associated action deliberately, that is, of consciously executing it. This allows us to make explicit a variable by which to represent one element of conscious causal contributions to actions: Conscious perceptual information [yes, no] can serve as a basic variable to represent the relevance of being aware of perceptual information versus being unaware (such as in semantic priming).[42] The fact that this ability can be exercised or not indicates a second variable to represent whether or not conscious effort is involved in interrupting or performing an action: Conscious execution [yes, no].

One last element of Bargh and Chartrand (1999) involves the third aspect of conscious action that can become automated: intentions or goals. Goals or intentions can become automated, or can be automatically induced, in a similar fashion as perceptual awareness and execution of action. Sometimes we are confronted with a similar situation on multiple occasions, and make similar choices on those occasions. Our goals in those kinds of situations may initially be conscious: we are aware of the range of options, we weigh them, and make a deliberate choice on which we consciously act. However, making the same choice, acting towards the same goal, in similar situations on multiple occasions can have the effect of automating that goal. When confronted with a similar situation in the future, we can automatically act towards the same goal,

[42] Perceptual awareness of environmental features would be better represented, in a more complete system, as involving degrees. Binary variables, however, don't do a great injustice to the research, since many experiments are measuring awareness in a binary fashion, as present or absent. I am here going to use binary variables because it is simpler as a means to make my point, and because this representation really serves as an initial starting place from which to further elaborate on and refine the representation, not as a final polished product.

without deliberating about it. We are no longer consciously aware of the goal, but still act in ways compatible with achieving it.

Sometimes, the goal or intention is activated automatically without ever having been conscious in the first place (Bargh et al 2001). These cases are interesting because it turns out that even though these goals were not deliberate, once we become aware of the fact that we are acting on them, we can consciously choose a new goal and alter the behavior. A frequently investigated instance of this is spontaneous trait inferences (STI's), made on the basis of very quick visual judgments of other people, on which we rely disproportionately often when something has threatened our identity or self-esteem. The restoration of self-esteem through the use of STI's is considered an automatically invoked goal, one of which we are not consciously aware. STI's usually involve the activation and use of a social stereotype that is judged to apply to the person in question because of immediately available visual information. Categories for such stereotypes include age, gender, and race. Subjects most often make STI's when they are not paying direct attention to how they are judging and treating others, and when something has occurred which undermines their own self-esteem or self-identity.[43]

The goal or intention to restore the sense of self through use of STI's is automatically invoked and not a goal we are aware of acting on (indeed, it is notorious that subjects making STI's claim they are not doing so). Nor was it formerly an intentional goal which gradually became automatized – it was only ever an automatic goal in the past. But in (Devine, Monteith, Zuwerink, and Elliot, 1991), subjects who were made aware of what was happening – who became informed that they were in fact making STI's – and who were motivated to not act in response to these stereotypes, namely, to avoid sexism, racism, and ageism, were able to

[43] See, for instance, Newman and Uleman (1989), and especially Greenwald and Banaji (1995).

interrupt their use of STI's in judging other people, and thus eliminate or at least reduce the impact of these biases in their behavior towards others. Knowing that they were automatically doing this, and having the deliberate goal to act otherwise, is sufficient for the automatic behavior – STI's manifested in treatment of others – to be derailed.

This is interesting because it indicates that the original and otherwise automatic intention can be interrupted or changed, and it can be interrupted by the subjects having a consciously held intention: namely, to not engage in or use STI's. The switch from automatic intentions to conscious ones may result in a circumvention of the prejudicial behavior. Since we respond differentially in our treatment of others in these similar situations when we are acting based on a goal of which we either are or are not consciously aware, we have another candidate variable to represent conscious causal involvement. This variable would be something like Conscious held goal or intention [yes, no].

The final avenue of conscious causal involvement I will argue for relates to the manner in which an action is performed. As we saw with the case of driving, some actions can be executed in an automatic fashion. Experienced drivers no longer have to consciously look in the mirror to keep track of traffic flow; they simply do this without thinking. When I type, I no longer need to look at my fingers or at the letters on the keyboard, although I could if I wanted to. The manner of execution of the typing is automatic. This is also seen in the case of musical performances.

However, even though many actions can be performed in either an automatic or a conscious fashion, the manner in which these same actions are executed will differ between the two cases. One experiment (Strayer, Drews, and Johnston 2003) considered people engaged in a driving simulation, some of who only concentrated on driving, and some of whom were then also involved in an engaging telephone conversation at the same time. The first group was thus

driving with conscious attention; the second group was driving automatically and with attention directed at the phone conversation. The results were, anecdotally, not surprising in the least: drivers who were driving automatically more often failed to notice traffic signs, and failed to respond sufficiently quickly to changes in traffic that required fast reactions. While the action of driving itself can be automatically or consciously executed, there are marked differences in these performances: the automatic performance is less responsive to new environmental information, less quick to react to changes in the task, and so on. This lays the ground for the third variable representing conscious involvement: Conscious Execution [yes, no].

3.5 Three variables representing conscious involvement in action

To summarize where we have come so far: we have found three avenues by which conscious awareness can be causally involved in action, which allow for direct empirical verification of their causal contributions to agency. There are three key elements to my alternative proposal, each of which marks a distinct causal role for conscious awareness in action.

1. Conscious intentions or goals, which includes conscious awareness of the means by which they can be achieved.
2. Conscious perceptual awareness of relevant environmental information, including information about oneself in relation to the environment.
3. Conscious execution of action, including monitoring of ongoing processes until completion/goal is achieved.

It's important to note that these three avenues of causal involvement are conceptually distinct, and can occur in conjunction or in any combination of the above. Fully conscious action involves each of these three elements. When an action lacks at least one of these elements, it is partially conscious and partially automated. Only if an action lacks all three elements would it count as

one to which conscious awareness did not causally contribute. The level of automaticity of any given action can be adjudicated by how many of these elements are absent.[44] These variables include only behaviors which could be consciously interrupted under the right conditions.[45]

We are now fully set up to incorporate variables 1 through 3 into an interventionist framework in order to ascertain specifically whether or not these are causes, and if so, what they are causes of. The values they take will be contingent in the specifics on the particular experiment involving the variable. As such, we can incorporate more of these experiments using the weakest or least specific values for the potential cause variables, namely presence or absence.[46]

Here is the start of a response structure, where we have yet to fill in values for the effect. Each variable is set to take values so that all combinations of values for all three variables are included. As discussed in section 3.2, the variable for Action outcome will take a variety of values, depending again on the specifics of the experiment in question, but so long as there is a single consistent set of values that the Action outcome variable can take for any 2 given rows in the response structures, this is not problematic. We'll see how this plays out below.

---

[44] A more detailed look at these variables could also represent the level of automaticity of an action in terms of degrees for each of the three variables. This would mean that, instead of values of either yes or no, each variable could take a multiplicity of values ranging between completely conscious and completely automatized. Again, this simplification into binary values does not affect the point I want to make here.

[45] See the end of this section for a discussion of modal considerations of fully automatic actions and interruptibility, or, why interruptible and uninterruptible behaviors have a distinct underlying causal structure and should be considered in distinct response structures.

[46] Although there is not space to do so here, it is not problematic to further enrich this representation by incorporating further variables, by incorporating further values for these variables, or by representing values with percentages instead of binary yes/no.

Table 2: Response Structure for 3 Awareness Variables

	Conscious goal/intention	Perceptual awareness of relevant info	Conscious execution	Action outcome
1	Y	Y	Y	
2	Y	Y	N	
3	Y	N	Y	
4	Y	N	N	
5	N	Y	Y	
6	N	Y	N	
7	N	N	Y	
8	N	N	N	

We can now ask, what would count as a row in this structure – what kind of situation would involve these three variables taking this particular combination of values? We choose a variable to investigate as a potential cause. The next step is to find two rows, where the values for the variable of investigation change, and the values for the other two variables are held fixed, and where the experiments or situations that instantiate each of the two rows have the same specific action outcome variable. There will be, in the example above, multiple such pairs of rows, with the other variables held fixed at different values for each pair. Finally, we simply fill in the blanks with experiments and look to see if the value for the action variable changes also. By doing this, we will be able to ascertain if each of the variables meets the criteria for counting as a cause of that effect.

Some of these rows have intuitively clear cases that count as instantiating those variable value combinations, where we needn't even find experiments in order to fill in the rows. To see whether perceptual awareness has a causal effect on the performance outcome, we can compare rows 1 and 3. In these rows, there is a conscious goal, and a consciously executed action, but in one case the subject has conscious perceptual information relevant to the task and in the other

she does not. It is not hard to see that the action outcome will differ in these two cases. Let's consider two comparison cases of subjects seated at a table, instructed to reach for a coffee cup in front of them. In one case, the subjects can see the cup; in the other, the subjects are blindfolded, lacking the relevant perceptual awareness. The action outcome can either be Reaches in the correct direction, or Makes correct hand shape as reaching. We would obviously see a marked change in performance between these two cases: subjects can easily reach for and grasp the cup correctly when perceptually aware of where it is. Subjects reach in the correct direction and assume the correct hand shape before reaching the cup; they do neither of these things in the case where they lack perceptual awareness.[47] This, and nothing more, is what is required to demonstrate that there is a test pair for Perceptual awareness in this system: that conscious awareness of the environment, perceptual information relevant to a task, is causally influential in behavior.

We can thus claim that the following, excerpted from the main response structure above, constitutes a test pair.[48]

---

[47] Someone might be concerned that blindsight subjects provide a counterexample to this otherwise straightforward claim. Blindsight subjects are capable of better-than-chance performances on tasks where they lack conscious perceptual information, making blindsight an intriguing phenomenon for understanding conscious and nonconscious action. However, it is easy to overestimate the importance of blindsight studies. Such subjects do not perform nearly as well as regularly-sighted subjects with perceptual information, so that even if we were to use blindsight performance in row 3, we would still see a difference in the outcome. The only thing which would need to be changed about the analysis as here provided is that instead of using yes and no as the outcome values, we should use percentages instead.

[48] Note that while I have simplified the results in the Action column to be binary, rather than involving a change in percentage as a more accurate representation would, this has no bearing on the present argument.

Table 3: Test Pair for Perceptual Awareness

	Conscious goal/intention	Perceptual awareness of relevant info	Conscious execution	Reaches in Correct Direction [yes, no]
1	Y	Y	Y	Y
3	Y	N	Y	N

This is sufficient to demonstrate that Perceptual awareness is a causal influence on action, and provides information about some circumstances under which it exerts such influence.

It's worth bearing in mind that for some potential test pairs, we may find that some variable does not count as a cause – in other words, the pair of rows is not a test pair. While this is usually an interesting result, the task of ascertaining whether or not there is a test pair is not completed. We should then check for test pairs using other values of the other variables, using other rows to see if it is possible to locate a test pair. Those variables will still be held fixed, but will be held fixed at different values.

To find a test pair for Conscious execution, consider the example from the previous section where subjects first had their self-identity or self-esteem threatened, and then responded automatically by using spontaneous trait inferences (STI's) in their judgments of other people, which affected the way they treated others. In this case, the value for Conscious intentions or goals is no, the value for Conscious perceptual information is yes (subjects make STI's on the basis of immediately available visual information of which they are aware, such as age, gender, and race), and the value for Conscious execution is no. This can be contrasted with the further research wherein subjects who were made aware of the fact that they were using STI's in their judgments and who did not wish to perpetuate prejudicial stereotypes could, by consciously intending not to be prejudiced, manage to avoid STI's. In this case, the value for Conscious goals

or intentions is now yes instead of no. The value for perceptual information is fixed at yes, and the value of Conscious execution is no.

This last point requires a little justification. While subjects were able to avoid using STI's, these STI's never did become fully conscious, and so the manner in which the intention to avoid STI's was enacted did not have to do with subjects deliberately avoiding them, just as the subjects did not deliberately implement STI's in their actions towards others. There is every reason to think that, with the same self-threatening priming, the STI's were automatically activated, and then not implemented in action. This means that the comparison between the two cases, acting on STI's and not acting on STI's, is one where two variables are held constant, the value for Conscious intentions is wiggled, and there is a change seen in the effect variable of Acts on Stereotypes [yes, no].

Table 4: Test Pair for Conscious Goal/Intention

	Conscious goal/intention	**Perceptual awareness of relevant info**	**Conscious execution**	**Acts on Stereotypes**
2	Y	Y	N	N
6	N	Y	N	Y

Rows 2 and 6 constitute a test pair for Conscious goals or intentions.

The Conscious intentions or goals variable thus has a test pair in the response structure, as does Perceptual awareness. Both of these variables are causes of the action outcome variable, where the values for that effect variable differ between the test pairs, but are the same within each test pair. What of Conscious execution? Conscious execution as a variable has a variety of experiments that contrast consciously executed actions with non-consciously executed, or automatic, ones. A few brief examples will suffice to demonstrate that there is no difficulty in finding test pairs for Conscious execution also.

Automatic behaviors that used to be consciously executed are a straightforward and easy way to demonstrate that there are test pairs for Conscious execution as a cause of the action outcome variable. Several of these have already been introduced in the earlier part of this section on automatism. Consider the case of automatic versus attentive driving. In both cases, Conscious goal has a value of yes and Perceptual awareness has a value of yes. In the automatic case, conscious execution has a value of no, whereas in the attentive driving case it has a value of yes. Automatic versus attentive driving thus instantiates rows 1 and 2, and if there is an effect on the action outcome variable, they constitute a test pair for the variable Conscious Execution.

And, unsurprisingly, there are numerous ways in which conscious versus automatic execution of actions like driving has an effect on the performance of the action, and on performance of other actions as well. Recall the study from the previous section demonstrating that attentive performance of driving allows for faster response times to sudden occurrences requiring reaction as compared to automatic driving while engaged in a phone conversation. In contrast, automatic driving allows for better performance on *other* tasks performed simultaneously than does attentive driving. This illustrates what is called the limited supply thesis of Bargh and Chantrand (1999): we are only capable of so much conscious involvement, and so whenever one task can be performed automatically, other tasks can be performed in addition, or can be better performed. This means that attentive performance of one task has a causal influence on the performance of a different task.

Table 5: Test Pair for Conscious Execution

	Conscious goal/intention	**Perceptual awareness of relevant info**	**Conscious execution**	**Responds to traffic signs**
1	Y	Y	Y	Y
2	Y	Y	N	N

It's also worth noting that, even when all three variables have the value 'no', i.e. when the action is fully automated, there is still a modal consideration regarding potential conscious involvement. For fully automated actions that became automated as a result of ongoing practice in similar situations, a particular performance of an action may be entirely automatic, and yet that action will still differ from another entirely unconscious one because the former could be consciously interrupted and performed in a different manner, whereas the latter could not. This potential for interruption is essentially the potential for an intervention, namely an intervention on the value for some variable representing conscious awareness. By changing the value of this variable, from automated to conscious, there is an effect on the action in how it is performed (in the case of interruptible behaviors, that it is inhibited). In contrast, for actions that are unconscious but cannot be interrupted with any amount of conscious effort, like the stimulus priming case, there is a different causal structure involved in the performance of the action: there are no possible interventions on conscious awareness, and thus no means by which to have an effect on the performance. Even though two actions may have the same values for all of the variables, they may exhibit a different set of causal relationships between those variables.

This difference in causal structure for interruptible and uninterruptible automatic actions helps us get a grip on the causal role of conscious awareness in action. In the case where conscious interruption is possible, conscious awareness has a causal role, whether or not it is exercised on any given occasion. The variable representing that facet of conscious involvement is causally connected to action, regardless of the value it takes. In the case where no conscious interruption is possible, the causal structure itself is different: there is no arrow from a variable for conscious awareness to the behavior in question. In the first case there must be such an arrow, since intervention on the variable has an effect on the performance. Even though two

behaviors may both be performed on a given occasion in an entirely automated way, there can still be an underlying difference in causal structure between the two, which manifests in this difference under manipulation. This is an example of how this approach provides an empirically enriched view of agency compared to views that make claims about conscious agency in general.

This discussion provides the materials to fill in the effect variable column of the response structure seen earlier. Not all of these rows have been filled in here. Each variable under consideration, however, has at least one test pair which *is* filled in: each has met the criteria for counting as a cause of the action outcome variable. This demonstrates, according to the interventionist approach and current scientific research, that each of the three awareness variables has a causal effect on action.

3.6 The causal efficacy of conscious awareness

The conclusion I have reached here is in stark contrast to other claims about the causal inefficacy of conscious agency (again, see Libet, Wegner, and Gazzaniga for examples of such broad claims), in spite of the fact that I have considered some of the same kinds of experiments as these authors (particularly Wegner). This means that the difference between their conclusion and mine rests primarily on the inferences made from the actual research to such conclusions; my approach is the only one that explicitly addresses this inferential question. This means that my approach to incorporating scientific results into a causal structure for conscious agency is stronger and better justified in at least two regards.

First, instead of utilizing single experiments as evidence for the non-involvement of conscious awareness in action, this approach requires that at least two different experiments be

compared, and that the differences between the two substantiate the claim of causation. While these authors certainly do offer more than one experiment in support of their claims, they offer these experiments as *individually* counting as evidence against conscious causal influence. They can thus stack up piles of single experiments without providing a contrast class against which to adjudicate the significance of these experiments. Such an approach does not meet the evidentiary standards for demonstrating that something is not a cause (indeed, it doesn't even recognize the issue of appropriate evidentiary standards for making such causal claims).

Second, I am making a rather detailed and specific causal claim, not an overarching one about what conscious agency in general does, or that we are always consciously involved in our own actions. Many of the philosophical positions that hold conscious awareness to be rarely or never causally efficacious in action, and offer specific experiments as evidence for this claim, are really responding to a straw man argument. They implicitly argue against some kind of absolutist position, according to which conscious awareness is *always* causally involved in action or is the *sole* cause of action. It is only if one is arguing against a position like this that it makes sense to offer individual counterexamples where conscious awareness is not involved in action, or is not the sole cause.

The approach I develop does not make universal claims about *the* causal role of conscious awareness. There may well be nothing substantive and true that can be said generally of all conscious and nonconscious action. Instead, I offer distinct, individually less significant but as a result better-substantiated, causal roles for different specific elements of conscious awareness that are involved in action. Experiments showing that conscious awareness is not involved in the execution of some particular task fail to provide evidence of a general lack of conscious causal involvement in action. Quite the contrary, as I have shown above, those kinds

of experiments are actually part of the evidentiary base for establishing that there is a causal difference between conscious and nonconscious action.

The detailed and piecemeal approach to demonstrating a constrained but significant causal role for some particular aspect of conscious awareness thus accomplishes several things. Instead of starting with a grand, strawman claim of broad generality to the effect that conscious awareness is always causally efficacious, which is easily undermined with a single counterexample, I work in the other direction, demonstrating that under specific circumstances, conscious awareness is causally efficacious. By conjoining more such detailed and piecemeal demonstrations of causal efficacy, we can develop an entirely more useful picture of the causal structure of conscious agency, one that is fleshed out in detail and does not ignore relevant differences between kinds of actions. The demonstrations in this chapter have only managed to cover a small area. But these demonstrations illustrate how to approach the scientific study of conscious agency in a new fashion, providing a general method such that other experiments and variables that I have not addressed here could nevertheless receive the same treatment.

By making explicit the need for test pairs, and thus for a wider range of experiments to compare against one another in order to ascertain whether or not a variable involving conscious awareness is a cause, the range of relevant experimental results expands to cover a great deal more than when the focus is solely on the sense of agency (recall chapter 2). The upshot of this is that there turns out to be ample evidence that conscious awareness is integrally involved in action. This expanded range of relevant evidence opens the opportunity to add complexity and structure to our understanding of the causal processes leading to action and involving awareness: details about the conditions under which awareness must be involved, could be involved, or could not be involved, how awareness fades out of certain actions over time and what allows it to

do so, how those actions differ as a result of decreased involvement of awareness and under what conditions awareness can still intervene, and behaviors in which awareness is never consciously involved and what precludes it from being so.

The conclusion of this chapter is that there is a variety of empirical evidence that supports the view that conscious awareness is causally involved in human agency in at least several specific ways. When this evidence is considered carefully in terms of a well-established and widely-accepted method of causal analysis, it can be systematized in terms of several variables that are at a mid-level of generality – neither as specific as the experiments themselves, nor as general as to take in all of conscious agency at a go. This treatment with the tools of interventionism was intended to be dual: both to establish the fact that the causal contributions of conscious agency are empirically supported, and to demonstrate some broader epistemic considerations by which such an interventionist treatment could be applied to other issues as well. The metaphysically deflationary but empirically enriched thesis of this dissertation is that according to our best currently available research and methods of causal analysis, there are interesting and substantive roles for conscious awareness in human action, and that these roles are more or less of the same kind as what we would naively expect. This chapter establishes the 'empirically enriched' portion of that thesis.

CHAPTER 4: Causation and counterfactual robustness

I have distinguished two dimensions along with causal questions can be understood: a metaphysical dimension, along which we ask what causation is and what kinds of relationships qualify as causal; and an empirical dimension, along which we address the actual causal relationships that exist in the world. I have just argued that empirically, we have substantive reasons to think that conscious awareness plays at least several important causal roles in action. This means that concerns about whether or not conscious agency is genuinely causally efficacious are not predicated primarily on empirical research on agency. Rather, these are concerns that, in one way or another, turn on metaphysical considerations of what ought to count as a cause: namely, whether higher level causes can be genuine causes, or if all genuine causation instead occurs at lower levels.

I've already shown that, given our best methods of causal analysis, conscious awareness has an empirically verifiable role in action. I will now demonstrate that higher levels causes are not metaphysically suspect, nor merely supervenient on and causally derivative from lower level causes. We ought to ascribe independent causal force to many higher level causes. This chapter and the next provide the foundation of causal metaphysics for the empirical analysis in chapter 3.

Briefly, the general problem is that conscious awareness seems to be a 'higher level' feature of the world, and as such, is composed out of, instantiated by, or at least supervenient on 'lower level' features of the world. These lower level features are themselves causally

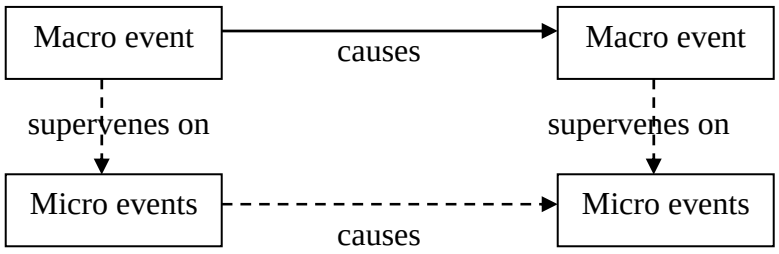
efficacious, and the higher level features (which include but are not limited to elements of conscious awareness) are in one way or another dependent on those lower level features. As such, there seems to be no causal efficacy for the higher level features as such, but only for the lower level features on which they depend. If this were the case, all conscious agency would be merely apparent, entirely dependent on law-like interactions between microparticles. The most pressing and intuitively compelling version of this problem is the Causal Exclusion problem. The intuition on which Causal Exclusion rests is that of microphysicalism, which this chapter debunks.

4.1 Microphysicalism and causal explanation

Microphysicalism is the view that genuine causal efficacy is only to be found at the microphysical level, so that the apparent causal efficacy of higher-level, macrophysical events is at best derivative and at worst completely epiphenomenal. Microphysicalism, alternatively called causal fundamentalism (Schaffer 2003, Woodward 2008), or more colorfully, microbangings (Ross, Ladyman, and Collier 2007), is based on intuitions that are both widespread and *prima facie* compelling (see Kim 1998). Even advocates of the causal relevance of higher-level explanations, including Jackson and Pettit (1991), Pettit (1995a and 1995b), Salmon (1984), share microphysicalist leanings: when both micro and macro explanations are available, we are to prefer the former as providing the genuine causal story. In this chapter I argue that for some kinds of phenomena (namely, counterfactually robust ones) this is false: macro explanations have more explanatory power than micro ones, and because of its failure vis-à-vis such phenomena, microphysicalism as a general stance is false.

We can understand microphysicalism as entailing a causal hierarchy of ontic precedence, where the further down in the hierarchy a cause is, the more ontically genuine it is as a cause. This entails that causal explanations following the solid line in figure 5 are merely temporary stand-ins for causal explanations following the dotted lines.

Figure 5: Macro/micro causal and supervenient relationships



The solid line indicates the causal relationship originally used to explain the macrophysical effect on the right hand side: it was caused by the macro event on the left. However, we (microphysicalist or no) know that the higher-level macrophysical event[49] cited in this explanation supervenes on some set of lower-level micro events, and we know the same about the macro effect. We know that there must be some kind of causal relationship between the micro cause and effect. The central microphysical assumption I criticize in this chapter is the view that causal explanations citing the dotted lines are more genuinely explanatory, because they identify the 'real' causes, than causal explanations that cite the solid line.

By beginning with an evaluation of causal explanation in the context of a strictly physical system, I can avoid many entrenched philosophical intuitions about what the correct answer should look like for mental causation. I will develop parallels between higher-level causes in

[49] I am here addressing individual events, not types of events.

physics, such as temperature, and higher-level causes in philosophy of mind and psychology, such as conscious awareness, to show how insights about causation in physics can illuminate causation in the case of conscious awareness. In particular, I will utilize the representational framework of phase space to explicate the problem of higher-level causation and explanation with respect to thermodynamics and statistical mechanics, and then demonstrate how this framework can be used to explicate the parallel problems for psychological causation. This chapter is thus intended to clarify the role of mental events in causal explanations, and to explain why explanations invoking mental events cannot or should not be superseded by explanations that instead appeal to the microphysical supervenience base of those events. While I will not explore them here, these arguments also have implications for higher-level causes in other disciplines as well.[50]

Approaching the problem of microphysicalism from the perspective of causal explanation has the advantage of making explicit the need to keep track of a single explanandum when comparing causal explanations. Part of the intuitive appeal of problems such as causal exclusion, I'll argue, rests on a subtle conflation of explananda. It's well-known that pragmatic considerations may sometimes lead us to rely on higher-level causal explanations, because they are more convenient to find. However, causal explanations can provide more than merely epistemic or pragmatic means of comparing two explanations. Explanations that are specifically *causal* explain an effect by identifying a causal relationship between the effect and the cause of (or one or more of the causes of) that effect. It is the existence of a causal relationship in the

[50] Higher-level causes that are counterfactually robust, regardless of whether they are psychological, are susceptible to the kind of analysis I provide in this paper. This has real implications for debates about biological causes and their relationship to biochemical causes, and for the motivations of reduction throughout neuroscience (see, e.g. Bickle 2003).

world that provides the explanatory force. For this reason, examining causal explanations provides a handle by which to compare the causal relationships posited by two competing explanations of the same phenomenon, to see if one is genuinely derivative of the other (as per microphysicalism).

In 4.2, I develop an example from the science of thermodynamics to illustrate the notion of counterfactual robustness as it applies to explanations, and to point to the existence of counterfactual "buffer zones" around single instances of causation. I then move from thermodynamics to the case of psychological explanations in 4.3, showing how the same "phase space framework" developed in section 2 – an analytic framework for representing the relationship between macrophysical events and their microphysical supervenience bases – can be used to clarify the relation between micro and macro explanations. Section 4.4 demonstrates how a presumed lower-level causal explanation of a macrophysical event actually changes the explanandum: the lower-level explanation does not have the same relata as does the higher-level causal explanation, and so cannot be a replacement for it. The criticism in this section demonstrates that the very formulation of Kim's problem of causal exclusion is mistaken, and that psychological events are causes *qua* psychological events, not merely *qua* their microphysical supervenience bases. 4.5 specifies two ways in which the counterfactually fragile explanations of the microphysicalist approach are explanatorily inferior to the counterfactually robust ones of the alternative view proffered here. Section 4.6 concludes.

4.2 Counterfactual robustness

The notions of counterfactual robustness and counterfactual buffer zones are critical to my demonstration of the legitimacy of higher-level causal explanations. I will introduce these new notions by examining the case of the thermodynamic vs. statistical mechanical analyses of the temperature of a gas in a box. This is a well-known example of higher and lower levels from an entirely different arena than that of mental causation. Physics provides a very useful framework, that of phase space, for describing the same problem of causal explanation at higher versus lower levels as we encounter in the case of mental causation.

Consider a volume of gas in a rigid, perfectly insulated box. The gas has a number of macrophysical features, including temperature. It also has microphysical features, such as the precise location and momentum of the particles of the gas. Any given macrostate, such as having the temperature 70 degrees Fahrenheit, is compatible with a large number of underlying microstates. Many different microstates, where the particles have different locations and momenta (momentum is simply mass times velocity, where mass is constant), could all give rise to the same macrostate of being at the temperature 70 degrees, so long as those microstates all have the same mean kinetic energy. As long as the average of all the velocities is the same, there can be a great deal of variation in how fast each individual particle is traveling. Every actual macrostate of the box is instantiated at any given moment by one and only one microstate, but we don't know which microstate that is, based only on knowing the macrostate: the macrostate only allows us to narrow down the possible microstates to a given range. This box thus provides

a standard example of weak supervenience,[51] where the temperature supervenes on particle locations, masses and velocities. There could not be a change in the temperature without there being a change in the microstate – in order to go to 69 or 71 degrees, the microstate must change in such a way as to alter the mean kinetic energy of the particles in the box.

The possible microstates of the box of gas can be described in terms of phase space. Phase space is a means of representing the exact microstate of such a system. Each particle of gas in the box has 6 degrees of freedom, three to specify location (in three dimensional space) and three degrees to specify momentum (the product of mass, which is constant, and velocity, which can vary in three dimensions). The dimensions of phase space for the box of gas will be 6N, where N is the number of particles in the box of gas. Each single point in phase space, then, represents one unique microstate, completely specifying the location and momentum of each particle of gas in the box.

Phase space also provides a way to represent possible macrostates. Rather than being individual points in phase space, as microstates are, macrostates are represented by regions or volumes of phase space. The volume that represents a macrostate such as being at 70 degrees will include all of the points in phase space that correspond to microstates giving rise to a temperature of 70 degrees, and it will include none of the points for microstates that give rise to temperatures higher or lower than 70 degrees.

Using phase space for macro and microstates, we can now explore some explanatory issues related to these two levels. Let's say we open a small hole in the side of the box and allow it to radiate. This is called black body radiation, and is given off by all bodies that have a

---

[51] This refers to the most basic definition of weak supervenience (see McLaughlin and Bennett 2008), where there cannot be a change in the higher-level event or property without a change in some lower-level event or property.

temperature. The box's internal temperature determines the frequency of the radiation: the higher the temperature, the more energy the radiation will have, which means a higher frequency. We can ask what caused the emitted radiation to be at one frequency rather than another. The answer to this question cites the temperature of the gas just before the hole was opened. In this situation, we produce a causal explanation of a macro or higher-level feature of the world – why the radiation had the particular frequency that it did – by adverting to another macro feature, the temperature inside the box.

But the temperature at the earlier time does not completely specify the microstate of the box prior to the hole being opened, since the microstate is underdetermined even though the macrostate is precisely defined – the microstate could have been any one of a number of possibilities, without having a bearing on the macrostate. If we think that genuine causal productiveness must be located at the microphysical level, then we are forced to count the original explanation of the macro phenomena of the radiation frequency in terms of the prior macrostate of the gas only as a stand-in or *prima facie* explanation. According to microphysicalism, it fails to cite a genuine cause, and is merely pragmatically useful if we lack causal information about the micro level. Even if this micro-explanatory method fails to account for the relationship between macrophysical *variables* or types of causation, there is a common – and, as I am arguing, mistaken – presumption that it provides an explanation of individual *instances* of causation.

Let's cash out the microphysical explanatory hierarchy in terms of the thermodynamics example: our first explanation might describe the macrostate of the boxed gas in terms of its temperature, and how the gas being at this temperature, rather than hotter or cooler, leads to radiation of the frequency emitted by the box, rather than radiation at a higher or lower

frequency. On our first explanation, there is no invocation of microstates, since the causal explanatory force comes from the prior and subsequent macrostates, specifically, the relationship between gas temperature and radiation frequency. To find the microphysicalist causal explanation, recall Figure 5. Microphysicalism suggests that causal explanation starts with reference to the macro cause and effect: the temperature and frequency of the radiation. It would then have us redescribe those macrostates at the microphysical level, by having us demarcate the precise microstates, the precise points in phase space occupied by the gas before and after the hole was opened. Such an explanation would provide the individual and excruciatingly detailed histories of each particle: we would be given the path of each individual particle in the box, bouncing around on the inside in such a way that it turned out that only specific frequencies of blackbody radiation were observed just outside the hole in the box.[52] This constitutes a description of the trajectory of the system through phase space from the point corresponding to the microstate on which the macro cause supervened, to the point corresponding to the microstate on which the macro effect supervened.

Because macro states supervene on micro states, microphysicalism capitalizes on a strong intuitive pull towards the idea that substituting the micro for the macro in an explanation should be acceptable, even explanatorily preferable, since it would cite the 'genuine' lower-level causes, not the supervenient higher-level ones. But this intuition is compelling only because it fails to

[52] It is the relationship between being a single point versus a volume in phase space that is equivalent here to micro versus macro. Even though the radiation frequency is plausibly microphysical, it should be thought of as macrophysical in this setting, in that there are multiple specific microstates that involve radiation of the same frequency, so that each such distinct state corresponds to a single point in phase space, but the collection of points that all instantiate different specific occurrences of radiation of the same frequency is a volume of phase space.

keep track of what, precisely, is being explained in the macro and micro cases (this will be taken up again shortly).

This is a crucial point: when we seek an explanation of a macro phenomenon such as temperature, what we are looking for is an explanation of why a system is in a particular *region* or volume of phase space. This is not the same thing as seeking an explanation of why the system is at a particular *point* in phase space. In the case of temperature, it was not merely the fact that the gas was in a particular microstate that led to its radiating at a particular frequency. The gas could have been in a slightly different microstate, and the subsequent macrostate, the frequency, would have been unchanged. Had the initial microstate been slightly different, there would have been some difference in the final microstate as well, and thus in the micro explanation. But this can be true despite both macrostates, and the macro explanation, remaining unchanged. An explanation for the change from one point in phase space to another point in phase space is not the same as an explanation for the change between one volume of phase space and another volume of phase space. Furthermore, a property of the macrostate which required explanation is the fact that it is not susceptible to small changes in the microstate.

I propose the notion of counterfactual robustness to characterize the difference between explanations in terms of points of phase space and explanations in terms of volumes of phase space. The micro explanation is counterfactually fragile, while the macro explanation that microphysicalism would have it replace is counterfactually robust. If we take the macrostate at a given instance of time, and specify with complete precision the microstate that actually obtains at that moment in time as part of our explanation, then our explanation will concern no more than the relationship between two very exact microstates. This explanation is fragile, in the sense that any alteration in the first microstate will make the explanation no longer applicable. It is the first

microstate which explains the later one; a change in that initial microstate means the later state will also be different, so that the explanation citing the first microstate no longer applies – even though the initial microstates are both instantiations of the same macrostate. This kind of robustness is counterfactual because while one microstate did in fact occur, the fact that it was this microstate rather than another does not make a difference, either epistemically or, most importantly, ontically, to the effect to be explained. At the start of our experiment, had any of those other microstates obtained, our explanation of the end microstate by its causal relationship to the initial microstate would have to be developed from scratch. Under minor counterfactual changes in the initial state, the explanation does not survive – it is fragile under such changes.

The implications of minor counterfactual variation deserve to be carefully stated. When we seek an explanation of a single instance of a macro phenomenon, it is true that on the precise occasion of the event for which we seek an explanation, that macro phenomenon is instantiated by a single microstate; this is a very precisely defined and exactly delineated state at the micro level, whether or not we know which particular state it is. For the micro explanation, if some minute feature of the microstate had been different, then we would no longer have an explanation – even small perturbations in the exact microstate would alter both the explanans and the explanandum. However, the event for which we originally sought an explanation would *not* have been altered by many such small changes in the microstate. There are usually any number of small differences that could have occurred without altering the causal relationship that figures in the explanation of the macro instance: the state of the box would have still transitioned from the same initial and ending regions of phase space.

Even though these small changes did not actually occur, they could have – and the fact that the original event was counterfactually robust with respect to small microphysical changes is

also a significant part of that for which we seek an explanation. To avoid spurious precision, we must acknowledge that we originally sought an explanation for a relevantly macrophysical occurrence: had the microstate of that process been different than it actually was, within the bounds of the macro volume of phase space, nothing in the macro causal relationship, and thus the macro explanation, would have changed.[53]

This exemplifies what I will call a counterfactual buffer zone. This buffer zone includes the microstates that could have occurred without altering the macro cause or effect. It encompasses the regions of phase space from which, and to which, the system moved, including but not limited to the micro-trajectory that actually happened to occur. Thus, we would be wrong to assume that the macrostate for which we are seeking an explanation is identical to some given microstate. Instead, it is identical to that microstate *plus* the counterfactual buffer zone around it, encompassing trajectories which could have but did not occur, and whose occurrence would not have altered the causal relationship.

4.3 Psychological phenomena are counterfactually robust

Counterfactual robustness and counterfactual buffer zones are general notions with wide application. I will now show how mental causation can be understood as providing causal

[53] The argument for counterfactual robustness bears resemblance to the arguments offered by Woodward (2008) regarding the stability or sensitivity of causal generalizations and the range of conditions under which causal generalizations hold. While Woodward and I both reach a similar conclusion – that in some cases, higher level causal claims are more explanatory than lower level ones – we do so on the basis of different reasons, and the notions of scope and stability of generalizations that he utilizes are not the same as my notion of counterfactual robustness and fragility. Related to Woodward's view, and to mine, is the discussion on generalizations in Mitchell (1997), and Mitchell (2008a) and (2008b).

explanations with counterfactual buffer zones for counterfactually robust events. The explanandum and the explanans in psychological explanations are counterfactually robust.

A quick word is in order in anticipation of a possible concern. There are disanalogies between the physics of gases and the study of conscious agency. Perhaps most striking is the lack of an account of how the two levels relate in the case of awareness and neuronal or particle activity. Physics has not completely reduced thermodynamics to statistical mechanics, and it may not be possible to ever do so (Callender 1999, Wook-Yi 2003).[54] Yet we can nevertheless give some kind of positive account of the relationship between them, in terms of the mean kinetic energy of the particles. No analogous account exists as yet for the relation between awareness and neuronal activity.

However, this difference of a known versus unknown interlevel relationship doesn't change the implications of my argument. If anything, it strengthens the criticism: if we can't substitute micro for macro explanations in cases where we have a relatively clear notion of the interlevel relationship, we will fare so much the worse when we try to do so in the case where we lack a good account of the interlevel relationship. All that is needed is a representation of the phenomenon in question in terms of points versus volumes of phase space. We need not be able to say what the specific relationship is between levels in order to nevertheless claim that there is some kind of supervenient relationship at hand, which is sufficient to distinguish between a higher and a lower level. The temperature of the gas is a property of the system at one level, the positions and momenta of the particles of the gas are at another level; conscious features of agency, including decision-making, perceptual information-processing, and execution are at one

[54] Simply knowing the interlevel relationship, in terms of mean kinetic energy of the particles, is insufficient for a complete reduction.

level, the physical processes of the nervous system and sensory apparatus are at another level. In order for the criticisms in the next two sections to apply to both the box of gas and to mental causation, all we need is a way to understand the case of conscious agency in terms of volumes and points in phase space.

In fact, Jackson and Pettit (1990) are working with just such an assumption about the relationship between the higher and lower levels for mental causation. Jackson and Pettit attempt to reconcile microphysicalism with the apparent strength of causal explanations involving psychological events (e.g. decisions, consciously executed actions). In order to salvage explanatory force for psychological events, they draw a distinction between two kinds of causal explanations. Even though only microphysical causes are causally *efficacious* in bringing about their effects, they claim that there is another kind of causal explanatory relation, whereby causes do not actually produce their effects, but where the cause can nevertheless be relevant to the effect in a way that makes it useful for explaining that effect. Jackson and Pettit refer to this attenuated explanatory role as causal *relevance*. They state their "solution" to the explanatory problem posed by microphysicalism thus:

> The realization of the [causally relevant] property ensures – it would have been enough to have made it suitably probable – that a crucial productive property is realized and, in the circumstances, that the event, under a certain description, occurs. The property-instance does not figure in the productive process leading to the event but it more or less ensures that a property-instance which is required for that process does figure. A useful metaphor for describing the role of the property is to say that its realization programs for the appearance of the productive property and, under a certain description, for the event produced. (1990, 114)

The theory of program explanation is one way to express the difference between explanations that provide an account of a trajectory from one point in phase space to another (in other words, a micro causal explanation), which they call causally efficacious explanation, and explanations that provide an account of a system developing from one volume of phase space to another

volume of phase space (or, a macro causal explanation), which they label causally relevant explanation. Programming for, or guaranteeing the occurrence of one of the right microstates, is simply a matter of defining some volume of phase space.

This is one way, however awkward, to reserve some kind of explanatory role for the mental while accepting that only microphysical causes are genuinely causally efficacious. At the same time, Jackson and Pettit still accept the causal explanatory hierarchy of microphysicalism (Figure 5), according to which micro explanations are always to be preferred where both micro and macro explanations are available. Where Jackson and Pettit fall short is in failing to recognize the existence and relevance of counterfactual buffer zones around the causal trajectories. Because of this, they unnecessarily concede that all causal efficacy is at the microphysical level, and their notion of causal relevance is merely epistemic: higher level causes don't cause anything, they merely serve to inform us about lower level causes.

Those considerations noted, I turn back to the representation of psychological explanations in terms of phase space. The consciously made decision to raise one's arm on a given occasion will serve as a higher-level or macroscopic occurrence; the neuronal processes, or biochemical, or even microparticular, processes on which that event supervenes on this particular occasion will be the corresponding lower-level event(s). However we choose to describe the lower level, this choice will set the degrees of freedom for phase space. For instance, if we use the state of the microparticles that compose the brain, nervous system, and muscles for the subject at the time of making the decision, then a point in phase space will represent a complete specification at a single moment in time of each of the parameters for each particle (namely, position and momentum). If we instead choose to use neuronal processing as the lower level, then a point in phase space will represent the complete specification of the state of the entire

nervous system at one moment in time, by specifying the values for whatever variables we choose to use (such as firing or not firing for neurons). For simplicity, and because it best characterizes the arguments for microphysicalism, I will use the exact state of all the microparticles as the basis for phase space.

Making a decision to move one's arm constitutes a region or volume of phase space, the parameters of which are set out by the microphysical supervenience base (the momenta and velocities of particles). The presumed effect, the subject's arm moving, is also represented by a distinct region of phase space. Each point in the region that represents the decision to move one's arm is a separate, complete, specification of a microstate of the subject at a moment. The subject can only occupy a single point at any given moment, and the trajectory through phase space represents the ongoing changes in the subject's state. Many different microstates can each be an instance of making a decision to move one's arm, just as many different microstates can be an instance of the box of gas having a temperature of 70 degrees. There are many small alterations in particle positions and momenta that would have no effect on the macrophysical phenomenon of making such a decision. Similarly for the effect: many microstates are compatible with the raising one's arm, which is why it is represented by a region of phase space.

We can translate these causal relationships in terms of movements within phase space. To say that making the decision to raise one's arm causes one's arm to go up amounts to saying that the points within the region of phase space representing that decision usually[55] evolve to subsequently occupy points in the region of phase space representing the moving of one's arm.

---

[55] This statement has a caveat to acknowledge that there are occasionally external factors that prevent such a decision from causing the arm to go up, such as paralysis or physical restraints. Such factors can also be represented in terms of phase space as barriers that prevent passage from one region of phase to another. This unnecessarily complicates the presentation without altering the conclusion, so I simply note it here.

There is something about being in that particular area of phase space such that the system subsequently evolves to be in another area of phase space.

The system starts out at a single point, and subsequently evolves to occupy a different point: this describes the time course of a subject first deciding to move her arm at some moment, and then subsequently moving that arm. The sequence of successive states for every single particle in the system is encoded in that trajectory. If even one particle had had a different location, or a slightly different momentum, at any instant of time, the trajectory itself would have been different.

Now we can compare micro versus macro explanations of why the subject's arm moved. A macro explanation of why the subject's arm moved on this particular occasion will cite the fact that the subject had previously decided to move her arm. Such an explanation could cite how systems within the region representing such decisions evolve to the region representing arms moving, but it need not do so explicitly. We certainly give macro causal explanations all the time without relying on phase space. What is important for these purposes is that these explanations *can* be represented in terms of phase space, whether or not they were originally couched in those terms. The explanatory power of such an explanation rests on the causal relationship between the two regions of phase space, and the fact that the actual trajectory that occurred was one such trajectory between the two regions.

A micro explanation translates the macro cause and effect into their respective microphysical states, describing the exact microstates of the subject at the beginning and end of the time frame in question. Such an explanation then cites the specific trajectory between these two points, and the explanatory power rests on the details of how the single trajectory moves through phase space, involving the exact particle interactions and movements that constituted the

precise microstate on this given occasion. Such an explanation is what the microphysicalist must endorse as providing the genuinely causal relationship while the macro explanation provides only a contingent and derivative causal relationship.

It should be clear now how the phase space framework allows us to understand the relationship between macro and micro explanations both for boxes of gas and for psychological events like a decision to move. Looking more closely at the latter kind of case, I will now demonstrate how, for single instances of causation,[56] the micro explanation subtly changes the explanandum in the switch from macro to micro level, and how it fails to accommodate the counterfactual robustness of the original explanandum.

4.4 Microcausal explanations alter explananda

As we've just seen, we can compare macro and micro explanations of events using phase space to represent the differences between these explanations. I will now argue that the micro explanation cannot serve as a straightforward replacement for the macro explanation because such a replacement forces us to *alter the explanandum* in causally significant ways. In brief, the micro explanation is restricted to explaining only what occurred, exactly as it occurred; in contrast, the macro explanation explains what actually occurred, *plus* the counterfactual buffer zone around that actual occurrence.

[56] My argument here addresses only single instances of causation and the explanations we provide of those single instances. Yet the same points hold *a fortiori* for type-level causation, including causation between variables.

Recall from the previous section that when an instance of causation occurs that could be explained with a macro explanation, that macro explanation covers not only what actually did occur on that occasion. It also explains the fact that other microstates could have occurred on that occasion without altering the cause or effect. The specific features of the initial microstate are irrelevant to what we want explained; part of the macro phenomenon is this indifference to which of numerous microstates actually instantiated it. Microphysical explanations provide a spurious kind of precision, one that actually obfuscates what we originally intended to explain.[57]

The fact that certain microphysical features could have been different, and yet the macro phenomenon could still have occurred, is part of what we originally sought to explain with the macro explanation, and it is something that no single micro explanation could capture. The explananda differ: one is a single trajectory through phase space, the other is movement from anywhere in a given region of phase space to anywhere within another, different, region of phase space. Every actual occurrence in the world of temperature is also an occurrence of a microstate which gives rise to that temperature, but this does not mean that it is a handful of microstates which we seek to explain when we look for the causal explanation between temperature and subsequent frequency of radiation. There is a coherent relationship between each of the microstates in that handful, and part of what we want to explain is not merely how a collection of initial states leads to a collection of subsequent states. The explanandum *includes* the fact that *the differences between microstates in the volume are not causally relevant* to the phenomenon being explained. Micro explanations fail to note that the single trajectory that did occur had a

[57] This is related to the argument on page 348 of Kitcher (1984), where he discusses how the "gory details" of microphysical explanation fail to meet some explanatory criteria (see also Kitcher 1991).

counterfactual buffer zone around it that is defined by including those trajectories that could have occurred without altering the cause or effect.

This applies both to the thermodynamics case and to the decision to move one's arm. The original macro explanandum doesn't include only the precise microstate that happened to occur. The other trajectories that also started and ended within the relevant regions of phase space, namely the counterfactual buffer zone, need to be included in the explanandum. This is because these microstates are indistinguishable and irrelevant, considered in terms of the original explanandum, to the decision to move one's arm and the consequent arm moving. When we are explaining a macro event, we need to cite the microstate it is actually in, plus the fact that this microstate is part of a collection of microstates that are related in just such a way as to lead to another collection of microstates which constitute the effect. If we fail to include this extra part, we have not explained what we originally set out to explain – we have merely redescribed what happened. The same holds for the case of the decision to move one's arm.

When we ask for a macro explanation involving higher-level quantities such as temperature, we seek to explain the exact trajectory that occurred, *plus* the counterfactual buffer zone around that actual occurrence. For this reason, causal explanations that translate a macro phenomenon into the microstates that constitute occurrences of that phenomenon are not explaining the same explanandum. In this regard, explanations at the macro-level are counterfactually robust, whereas explanations at the micro-level are counterfactually fragile. Counterfactually robust explanations cannot be replaced without loss by counterfactually fragile ones, in part because that which is being explained in either case differs. Explaining a macro phenomenon by explaining the micro phenomenon on which it actually supervened in a specific instance, without regard for that on which it could have supervened, misses something crucial.

This argument can be understood as the claim that psychological occurrences such as decisions to move one's arm are not token-identical with the microphysical occurrences that happen to instantiate each psychological occurrence. Instead, the higher-level potential cause – the decision to move one's arm – should be thought of as identical with the actual microphysical occurrences which instantiated them on given occasions *as well as* the counterfactual buffer zone around those actual occurrences, the zone where changes within it would not affect either the fact that there was a decision to move one's arm, or that one's arm then moved. This criticism entails that the explanatory power of macro explanations is not simply due to their pragmatic or epistemic convenience. Even if we had both a macro explanation of a macro phenomenon, as well as a micro explanation involving its supervenience base, we could not use the two interchangeably: they explain different things.

This argument also provides a new approach to the so-called causal exclusion problem (e.g., Jaegwon Kim 1998). Kim's formulation begs the question. We need not feel compelled to solve the causal exclusion problem, since it is predicated on a subtle conflation of two distinct explananda. The general form of the causal exclusion problem is well-known. "[T]he problem of causal exclusion is to answer this question: Given that every physical event that has a cause has a physical cause, how is a mental cause also possible?" (1998, 38). The problem is that any mention of mental causes appears to be unnecessary, since anything a mental event could cause already has a sufficient physical cause. Mental properties and events, in particular those that we attempt to capture with psychological predicates such as 'decision', supervene on physical events. The problem of causal exclusion is taken to entail that either mental causes are systematically overdeterminative, or the supervenience base does all the genuine causing, so that supervenient phenomena are causally inert.

According to my analysis, Kim has missed the mark when he simply substitutes the microsupervenience base for the macro event to be explained. If there were nothing more to the macro event than the token of the microsupervenience base on any particular occasion, then the causal exclusion problem would be extremely serious. However, given that the microsupervenience base leaves out a very important part of the original explanandum, it is not such a dire situation as we may have thought. Mental causes aren't overdeterminative *of the original explanandum*, however much they would be overdeterminative of the altered explanandum. Absent an argument to the effect that we ought to prefer the modified explanandum, the causal exclusion problem lacks teeth.

A consequence of one way of solving the causal exclusion problem leads to a new problem with higher-level causes, identified by a number of authors[58] and labeled 'quausation' by Horgan (1989).[59] One response to the causal exclusion problem is to accept at least token-token identity between higher-level macrophysical events such as decisions to move one's arm, and the lower level microphysical events on which such decisions supervene (usually, the particles that compose such processes and interact according to the laws of fundamental physics). This token-token identity seems to imply that a macrophysical event does cause some effect, but it does not do so *qua* macrophysical cause – it is a cause only by dint of the cumulative causal effect of the microphysical events which instantiate it. The concern is that something like a conscious decision could only causally influence anything *qua* its microphysical instantiation, and so while conscious decisions might plausibly be causes, they are not causes *qua* decisions, but only insofar as the microphysical instantiation is causal.

---

[58] See, for instance, McLaughlin (1989) and Hardcastle (1998).
[59] This same problem, under a different name, has also been raised by Pettit (1993) and Goodin and Smith (2006).

The argument presented in this section implies that it is not *qua* microphysical events that something like a conscious decision to move one's arm manages to be a cause, since that would only cover a fraction of the counterfactual buffer zone that requires explanation. It is actually *qua* macrophysical, *qua* region of phase space, that it has the causal effects that it does. The microphysical is causally efficacious in bringing *something* about, but not in bringing about the effect we originally set out to explain. The initial starting point for the trajectory can only be conceived of as a cause of the final point; it is unable to cause the entire buffer zone around the effect.

4.5 Further shortcomings of micro explanations

I've shown in 4.3 that replacing macro with micro explanations is equivalent to replacing counterfactually robust explanations with counterfactually fragile ones. In this section, I will demonstrate how, even when they share the same explanandum, replacement of counterfactually robust causal explanations with counterfactually fragile causal explanations results in the loss of much explanatory power. There are two primary criticisms I'll levy against such replacement. First, lower-level, counterfactually fragile explanations are unable to account for the connection between each of the trajectories from and to the relevant volumes of phase space; as a result, they fail to explain the fact that the macro phenomenon is counterfactually robust. Second, the proposed micro explanation must necessarily include extraneous causal information, information about processes or interactions that did not causally contribute to the effect.

My first criticism consists in the fact that counterfactually fragile explanations are unable to account for the relationship between the points in each region of phase space; consequently, they fail to explain the robustness of the macro phenomenon.

Essentially, the problem is that part of what an explanation ought to provide is an account of the relationship between the points in the regions of phase space that constitute the cause and effect; micro explanations are unable to do so, while macro explanations can. Each point within the region constituting the cause bears a relationship to other points in the same region, namely, that these points are ones between which the system can vary without altering the macro phenomenon in question. In the gas example, the relationship between distinct microstates that all give rise to the same macrostate can be cashed out specifically: the inter-point relationship between the various points in phase space is that of having the same mean kinetic energy. The macro causal explanation accommodates this relationship by reflecting the coherence of all these points as a macro event.

Micro causal explanations, on the other hand, must address a single point in the cause or effect region at a time. This involves trajectories that start and end within the relevant regions of phase space, but it does not address the relationship between the different points within each region, such that they are all part of the same region. The fact that these micro points together constitute a coherent macro phenomenon is ignored. It is true that an even better macro explanation would be able to provide information about the nature of the relationship between microstates for macro phenomena such as deciding to raise one's arm. Nevertheless, the macro explanations that we can currently provide of such events, even while lacking this information, are still more explanatory than their micro counterparts.

A defender of micro explanation may attempt to alleviate this problem by compounding single trajectories. The disjunction of trajectories from points within the region that constitutes the cause, to the points in the region that constitute the effect, may account for each individual trajectory that could occur within the scope of the macro explanation. Yet explanation via a disjunction of trajectories is not equivalent to explanation via a region of phase space.[60] Explanation via region accommodates the fact that small movements around a given point aren't causally relevant to the macro state – these points are related in such a way (whether or not we know what that relation is) as to give rise to the same macro state. In the case of disjoint trajectories, the points in the disjunction bear no relationship to one another – they have been put together but do not cohere.[61] A handful of threads are not a rope. If one wishes to add to such a disjunction some kind of characterization of how the points in the disjunction relate, in particular if one wishes to characterize the trajectories not by listing points one by one in a disjunctive fashion, but by providing a description of the area they are in, then one is no longer using a disjunction of trajectories. One would simply be using the region of phase space, a covert version of the macro explanation.

---

[60] A technical sidenote: if one is using a phase space representation with continuous, not discrete, variables, then the micro trajectories that could be compounded will always be at most denumerably infinite, while the points in any volume of phase space will be nondenumerably infinite. The volume of phase space that constitutes the cause or effect need not be completely contiguous – there may be several distinct subvolumes that collectively constitute the region of the case. The relationship between the points in these subvolumes, and the subvolumes themselves, is still incorporated in the macro explanation.

[61] This criticism against micro explanations is closely connected to the debate about laws in the special sciences (see, for instance, Fodor 1974). Those who argue against the reduction of such sciences to physics often claim that the only move open to reductionists is to use disjunctions (of laws or of natural kinds), which are not equivalent to what was supposed to be reduced (in that they are either not lawlike or no longer natural).

This inability to account for the relationship between points in regions of phase space means that counterfactually fragile explanations do not explain relevant features of the macro phenomenon. As I have already argued, the macro phenomenon in these cases is counterfactually robust: it includes the actual microstate that did occur plus the counterfactual buffer zone around that point. An adequate explanation of such a phenomenon needs to at least acknowledge this feature of the explanandum. Counterfactually fragile explanations lack the means to do so. They are unable to convey information about the counterfactual buffer zone because, by definition, they are confined to explaining the relationships between single points.

This is a different point than that of the previous section: it is not merely that the counterfactually fragile explanation changes the explanandum when it replaces a counterfactually robust explanation. The problem is that, even for that which the micro explanation does explain, it fails to explain enough about that explanandum. This means that, if we have both a counterfactually robust macro explanation and a counterfactually fragile explanation of the same robust event, the micro explanation would be an inferior explanation than the macro one would be. The explanatory advantage of macro explanations is not merely pragmatic or epistemic. Counterfactually fragile explanations should be deployed for events which were genuinely counterfactually fragile themselves; counterfactually robust explanations should be used for events which were counterfactually robust. In the philosophy of mind, this will usually translate into using macro explanations for counterfactually robust, psychological, events.

My second criticism is that micro explanations must include extraneous causal information, because they lack the means to ascertain which causal interactions were relevant to bringing about the effect and which merely occurred but did not assist in bringing about the

effect. Counterfactual buffer zones provide such means to ascertain the relevance of causal information. The process view of causation provided by Salmon (1998) and Dowe (2000) can be usefully applied to describe the time evolution of the trajectory through phase space. The movement from one point to another is equivalent to a small change in either the position or momentum of at least one particle in the system. Such changes come about because of conserved-quantity transferring interactions between the particles in the system. The quick and dirty way to understand these interactions is in terms of particles of gas bumping into each other or against the walls of the box: when particles run into each other, they change each other's velocities, and that change is a causal interaction. Thus, each movement from one point to another in phase space, each step along the trajectory of the system, constitutes causal interaction in the Salmon/Dowe sense, which is the sense of 'causal production' that microphysicalists claim as the only real kind of causal interaction.

What is crucial to the comparison of macro and micro explanations is that some of these causal interactions, some of the precise details of these micro trajectories, are not actually causally relevant to the production of the final state to be explained.[62] Given that a slightly different trajectory could have occurred, and the cause and effect nevertheless have remained unaffected, we know that on any given trajectory, some of the causal interactions in its time

---

[62] Here again it is useful to approach the issue of causation via the issue of causal explanation. Doing so forces us to keep in mind a fixed explanandum, and how well two different causal factors explain that explanandum (rather than, for instance, how 'causey' either causal factor seem to be). Thus, my point here is not merely an epistemic one about explanations, and how it is easier to find macro explanations than micro ones. Causal explanations cite causes, and we can use such explanations to track causes, even though explanation and causation can, under some circumstances, come apart. Rather, while it is not conclusive, even if we had both a micro and a macro causal explanation in hand, the fact that the macro one would be more explanatory of the explanandum at hand should tell against simply assuming that all causation is at the microlevel. There are causes at that level, certainly, but, if the explanations are any indicator, not causes *of the effect we wanted to explain*.

evolution are causally irrelevant to the actual end state to be explained. We do not necessarily know which ones were irrelevant, but we know that at least some of them must have been, since they could have occurred differently without affecting the explanandum.[63]

True, these interactions were genuinely causal, in the sense that they were causally relevant to *something*. But they were not causally relevant *to the explanandum at hand*. Not everything that is causal and occurs in the vicinity of these initial and end states will have causally participated in moving the system from the initial to the end state. Two particles which collided and rebounded in certain directions could have collided and rebounded in other directions, or not have collided at all, without the end state changing at all. That particular interaction, and others like it, would have to be monitored in order to provide the full micro explanation, but such an explanation would be blind to the fact that some of these interactions were unnecessary to the phenomenon being explained.

As such, in any single trajectory, we can trace out the causal interactions that take place, but if we are confined to the micro explanation, we have no means by which to ascertain which ones were genuinely necessary for the effect to have occurred, and which could have been altered or eliminated without altering or eliminating the effect. We are essentially including extraneous causal information in our explanation, causal relationships between particles that did not contribute to the causal relationship that figures in the explanation. Macro explanations, however, provide exactly the kind of means necessary to discriminate between causally relevant and irrelevant interactions, namely, the counterfactual buffer zone, the zone within which small changes have no effect on the explanandum.

---

[63] A related point is made in Craver (2007).

Thus, micro explanations are unable to account for the relationship between the points in phase space that constitute the regions of the cause and effect. They are also unable to account for the counterfactual robustness of the event they purport to explain, which means that even if we had both a macro and a micro explanation of the same event, the macro explanation would still be more explanatory of the explanandum. Finally, counterfactually fragile explanations must always include extraneous causal interactions because they lack the means to ascertain which ones were extraneous. It is the counterfactual buffer zone that provides just such grounds to determine which causal interactions were extraneous to the phenomenon to be explained.

4.6 Conclusion

I have argued that a consequence of microphysicalism for causal explanation, the causal explanatory hierarchy of Figure 5, is misguided. At least for counterfactually robust phenomena, the causal relationships provided by macro explanations can be more explanatory than those provided by micro explanations. There are a number of consequences of this failure of microphysicalism to accommodate causal explanation. The first is simply that we have reason to be skeptical of microphysicalism itself. While my arguments here are by no means decisive against microphysicalism, they are further evidence that it is an untenable position: the fact that microphysicalism's direct consequences for explanation are so wrong is indicative of a problem with microphysicalism.

The motivating problem for this analysis, microphysicalism, has posed problems not only for mental causation but also for causation in the so-called 'special sciences' (essentially, anything other than particle physics). While I have specifically addressed the cases of

temperature and of conscious agency, my analysis also applies to many other higher-level causes as well. Any cause that can be construed as 'higher-level', in that instances of these causes supervene on instances of something else (the lower level), will encounter the same problem, and be labeled causally superfluous by microphysicalism. However, by dint of this same supervenient relationship, these causes can be represented as regions of phase space, comprised of points in phase space representing microstates, such that the same kind of analysis applies. Many other macrophysical causes, especially causes that are particular to biology and psychology, will be higher-level by dint of supervening on the microphysical details of their individual instantiations. Thus, while the present argument focuses on psychological causes, the solution I offer in this paper should vindicate the possibility of genuine causal influence for higher-level causes more generally.

This chapter vindicates the role that psychological events already have in causal explanations. We have found ourselves, in matters of practice, unable to do without these kinds of explanations, and unable to actually replace them with lower-level explanations. I have argued that such a replacement, even if it were tractable, might not be a good goal: causal explanations involving higher-level features of the world can provide counterfactually robust causal explanations of counterfactually robust events. The hierarchy of causal explanation that is a consequence of microphysicalism is wrong, which provides reason to think that microphysicalism itself is ultimately misguided.

CHAPTER 5: Variables, levels, and downward causation

This chapter continues the project of providing a metaphysical foundation for the attribution of causal efficacy to higher level causes, including although not limited to conscious awareness. I have now addressed the issue of causation as a relation between single events, and demonstrated (chapter 4) that for counterfactually robust phenomena, the microphysical causal relationships are not equivalent to, nor substitutable for, macrophysical causal relationships. In this chapter, I will address three more concerns that one might have about the empirical approach I developed in chapter 3. First, one might be concerned that the choice of variables by which I represented the three avenues of conscious causal involvement in action begged the question, and that if other variables were utilized instead, the appearance of conscious causal efficacy would be undermined. Second, one might be concerned that while the variables representing awareness appear to causally efficacious, it is still the microphysical supervenience base of awareness that does the 'real' causing; awareness is not causal by dint of being awareness, but merely by dint of the interactions between particles. Finally, one might be concerned that the kind of causal efficacy I attribute to conscious awareness requires a metaphysically suspect kind of downward causation, where higher level causes like awareness must causally influence lower level causes such as neuronal activity.

This chapter addresses all three of these concerns. By taking a close look at what would be necessary to replace the higher level variables with lower level ones, I uncover a dilemma. If

such a replacement does not change the events included in the variables, then the lower level variables are simply a parasitic redescription of the higher level ones. If such a replacement does change the events included in the variables, then the original variables have not been reduced themselves. Instead, the phenomena being modeled has been parsed in a different way, and the new parsing must be shown on a case by case basis to be empirically superior to the original one. Either way, there is no basis from which to claim that a lower level series of replacement variables are either ontologically more fundamental or even that such replacement exists.

This leads into an exploration of the relationship between causation as a relation between singular events, and causation as a relation between variables that are collections of such events. I demonstrate an interdependence between singular and variable causation: variables causation is ontologically dependent on singular causation, yet singular causation is epistemically dependent on variable causation. This second direction of dependence provides grounds to show that when higher level causes such as awareness are causally efficacious, they are efficacious by dint of the features that make them members of variables like the three from chapter 3. This means that variables representing awareness are not merely efficacious by dint of their microphysical instantiations, but by dint of being instances of awareness.

I then consider the issue of downward causation. The existence of downward causation is the subject of considerable debate, and compelling arguments against it have been made. I acknowledge that there are versions of downward causation that are metaphysically incoherent, but that certain kinds of systems can display a metphysically viable version of downward causation. In sufficiently complex systems there is a kind of causal articulation among the parts of the system, such that entities or processes at a higher level could have a direct causal influence on entities or processes at a lower level, so long as those lower level entities or processes are not

in the supervenience base of the higher level cause. This demonstrates the metaphysical possibly of downward causation, and human agents certainly possess the relevant kinds of complexity to allow for such causal articulation.

Section 5.1 addresses the criticism that I have made these variables appear causal by failing to include the 'right' variables, either neuronal or microphysical ones. In section 5.2, I examine how general or variable level causation allows us to evaluate counterfactuals for single instances of causation, and how such evaluations allows us to hone in on what it is about an instance of causation that was efficacious. Section 5.3 develops a theory of causally articulated downward causation in complex systems. The conclusion in 5.4 wraps up metaphysical leg of the dissertation that includes chapters 4 and 5: we should feel confident that higher level causes, including but not limited to those involving conscious awareness, can be genuinely causal, and as such we have a solid metaphysical foundation for the empirical analysis in chapter 3.

5.1 Variables and levels

Recall the three variables I defended in chapter 3: consciously held goals or intentions, conscious perceptual information relevant to goals, and conscious execution of movement. A critic might take issue with my choice of variables: by including only these variables, rather than others, I have spuriously generated the appearance of causal influence but at the cost of the begging the question. The charge that I have failed to include other causally relevant variables might take two forms, depending on what one takes to be the 'right' variables. The first version is that I have failed to include a common cause of both the awareness variables and the action outcomes. If this charge were true, then some additional, presumably nonconscious, variable(s) would actually

be the cause of the awareness variables as well as the cause of the action outcome, and conditioning on that common cause would demonstrate that the awareness variables are not in fact causal. This criticism is equivalent to claiming that the system of variables I used in chapter 3 fails to satisfy Causal Sufficiency, and thus fails to satisfy the Causal Markov Condition, an important prerequisite for use of the interventionist analysis. This first version of the charge is briefly addressed in a footnote in chapter 3, and I will not take it up here.

The other version of this charge is that the variables I used are only stand-ins for other variables. These other variables are to be preferred, because they are more ontologically fundamental. This charge concerns the grain or level at which my analysis was pitched. I utilized fairly coarse-grained variables in my analysis, variables that were defined broadly to include a number of distinct sorts of events that are large-scale in size (we don't individuate instances of perceptual awareness by reference to microparticle interactions, for instance). Instead, a critic might argue, I ought to have utilized finer-grained variables: each such coarse variable should be replaced with multiple others which each represent finer-grained events and the causal relationships between them. These finer-grained variables would represent events that are measured at a smaller scale – conscious intentions would be replaced by variables for sensory input to various parts of the brain, variables for certain kinds of processing and the chain of influence between populations of neurons, culminating in firings of neurons to stimulate muscle movement of the right sort. The idea is that it is these finer-grained, smaller-scale variables which are 'really' doing the causing that I have spuriously attributed to the coarser-grained, larger-scale variables.

I will speak of fine grained and coarse grained variables, which should be understood in this particular context as also indicating a difference in level on other level breakdowns: it is not

merely that one variable might be subdivided into two or more individual variables. It is primarily that the new subdivided variables are *also* lower level with respect to the original coarse-grained variable: they might be at a smaller scale, they may be components of a mechanism for the higher level entities, and so on, depending on how the levels are parsed.

This section (5.1) addresses the issue of high-level versus low-level causation as it specifically relates to causation as a relationship between variables. Can the 'higher-level' variables I have defended be replaced with 'lower-level' variables? I argue that while instances of these variables can be redescribed in lower level terms, the *variables* themselves involve more than simply a list of instances and as such cannot be replaced; it is a mistake to think that the *variables* are somehow necessarily coarse-grained or higher level because they pick out their instances with a higher-level description

5.1.1 Variable extension and intension

It will be helpful to think of causal variables in terms of sets. Any given causal variable can be understood as a set that collects together the events in the world that count as instances of that variable. Instances of variables should thus be understood as events where the variable takes one of its possible values. Causation is primitively singular: single events cause other single events. What it is about a particular event, though, by dint of which it is able to cause another event, may occur on more than one occasion and in more than one kind of setting; that by which one event causes another can occur in multiple single instances of causation.[64] A causal variable can thus

[64] If this explication seems a bit cryptic, there is a longer and more detailed discussion of this notion in section 5.2 of this chapter.

be understood as a way of grouping single instances of causation together according to common features by which each of those single causes led to each of those single effects.[65]

There are two characteristics by which we individuate causal variables, by extension and by intension. The extension of a variable is simply a list of all the members of the set: a list of each individual event in the world that counts as an instance of that variable. The intension of the set provides the characterizing feature that distinguishes instances of the variable from all other causal events in the world. It is the rule by which we decide whether or not some new event is a member of the set (and can be used without having to consult the list of events constituting the extension). Generally the intension will be based on some feature events could have and which are common to all of the causal events in the world that are members of this set. It provides a handle on the set. We can use the intension to decide, when confronted with new causal events, whether or not they are members of the set, whether or not they are instances of the variable.

This set approach to variables allows us to compare variables to each other based on their intensions and extensions. We can easily see how two variables are different if their extensions differ – each variable represents a different set of events in the world. Variable extensions can also partially or completely overlap in interesting ways. One variable might have an extension that is a proper subset of another. These kinds of comparisons are useful because they allow us to keep track of certain logical relationships between variables that constrain the ways in which we

---

[65] There are several assumptions about causation baked into this to which I will simply help myself at this point. In particular, one might object to the kind of causal realism I assume here, where variables are not merely any kind of aggregate of instances, but must be collections the instances of which bear some kind of relation to one another. The issue of whether causal variables are, or could be, natural kinds in this fashion is extremely interesting and far too lengthy for me to treat here. I will instead refer the reader to Woodward (2003), and note that according to interventionism, causes are that on which we can intervene to bring about changes in the world. If a variable merely collects unrelated events, then one could not define an intervention on that variable.

can incorporate these variables into systems. In order to treat two variables as capable of potentially standing in a causal relationship, these two variables must be distinct, and we can ascertain whether or not they are distinct in this way by comparing their extensions.[66] If there are any instances that count as an instance of both variables, then the two variables cannot stand in a straightforward causal relationship.[67]

Alternatively, we can compare variables in terms of their intensions, the rules or guidelines by which we ascertain whether or not a new causal event counts as an instance of a variable. We could find that two variables have the same extension, but different intensions – the same events in the world count as instances of each variable, but do so in virtue of distinct features of those events. In such a case, we can imagine two possibilities regarding future events: it may turn out that having identical extensions so far has been a result of the fact that the two intensions are really picking out precisely the same features of events in the world – they are really redescribing the same events. In such a case, we would want to simply say that we only have one variable, described differently, rather than two genuinely different variables. On the other hand, we could find cases where the extensions of two variables have matched thus far – for all of the instances we've already found – but which could come apart under some conditions

---

[66] This idea has already appeared in the dissertation. A part of my criticism of Libet in chapter 1, and of other such as Daniel Wegner (see also Andersen 2006) turned on how these authors individuated variables in such a way as to have a partial overlap of extension for cause and effect. For instance, in Libet's experiments, conscious volition was treated as fully distinct from and capable of standing in a causal relationship with neuronal processing. However, unless we commit ourselves to dualism, we must accept that at least some of the neuronal processing being measured was also associated with conscious decision-making: some events which count as instances of the latter variable also count as instances of the former variable. Thus, a single event could occur that would end up counting both as cause and as its own effect. It is a very common stricture on causal relationships that this is unacceptable.

[67] There is work being done on how to incorporate partially overlapping variables into a single causal system by Peter Spirtes and others at CMU. My understanding is that this problem is quite complex and not yet resolved.

that have not yet been realized, a case analogous to underdetermination with respect to current evidence but not with respect to all evidence. Some future experiment, as yet unperformed, would confirm one over the other. In this sort of case, we'd want to say that these were two different variables with two different extensions, even though all we had thus far seen of their extensions had been identical.

This means we have two ways by which to characterize a variable, depending on our interests and also our epistemic situation: in terms of the events in the world which count as instances of the variable; and in terms of the feature of those events by virtue of which they are or would be members of this set, i.e. instances of the variable. This gives us a clearer way of expressing the way in which some variable may be finer or coarser grained than others, and it also gives us a way to evaluate the plausibility of suggestions that a coarser grained variable could or should be replaced with a finer grained one (where a finer grained variable is lower level with respect to a coarser one). Finer grained variables might have extensions that are proper subsets of the extensions of coarser grained variables, or they might have intensions that that pick out events at smaller size or time scales than do coarser variables.

Consider Conscious Perceptual Information as a variable from the previous chapter, and some variable representing activity in V1, the primary visual cortex. A critic of my approach in the last chapter might say that Conscious Perceptual Information is a coarser grained, or higher level, variable than the one for V1, on a number of understandings of 'level': processing in V1 is a component or enabler of conscious perceptual information, it is arguably at a smaller scale if we consider the former to be at the scale of a full person, it certainly occurs at smaller time scales, etc. The details of how we differentiate levels will turn out not to matter, so I gloss over it here. So long as we accept the assumption that, insofar as there is any sensible way to distinguish

higher and lower levels in this case,[68] conscious perceptual information is higher level than neuronal activity in V1.

The challenge I am considering here, then, is that variables like Conscious Perceptual Information can be, and ought to be, replaced by variables such as one involving neuronal activity in the visual cortex. There may be multiple variables required to replace a single coarser grained one, but the idea is that such replacement can be done, and that such replacement will be at least as good a representation of the causal system, although some would probably claim it was a superior representation of the causal system by dint of its being more basic or fundamental.

Can we make sense of the request to replace higher level variables representing conscious awareness with lower level ones representing neuronal activity? Several common intuitions appear to support an affirmative answer to this question, but I will argue that we should answer in the negative, and be wary of those common intuitions. If we are physicalists of some variety or another, we should at the very least be committed to the assumption that the variables representing conscious agency are physical variables, where the contrast class would be ghostly or ethereal variables, perhaps. Each event that counts as an instance of Conscious Perceptual Information is also a physical event (where physical events can have modal properties, see the next section). This is basic to physicalism, since we could phrase the physicalist commitment as holding that there are no nonphysical instances of causings – every cause is a physical cause. But it does not follow from the bare physicalist commitment that finer grained or lower level variables are automatically preferable to coarser or higher level ones, nor does it guarantee that

[68] Our method of level differentiation need not imply that there are somehow 'really' levels out there in the world. We can use pragmatic or epistemic considerations to differentiate levels and my arguments here will not be affected.

we could even make such replacements. The problem with such replacement becomes clear when we compare the intensions and extensions of the variables, as I'll now do.

The key point that precludes such replacement in many cases is the distinction between individual events in the world that are causes and effects of other single events in the world, and variables that represent general or type-level causation. Descriptions of individual *events* can be given in higher or lower level terminology, as part of macrophysical and microphysical theories. But *variables* differ from each other because of differences in extension. We must not mistake differences between descriptions of individual events for differences between variables. Once we get clear on this distinction, we'll see that we cannot simply replace higher level variables like Conscious Perceptual Information with lower levels ones like Activity in V1 (just a reminder: in the next section I will address the issue of higher and lower level causes as individual events; this section focuses only on the case of variables).

5.1.2 Replacing higher level variables, maintaining extension

I'll start by trying to get specific about how we would go about replacing one variable with another, lower level, one. Consider the collection of instances that count as instances of Conscious Perceptual Information. In the rich causal nexus of the world, these event-instances are picked out and grouped together in virtue of specific features. It is the possession of these features that allows them to stand in the causal relationships that they do, and to be grouped into variables according to these features. In order to replace this variable with another (or multiple such others), the replacement would need to have precisely the same extension – pick out exactly all and only those events picked out by Conscious Perceptual Information. This is so both for events that have already occurred, as well as for future events that have not yet occurred (and

thus not been classified). To accomplish the sorting of future events, the replacement variable(s) needs to provide means by which to ascertain whether or not a new event should count as an instance of the replacement, and must yield the same answer as the original variable would have. If the original variables would have labeled an event as an instance, so must the replacement variable. And, for this replacement of variables to count as providing a more legitimate causal representation, the replacement variable(s) must be finer grained, at a lower level, than the original.

To be more concrete, if we want to replace Conscious Perceptual Information with multiple, lower level and finer grained, variables, such as ones involving neuronal processing, we might suggest a series of variables involving activation of the retina, leading to patterns of activity in V1, leading to further activity through the higher visual areas and into the frontal cortex. This replacement series of variables would need to pick out all and only the same instances in the world, all and only the events picked out by Conscious Perceptual Information. The series of variables must do this not only for events that have already occurred, but for future events. It must provide grounds to decide if a new event is an instance of these lower level variables that were formerly described as conscious perceptual information, and it must decide this without recourse to higher level descriptions involving perceptual information: to be a genuine substitution, only intensions involving lower level neuronal processes are admissible. Is this possible?[69]

---

[69] I ask whether this is possible but leave aside the question of whether or not this is a meaningful substitute to stand in for the original variable – whether or not it is conceptually coherent to treat the person-level variable of perceptual information as even potentially replaceable without loss by such subpersonal mechanisms. I tend to agree with McDowell (1994) and Machamer (conversation) to the effect that this substitution elides an important distinction. However, since I will be arguing that this substitution is not possible for other reasons, the fact

A plausible way to effect such a replacement would be to replace the higher level extension with a lower level one that includes all and only the same instances, but includes them *qua* lower level events by altering the intension. Consider what such a replacement intension would look like for Conscious Perceptual Information. This supposedly higher level variable picks out its instances with a higher level description: events which count as instances are identified according to higher level features or properties of those events. We identify instances of Conscious Perceptual Information not by looking to neuronal processes, but usually by grosser macrophysical measures such as asking subjects if they see something, or having them perform tasks which differ in the performance if they have or lack the relevant perceptual information. The intension of Conscious Perceptual Information is thus higher level. The extension is neither higher nor lower level, *per se*, since it is essentially a list of events in the world, which it would be awkward to claim are themselves either higher or lower level. Such events can be given a higher or lower level *description*.[70] The case where the new, reduced, extension is a proper subset of the original extension (by cutting out some of the smaller scale events that comprised the larger scale events in the original extension; essentially, more finely describing the events) will be dealt with shortly.

To proceed with the attempt at replacement, we assume that all of these instances are physical instances, and as such we can replace the higher level description of each event with a lower level description of the same event. By taking the extension of the original variable, and then redescribing every member of that extension using only lower level features, we could re-

---

that there are serious doubts about the conceptual coherence of the replacement only strengthens my position.

[70] This claim will be expanded and defended in section 5.2. To be clear, it would be an ensemble of events described in lower level neuronal terms that would comprise a single event described in higher level terms involving conscious awareness.

create a new variable that had as members the same events, but which would identify these events in lower level terms rather than higher level terms. We would start with the extension of Conscious Perceptual Information – the collection of all events in the world that count as instances of having (or lacking, for the value of 'no') conscious perceptual information.[71] We would then consider each such event individually, and redescribe it in solely lower level terms: nothing about conscious information would appear, but instead a precise specification of the state of the system that involves (for instance) the exact state of the retina and optic nerve in that particular event, plus the exact pattern of firing in the visual cortex, and so on. This specification of the lower level details would need to be very exact, but since we are only describing a single event, we can for now assume (as the critic suggesting this replacement does assume) that there exists such a complete specification of the event, even though we have little to no epistemic access to it.

We have now presumably redescribed each instance of conscious perceptual information entirely in terms of the lower level processes taking place for each individual instance. The extension of the original variable has been preserved (at least for events that have already occurred) in the new series of variables. The intension of the original variable has not – it was a

[71] There are some obvious difficulties here with this extension so understood: in particular, using such a sparse description of the variables as I have provided so far, we would have trouble ascertaining at the lower level what could count as an instance of not having some perceptual information. It looks as if every event in the world that is not an instance of having perceptual information is thus an instance of lacking it – including events such as rocks falling off cliffs, planets orbiting the sun, etc. There are parallels here with the case of confirmatory evidence for laws such as 'All ravens are black', which seems to include all nonblack nonravens. To be charitable to the attempt at replacement (which will ultimately fail for other reasons), I will assume that we can meaningfully delineate the extension to cases of organisms that could have such perceptual information but don't, and which specifies the information that is lacking (i.e. lacking information about the raisin on the table within one's reach).

higher level intension, and so cannot be included in the new lower level replacement variables. We have not yet determined the status of the intensions of the original and replacement variables.

This might sound a bit convoluted, but it essentially captures the philosophical move made by the majority of philosophers who deny the possibility of substantive higher-level causation, both for causes involving conscious awareness, and for any higher level causes at all (in particular the work and Jaegwon Kim and philosophers who endorse his conclusions). This move of replacement by redescription of individual instances has parallels in the subject of laws: the reformulation of higher level laws in terms of lower level ones is an ancestor of this contemporary debate about causation. In the laws case, a major problem encountered by this method of reducing higher-level laws to lower level ones was the messy, disjunctive nature of the result. What could be stated as a somewhat general higher level law had to be described at the lower level as a wild disjunction of singular statements. While this replacement accomplished a reduction of the original law in some sense – it dispensed with the higher level predicates appearing in the original version – it struck many as also having lost some of the lawfulness of the original law (Fodor 1974). Disjunctions of apparently unrelated occurrences don't seem capable of replacing a law, if we require a way to represent the lawfulness of the generalization (its modal properties), rather than merely it's holding true.[72]

A similar problem confronts the attempt to replace a supposedly higher level variable by redescribing its instances in solely lower level terms. Let's assume the task has been successfully accomplished: each instance of a higher level variable has been redescribed in lower level terms, so every event in the world that counts as an instance of Conscious perceptual information has

---

[72] This problem was not taken as fatal for the program of law reduction, however. There are other interesting connections between the debate about the reduction of special-science laws and the reduction of higher level causes, but I will not address them here.

been completely redescribed solely in terms of neuronal processes and events. The extension has been preserved, but there is more to replace before the *variable* has been successfully replaced. We have not yet ascertained the status of the intension of the original variable under the replacement by the series of lower level variables, and the intension plays a crucial role in adjudicating future instances. The situation thus far is that the extension has been preserved by redescribing each instance in lower level rather than higher level terms. The original higher level description was, essentially, due to the intension: it was by using a higher level description that the events were picked out as instances of the variable. Changing this description means that the original intension is lost.

This is a crucial point that gets glossed over by the intuition that replacement of the higher level descriptions of events with lower level ones is sufficient to replace the variable (stated in more familiar terms, this is the intuition that there exists some state of the body at a given moment which, if completely described in terms of particle locations and momenta, is equivalent to the conscious state of the subject). The intension plays an important role in the characterization, and especially in the deployment or use, of a variable. Not all events that will be members of some variable's extension have already occurred. Besides knowing what events are already members of the set of instances of a variable, we *also* need means by which to decide if future events, when they occur, should count as instances of the variable. In other words, we still need a method for sorting events into members of the set and nonmembers.

This is the heart of the problem for any attempt at variable replacement. It can be put in terms of a dilemma. Either we continue to use the same higher level intension to pick out new members of the set, or we must use a lower level intension to do so. If we continue to use the same intension, then we haven't actually replaced the variable at all. Redescribing instances at a

lower level would be purely parasitic on continuing to use the higher level variable to determine members of the set. This would be not a replacement, it would be post-hoc partial redescription.

But if, on the other hand, we turn to using lower level intensions, we run into several formidable problems. There are two broad possible outcomes to an attempt to find lower level intensions that can be used instead of the higher level one. First, it could turn out to be impossible to do so. There may be nothing discernible at the lower level by which to pick out all and only exactly the right instances that would replace the higher level variable, except a pure disjunction of events that we already know should count as members. In this case, we would have no lower level handle on how to differentiate events into members and nonmembers, which again renders the supposed replacement variables parasitical on the original variable. I will argue in 5.2 that this is the case, and, moreover, that this isn't merely an epistemic point about our limited abilities to access these lower level descriptions at a sufficient degree of precision. Rather, it is a function of the difference between singular and general causal relationships.

Even if someone is unconvinced by these reasons, however, the situation for the second possibility is not any better. Assume we find lower level intensions for the replacement variables. This is where we must tread carefully with respect to comparisons of the replacement and original variables. If these new lower level intensions do manage to pick out exactly all and only the same instances, even for future instances, as the original variable, then we actually have no reason to think that these are somehow *different* variables. If the current and future sets of members are identical, then we have every reason to suppose that this is simply the same variable under a new guise, two ways of expressing the same thing. It would amount to a purely metaphysical assertion (and here I intend the term in the derogatory sense, as something that could not possibly make a physical difference) to say that somehow the lower level version of

the same variable was *really* the right way to express it, was somehow more causal than the higher level description of exactly the same variable, even though this made no difference whatsoever to our treatment of the variable in any possible application.

If a reductionist tries to use replacement variables with different extensions than the original variable, there is a different kind of problem. An advocate of reduction could argue that it constitutes an advantage to such a reduction that the new extension includes only a proper subset of events in the higher level extension. Some of the finer grained events out of which a single coarser grained event was composed may have been causally uninvolved, and so the ability to weed such events (or parts of events) out would make the lower level extension better for representing the causal system. I won't address whether or not this would be an advantage, because even if it is, it is not an advantage that can be claimed by the lower level representation over the higher level representation. Such an advantage would only accrue to a system of variables that were capable of picking out their own extension without recourse to a higher level or coarser grained intension. And, as we've just seen, the lower level variables lack this, instead parasitically relying on the higher level intension to pick out instances. By doing so, they are unable to claim an advantage by having a smaller, more refined extension.

But this kind of difference-without-a-difference could be avoided if, for the second possibility where we find (or assume) lower level intensions, we allow that the extensions don't have to match – if the replacement variables pick out a slightly different set of events in the world then we haven't actually merely replaced the original variable. We have asserted something new: a new set of variables that may overlap extensionally to a large degree with the original one, but which do not pick out as instances the same events as the higher level variable. In this case, we haven't really replaced a higher level variable with a lower level one. We have

changed the parsing of phenomena in the world under our variables, and we would have to buy into the assumption that a lower level, smaller size scale parsing is more ontologically legitimate than a higher level one. Many philosophers do assume just this, but the ability of this approach to capture what is genuinely causally relevant will be problematized shortly.

Thus, the best possible outcome for such a replacement would be to merely redescribe the same variable in a new way, adding nothing to the original representation except for a more complicated and less epistemically convenient way of finding instances of the variables. Even if we could always accomplish this redescription, we lack a compelling motivation for why one description of the same variable should be preferred over another – if they genuinely pick out all and only the same events (as they do by presumption here), then we have nothing to gain or lose by going with either representation. If we are unable to always find such lower level replacement variables, then we are stuck parasitically using the higher level intension to sort and then afterwards redescribing instances in lower level terms. And it would need to be possible in *all* cases to make such a replacement. Finding a lower level intensional replacement for some, but not all, higher level variables, means that as a strategy, replacement is not generally effective – it is not something we can claim is available simply based on metaphysical considerations about how lower level causes are 'really' causal whereas higher level ones are not. If an advocate of lower level replacement is serious about such metaphysical claims concerning where the real causal oomph lies (as they must be, considering the fact that these replacements really make no difference in usage, since by assumption they pick out all and only the same events in the world), then it must be the case that such replacement is always possible.

Thus, it should be clear that simply replacing higher level variables with lower level ones is a poorly formed suggestion. Simply redescribing the extension of a higher level variable in

lower level terminology does not render the variable itself lower level, nor does it mean that we can find a suitable lower level intension to use in the future. We should not accept parasitical implicit uses of higher level intensions, with subsequent lower level redescriptions of extension, as legitimate replacements of variables.

5.1.3 Replacing higher level variables, changing extension

The only plausible possibility that emerged from these considerations is the case of finding lower level intensions for replacement variables that do not match the same extension as the higher level variable being replaced. This isn't a strict replacement of a higher level variable with lower level ones, since the extension of the variables will change: some events which previously counted as instances of the variable will no longer do so, and there might be events which did not count as instances of the original variable which do count as instances of the replacement variables.

This possibility deserves further consideration. There are several assumptions that must be made to claim that this kind of replacement of higher level variables will be available and will be an improvement over the original variable. One such assumption is that the real causal story always takes place at the lower level: this would underwrite the claim that such lower level replacement is not merely different from, but is an improvement on, the original variable. Another such assumption is that not only is there a difference between the extensions, and the lower level somehow captures the more genuine causal story, but that the extensions for the lower level will also be a better representation of the system, in the sense that it will be empirically better supported than the extension for the higher level.

The first assumption will be critically evaluated and found wanting in section 5.2. And the second assumption does not follow from the first: it is possible for a lower level set of variables to exist, be more genuinely causal, and yet be a worse representation of the system, because there is more to having a good representation of a system than pure ontological fundamentality (think of Borges' map that is so precise that is takes up the entire territory of the land that it is a map of). We would have to establish that the lower level representation is better on a case by case basis, and it is not difficult to imagine scenarios in which the different extension associated with the lower level intensions was actually less well suited to some investigative purpose than the extension associated with the higher level intensions.

But I now want to problematize the assumption that this kind of replacement constitutes a sort of reduction of the original variables, in the sense of reducing a higher level variable to a series of lower level variables. We might *rid* ourselves of the original variable, but this is not the same as *reducing* the variable.[73] The problem comes down to the fact that it is really events in the world about which we are talking; we represent these in terms of variables, and some aspects of these variables (namely, their intensions) can be meaningfully higher or lower level. But the same is not true of extensions. Since they are collections of actual events in the world, they are not intrinsically leveled – events can be *described* in lower or higher level ways, but this has to do with intensions, not extensions. Real events include the higher and lower and middle and all other levels together, insofar as it is meaningful to differentiate any of these for a given event. And it is by extension that we differentiate variables from each other (i.e., two apparently

[73] The term 'reduction' is heavily laden with philosophical baggage, and I am not addressing all possible uses of that term in this section. I am confining myself to a specific understanding of reduction whereby predicates or variables that are higher level are replaced without loss by one(s) at a lower level, where levels are taken to indicate at least (although not necessarily solely confined to) a marked change in the size scale between higher and lower levels.

different variables with the same extension are simply two ways to describe the same variable). This means when two variables differ, they differ by the events they group together as instances of a variable; they do not differ according to levels. Insofar as there is any meaningful difference between a purportedly higher and lower level variable, it is solely due to inclusion of different events, not due to a difference in level as such.

We must thus take care when we speak about higher level variables. A variable is higher level in the sense that the method for sorting events into instances and noninstances involves a higher level description of those events; it is intensionally higher level. Lower level variables, using different intensions to pick out their instances, may collect together an overlapping but nonidentical set of events as their instances.

What differs in such cases between the original variable and the replacement series of variables is *not* their level. It is their extensions. And the members of that extension are events in the world. The relevant difference between the original and replacement sets of variables is that they pick out distinct events in the world.

We can certainly see how one set of variables might be more accurate than another – the new replacement variables might indeed be preferable to the original ones. But if this is so, it is not simply because the replacement version is lower leveled than the original. It is entirely due to the fact that it picks out different events. It is the events themselves, as singular occurrences, which stand in cause and effect relationships; our representations in terms of variables will do a better or worse job based on how well they collect events into common groups as collectively standing in specific causal relationships (more on this shortly). The replacement variables are not better because they are lower level. In a very important sense, the *variables* themselves are not lower level, they are not leveled at all – they have distinct extensions, that is all. Different

variables parse the phenomena in the world into different categories. There is nothing surprising about one parsing being better than another (where 'better' must be understood in terms of specific goals and questions which we are investigating in the context of explanation), but there is nothing about these distinct variables that makes one alternative better than another *because of* its being lower level, rather than because it picks out distinct sets of events.

Thus, we can offer replacement variables, and those variables can pick out their instances with intensions involving lower level descriptions instead of intensions using higher level descriptions, *and* we can allow that these new variables are an improvement on the old variables being replaced. But what we haven't done is to reduce the original variables; they have undergone something more akin to reorganization. This kind of reorganization does not demonstrate that in general, such replacement with lower-leveled-intensions is either ontologically more fundamental, or that it even exists. We have criticized the original parsing of this phenomena, something that is not intrinsically leveled, and found that this particular collection of events, rather than that one, should be used as a variable.

5.1.4 Section conclusion

Given that the difference between supposedly higher and lower level variables is in how they parse the phenomena of the world into distinct sets of events, this discussion has consequences not only for the way we should think about the causal efficacy of conscious agency, but also for the causal efficacy of any higher level cause. When we consider general or variable causation, rather than singular causation, there is no fundamental level of causation, from which other levels are derived. What differentiates causal variables as they are found in the 'special sciences' from causes in physics is not that special science causes are compounded out of or summed from

causes in physics. It is rather that, from the same nexus of events in the world, physics cuts out one set of events as belonging to a given variable, and sciences like biology or psychology cut out a different (perhaps overlapping, but certainly not identical) set of events as belonging to a different variable. This means that, when we pose the question about what the basic causal relations are in the world, from which other causal relations are derived or on which they supervene, there is no compulsion to answer that the basic or fundamental variables are lower level or microphysical ones. Higher level, or macrophysical, causes, and the way they parse phenomena into variables, have as much claim to being basic or fundamental as do microphysical causes.

The upshot of this rather dense discussion is that my analysis of variables in chapter 3 still stands as demonstrating conscious awareness to be a legitimately causal factor in action. I've shown that a substitution of putatively higher level variables for putatively lower level ones would either have to parasitically utilize parts of the original higher level variables in order to be applicable, or would actually change the set of events under consideration. In that case, one would have to argue against my chapter 3 analysis using empirical evidence to demonstrate that a specific set of variables do a better job parsing the phenomena; a metaphysical argument about the preferability of neuronal causes won't cut it. There is thus no reason to think that the variables representing conscious awareness are essentially or problematically higher level.

This wraps up the discussion of reduction as applied to variables. The next step in establishing the metaphysical viability of higher level causes involves singular causation. A critic might respond to this section by arguing that causation is a relationship between microphysical elements of an event and not between the macrophysical ones, regardless of what happens with variables and general causation. This is at the heart of the Causal Exclusion problem. Different

ways of pitching this criticism will boil down to some version of the notion that the real causal story for particular instances of causation is always at a lower level, where the most legitimate causes are those at a microphysical level and the apparent causal efficacy of a higher level variable is spurious. I now turn to this reformulated criticism of my view.

5.2 Counterfactuals and singular versus general causation

I have now demonstrated how micro explanations could not in general replace macro explanations without loss, for token instances of causation (in chapter 4), and for causation between variables (previous section). Explanatorily, micro and macro explanations are not interchangeable. Yet a critic might still be concerned about the 'causal oomph' of the higher level causes I have presented. It might well be that we can use higher level causes in explanations, but that the causal efficacy of these causes is still dependent on their microphysical supervenience base. This problem was first raised by C.D. Broad, and has been aptly named the problem of quausation by Terence Horgan: do the higher level cases bring about their effects *qua* higher level cause, or merely *qua* their microphysical supervenience base?

One might also be concerned about my use of counterfactuals in chapter 4, and the modal connection between individual instances of a variable. How do I justify the claim that for a single instance of macro causation, there is a counterfactual buffer zone around it such that minor micro alterations would not alter the causal relationship? There is a quite legitimate concern about the means by which we could evaluate these counterfactuals for individual instances.

In this section I will outline the kind of interdependency that singular causation and general causation have with one another. By singular causation, I refer to particular instances of

something causing something: the dropping of a plate on a particular day at a particular time, which caused it to break. By general causation, I simply mean type-level or variable causation, causal relationship between groups of instances: dropping plates causes them to break. Actual causation in the world is, as has been discussed before, singular; in this regard, general causation is always dependent on singular causes that are collected together. Yet there is also a reverse dependency: the counterfactual buffer zone around an instance of singular causation[74] is defined in terms of, and known by means of, general causal relationships. Even though we have epistemic access to this buffer zone for a single instance by means of general causal relationships, it is not merely an epistemic buffer zone – it is a modal property of the instance of causation itself. This interdependency of singular and general causation has the consequence that genuine causation need not be solely microphysical, and illustrates how general causation helps delineate the counterfactual buffer zone for singular causation.

The upshot of this will be that single instances of supposedly 'higher level causation', where the entities that act as causes are at a macro level, are not only genuinely causally efficacious, but are also efficacious *qua* higher level cause. This has implications for a wide range of causes in the so-called special sciences, where causes are composite in some sense. At the end of this section, I will discuss the implications this has for the debate about the causal role of conscious awareness in action: namely, that conscious awareness is causally efficacious in action, as we saw in the last chapter, and that it is causally efficacious *qua* conscious awareness, not merely *qua* the microphysical entities and processes on which it supervenes.

---

[74] It is worth reiterating that I am using an interventionist view of causation, which is one kind of counterfactual theory, in making these claims, not a view such as the conserved quantity transfer of Salmon and Dowe. Such transmission theories would require a different analysis; see chapter 3 for reasons as to why this view of causation is unsuited to the topic.

5.2.1 The interdependence of general and singular causation

I'll start by considering the most obvious direction of dependence. General causal relationships are dependent on singular ones, in that a type-level cause, a variable, simply is the set of instances of causation. The general variable picks out single instances of causation in terms of some common feature that they share, by virtue of which they are members of the set (i.e. the intension, per the discussion in section 5.1). Variables can be distinguished by their extension, by the sets of instances that count as instances of that variable. This is an ontological dependency: variables are the collections of their instances, and the instances are causal-ontologically more fundamental.

But there is also a reverse dependence. Singular causation depends epistemically on general causation. But this epistemic dependence also reveals an ontological feature of singular causes. Variables don't merely randomly collect instances together and throw a label on top of them all.[75] The instances of a variable are collected by dint of some feature that they all share, some aspect of the singular causal relationship that can be indicated as common to them. It is here that we can see the difference between a counterfactual interventionist approach to causation and a process or conserved quantity transfer account (CQT). For a single instance of causation, as we saw in an earlier section, there may be a large number of individual causal interactions that take place, as part of the process from cause to effect. According to a CQT account, these interactions are all equivalently causal; taking a slightly different path from cause

[75] It is, of course, always possible to generate such random collections of instances and call them variables. But these are not variables that are useful for anything, or are suggested by research. I am here confining myself to the kinds of variables that we could actually encounter in science.

to effect, where a slightly different combination of quantities were transferred between processes, would mean that the causal relationship itself was somehow different. However, on an interventionist approach, this is not the case. Not all of the causal interactions that take place between cause and effect need be genuinely involved in the causal relationship. Some of the specific details of causal interactions could have been different, without altering the fact that the cause led to the effect.

These cases are counterfactually robust. In order to define the boundaries of the counterfactual buffer zone around the singular causal relationship, we can use the details that could have been different without altering, in a counterfactual sense, the instance of causation. But we only know *which* details could have been different by using general causal relationships. By using information about which variables this singular instance counts as an instance of, we hone in on which interactions were causally relevant to the effect, and which were not.

An example of this might be a series of neuronal processes whereby perceptual information about a scary looking face is sent immediately to the amygdala, provoking a fear response before the subject becomes fully aware of the perceptual information about that face. For a single such instance, there is some actual matter of fact (about which we realistically have little to no epistemic information on any given occasion) about the microphysical processes whereby the light from the face struck the retina, information about it was propagated in at least two directions (to V1 and to the amygdala), and a series of processes involving neuronal activity in the amygdala and elsewhere transpired. The way in which this instance actually took place on a given occasion is indefinitely precise. There was not only some fact of the matter about the neuronal activity, but also about biochemical activity, about the positions and momentum of the particles that compose the biochemicals, all the way down to the lowest microphysical level,

where the facts about what happened during that instance of causation could be described quantum mechanically).

This indefinitely precise series of occurrences that constitute the singular instance of causation, a scary face causing a pre-conscious fear reaction, is misleadingly precise, however, relative to this sort of causal claim. There is a counterfactual buffer zone around the details of this singular instance. We have reason to think that some of the neurons that were actually involved could have failed to fire, and yet the same reaction would have taken place; similarly, some of the uninvolved neurons could have fired, without changing the outcome. Some of the biochemical details, at each step, could have been different in small but physically significant ways, without altering the relationship between cause and effect in this instance. Many of the particles could have had different values for position and momentum, indeed a range of different values than the ones they actually had, without altering the causal relationship – the face would still have induced a pre-conscious fear reaction. Thus, specifying the microphysical details to a significant degree may be misleading or inaccurate, in that some of those details did not contribute to the outcome; they could have been different.

For the example of the amygdala, we do not know all of the features of the processes that could be altered, or the extent to which they could be altered, without changing the causal relationship. But we do have grounds for making claims about some of them: we can say with confidence that had some single particular neuron within a population fired, the chances that the effect would have been different is (for most such individual neurons) vanishingly small. We know this because of characteristics of variable-level causal relationships that are investigated in neuroscience. This buffer zone is thus an ontological feature of a single instance of causation, epistemically revealed by variable causation. While our knowledge of the extent of the buffer

zone is generally limited by circumstances, especially for cases such as the pre-conscious fear reaction, that buffer zone nevertheless exists (and is crucial for actual scientific practice).

For any individual instance of counterfactually robust causation, we need some means by which to evaluate counterfactuals, some means by which to find out at least something, albeit not everything, about the extent of this buffer zone. A single instance of causation does not carry that information with it, as it were: it does not provide its own means of counterfactual evaluation. The problem is that many existing accounts of counterfactual evaluation are unsatisfactory for a number of reasons. The primary problems with possible worlds are discussed elsewhere in this chapter: evaluation of counterfactuals requires that judgments of similarity in possible worlds be the same – that we have the same intuitions about what counts as sufficiently similar. Possible worlds cannot answer counterfactual questions to which we don't already know the answer, and so they provide us with no new information. What we would like is a means of evaluating counterfactuals that has applicability to scientific investigation, where the causal structure of systems is not know at the outset. Possible worlds, however, don't provide any additional means to adjudicate; we get out of them all and only what we started with.

5.2.2 Evaluating singular counterfactuals with general causation

My proposal is that we use the epistemic dependence of singular causation on general causation as a means by which to evaluate these counterfactuals, to gain information about the extent of the counterfactual buffer zone around single instances of causation.[76] This approach to singular

[76] As James Bogen has pointed out, my proposal here should be understood in terms of general causal relationships being that which make counterfactuals true or false; they are the means we have to ascertain whether or not counterfactual statements are true because they reflect certain

causation aptly describes common scientific practice. The sciences mark distinctions between singular instances that can, by use of general causation, be counterfactually evaluated, from singular instances that lack the kind general causation needed for evaluation, and thus cannot be counterfactually evaluated. General causation can be used to evaluate counterfactuals regarding singular causation when the single instance being evaluated counts as an instance of some sort of general causation. A single instance may count as an instance of multiple variables, depending on research interests and different ways of dividing up phenomena into variables. I will outline this approach here, although it deserves a much longer discussion in order to be fully developed.

Consider a single instance for which we would like to be able to evaluate some counterfactual: if a given particular aspect of the process had occurred differently, would the effect still have transpired in the way that it did? In other words, was the occurrence counterfactually robust with respect to that particular change? If we confine ourselves to only consider this single instance, the only means we have to evaluate the counterfactual are in terms of possible worlds, or, perhaps, in terms of covering laws. We would like to be able to evaluate the counterfactual using only events (this event and others) that have actually occurred, not ones in possible worlds. And it's good to avoid problematic accounts of physical laws if we can. Causal relationships between variables allow us to meet these desiderata. Let's assume that for this particular instance of causation for which we are trying to evaluate the counterfactual, the cause and the effect instances are members of variables about which we have information. The kind of information we need here is fairly straightforward: we have witnessed previous instances of this cause variable, as well as previous instances of the effect variable; we have some (it need

---

features of the world. I am not providing a semantic analysis of the truth of counterfactual causal claims.

not be exhaustive) information about conditions under which the cause and effect can occur; and, what would be most convenient, we might already have sufficient evidence to ground the claim that the one causes the other (for what counts as sufficient evidence for this, see chapter 3).

Knowing that variables of this kind cause effect-variables of the other kind is not necessarily enough to evaluate particular counterfactuals. But we can take a look at the other individual instances of causation that are relevantly similar to the one we are concerned with (where relevantly similar means that they are also instances of the same variable; relevance can be considered in terms of the intension of that variable). Comparison to these other instances and other variables provides means to evaluate the counterfactual question of whether the effect would still have occurred as it did had this particular part of the process differed in a specified way.

We can keep looking through other instances of cause and effect variables to which our instance belongs, to see if we find any that instantiate the counterfactual we want to evaluate, where the particular part of the process differs from the precise way in which it originally occurred. Metaphorically, we can think of this process as taking the different instances of the cause and effect that occurred, and laying them over one another. Only those features which all of the instances have in common are candidates to be genuinely necessary for the causal relationship between cause and effect. Those features of individual instances that are not shared by other instances are candidates for being in the counterfactual robustness zone: they could have been different without altering the causal relationship. By dint of its membership in cause and effect variables, the counterfactual for the original instance can be evaluated using the other instances of those variables. Does the particular part of the process described in the counterfactual change between different instances of the variables? If so, we have reason to think

it was not necessary for it to have transpired in precisely the same way it did. Does that part occur the same way each time? If so we have good reason to think that it must occur that way in order for the cause to bring about the effect.

This technique will not provide absolute answers to counterfactuals except in the case of deterministic causation. For probabilistic or indeterministic causation, it will provide something less than Yes or No but something more than No Idea. If the cause is probabilistic, it will not exercise its capacity, or causally contribute to bringing about its effect, every time it occurs. As such, we can't answer simply Yes or No to the question of whether the effect would have occurred, had the cause the occurred. But failing to meet this high of a standard shouldn't trouble us too much, because often enough the counterfactuals will be helpfully answered by being able to say that the effect would most likely have occurred, with such-and-such a percentage chance, had the cause occurred. Probabilistic evaluations of counterfactuals also give us traction on understanding a causal event.

To return to the example of the fear response via the amygdala, we could ask if the causal relationship between seeing the scary face and the effect of a pre-conscious fear response on a given occasion is counterfactually robust with respect to whether or not the scary face is consciously perceived at any point.[77] Let's take X1 as the subject seeing a scary face and having a preconscious fear response to it; and stipulate that the occurrence X1 includes the subject being subsequently consciously aware of the face. The counterfactual we want to evaluate is something

---

[77] This example uses a much more macroscopic change in the process to counterfactually evaluate than need be the case. The arguments offered here also apply to much smaller changes, such as individual neurons firing or not firing, or even single particles having different locations. However, those kinds of examples aren't as illuminating here because of the limited epistemic access we have to them. Using such an obvious change also demonstrates that counterfactual robustness zones can be quite large, encompassing many different microstates.

like this: had the subject not become consciously aware of the scary face, the face would still have provoked a preconscious fear response.

In order to evaluate this counterfactual in the scheme I've just described, we should look for the variables of which X1 counts as an instance. This is pretty easy: seeing a scary face is one, having a preconscious fear response is another. The experiments in which these variables are utilized provide us with a bank of other instances that we can check in order to evaluate the counterfactual. In particular, experiments in which the scary face is a masked prime, displayed for too short a period of time for the subject to become conscious of the stimulus, will be helpful in evaluating this counterfactual. Compare such instances to our original instance: the causes will be relevantly similar, in that they are all instances of displays of scary faces. Are there instances where the effect still occurs – the preconscious fear response – but the subject is shown the stimulus for too short a time to become consciously aware of it? Yes, there are. And it is precisely the existence of these other instances, these actual past occurrences, which provide us with grounds to say that X1 is counterfactually robust with respect to whether or not the face was consciously seen by the subject. If it had not been consciously perceived in the actual occurrence X1, then the subject would nevertheless have had a preconscious fear response.[78]

This claim about counterfactual robustness holds not for the variables, but for the individual instances that constitute those variables. We can say of X1 itself, the single case, that

[78] Again, the answer provided by the experiments might be probabilistic. But for this particular experiment, unless there was some other reason to think that this occurrence in this subject was an exception, there are good reasons to believe, and no good reasons not to believe, that had the face never reached conscious awareness, the subject would still have had the pre-conscious fear response. As such, this is a fallible epistemic guide – out of every 100 times we provide this answer, we may end up being wrong in a couple of them. But we should be wary of holding ourselves to problematically, and unscientifically, high epistemic standards for counterfactual evaluation.

it was counterfactually robust with respect to whether or not the face was consciously perceived. Had the individual event X1 occurred differently in this way, the same effect would still have occurred. This is a significant feat: no recourse to possible worlds, or to covering laws, was required in order to evaluate the counterfactual for a single instance. It is features of X1, by dint of which it is a member of certain variables, that allow this counterfactual robustness to be epistemically accessible.

Being able to isolate the features of single instances of causation that were genuinely necessary (in the sense that they are not counterfactually robust to alteration) is extremely valuable. It allows us to isolate what, out of the many causal interactions (in a Salmon/Dowe CQT sense) surrounding the event leading to the effect were actually efficacious in bringing about the effect itself, and which merely happened to occur in the vicinity, without being causally involved in bringing about that particular effect.

5.2.3 Evaluable and nonevaluable counterfactuals

A number of interesting things follow from this. One is that we will be unable to evaluate counterfactuals for occurrences of causation that do not count as a member of any causal variables about which we have information. Most such occurrences would still have some kind of counterfactual buffer zone – it would be extremely rare to find a causal relationship so fragile that any change would destroy it. But we would not be able to know anything about the buffer zone, and thus be unable to evaluate counterfactuals about what would have happened had something been different than the way it actually occurred. Science provides us with the kinds of information about variables that we need to evaluate many counterfactuals; insofar as some occurrence of causation is not a member of some variable, then science cannot provide that

information. Such ‘one-off’s’ must be treated differently than single occurrences that can be counterfactually evaluated.

An example will make this clearer. Consider the counterfactual evaluability of a speciation event, where one species ends up branching into two separate species. The speciation event is unique, in that this particular branching has never happened before and will never happen again. The precise conditions under which it took place will never be repeated. There are, no doubt, counterfactuals that we may suspect to be true or false, but which we are not able to properly evaluate. This might include details such as small changes in subpopulation migrations, or counterfactuals involving what might have happened if a single event, such as a fire or other catastrophe, had not occurred. We will often simply lack the information to evaluate these. However, even though it is a unique event, there will be other counterfactuals that could be answered about it, based on the speciation event’s membership in different variables. We already know that genetically isolated subpopulations under different selection pressures are likely to eventually end up as distinct species. If the speciation event is a member of a variable such as this – it includes genetically isolated subpopulations in different geographical areas, with different ecosystems in each area – then there will be at least some counterfactuals we will be able to evaluate. If the ecosystems to which the subpopulations migrated had been different than they were by being less similar to each other, then we have grounds, from other such occurrences, to say that the speciation would still have occurred.

If, on the other hand, the counterfactual involves genetic isolation but nearly identical ecosystems, we would have less to say about whether or not the event would have occurred in the same way. Insofar as we can consider individual instances of speciation events as members of variables, we can make generalizations about them, use them to evaluate counterfactuals about

other such variable members, and other activities involved in making scientific generalizations. Insofar as we are unable to evaluate such counterfactuals, we must treat these instances as completely unique, and we can only study them by finding out more information about these one-off's as unique historical events; information about other occurrences will not shed light on what was causally relevant and what simply happened to occur.

Compare this to the evaluation of counterfactuals regarding Condaleeza Rice first becoming the Secretary of State of the Unites States.[79] This is a one-off causal event *par excellence*: it could only ever possibly occur once, and the constellation of events that preceded it is vast and convoluted. Out of all the events that preceded her becoming Secretary of State, some were causally efficacious in bringing about her eventual Secretaryhood, and some preceded it but were not causally efficacious in bringing it about. We could try to gauge which events preceding her ascendance to Secretary of State were causally involved by considering the event as a member of some variable. We will most likely have very little luck with that, however. Consider one candidate variable, the set of events of people becoming Secretary of State of the US. This already includes fewer members[80] than is convenient for evaluating complex counterfactuals about the kinds of events that are or are not causally involved in rising to the office of Secretary of State. Further, we have reason to think that a significant portion of these will not be helpful in evaluating counterfactuals about Condaleeza Rice. The sorts of thing it takes to end up in such an office have presumably changed since the late 1700's and 1800's – for instance, the thought of a female or black Secretary of State during most of US history would have been overwhelmingly

---

[79] This example is due to James Bogen (private conversation), although he and I disagree about what it demonstrates in this regard.

[80] There have been 67 full Secretaries of State, according to Wikipedia, which does not include temporary acting Secretaries while replacements were sought. At the time this section was originally written, Rice was the Secretary of State.

improbable, to the extent that the circumstances and events which were actually causally involved in Rice's becoming Secretary of State could not have occurred then, or if they had occurred, would not have been causally efficacious in bringing about Secretaryhood. We could instead compare Rice to other contemporary secretaries of state in other countries, but this would run into a similar problem: too small a set of other instances, and too widely varied circumstances in which each become secretary of state, including cultural, economic, political, and other such factors.

The upshot, then, is that we lack the ability to evaluate counterfactuals about what could have been different in Rice's past such that she could still have ended up becoming Secretary of State at the time that she did. We can easily and vacuously say there must have been something that could have been different – she probably could have had the ham sandwich instead of the turkey on some occasion of a meeting with the President, to illustrate the vacuity involved in asserting the existence of an unspecified counterfactual robustness zone. But we cannot say of specific events or circumstances that they could have differed or in what ways they could have differed without destroying the effect in question.

5.2.4 Solving the problem of 'quausation'

There is a consequence more directly germane to the metaphysical question of whether or not any mental process could be a cause *qua* mental process. Being able to evaluate counterfactual robustness buffer zones for single instances of causation means that we are better able to isolate just what it was about an event that was causally efficacious, and what about the event happened to occur but was not genuinely causal in bringing about the effect. This consequence is key to understanding the causal contribution of higher level causes, and to answering the question of

whether they are causally efficacious merely because of the causal abilities of their microconstituents, or whether they are causally efficacious *qua* higher level cause. The epistemic dependence of singular causation on general causation justifies the claim that single instances of conscious awareness can be causally efficacious, as instances of awareness and not merely as brute physical occurrences. I'll spell this out by first looking at the problem of quausation as laid out by Terence Horgan.

The problem is this: "Even if individual mental events and states are causally efficacious, are they efficacious *qua* mental? I.e.., do the mental types (properties) tokened by the mental events and states have the kind of relevance to individual causal transactions which allows these properties to figure in genuine causal explanation?" (Horgan 1989, 47). The concern is that the causal oomph, the actual causing, appears to be entirely done by the physical interactions of the particles that instantiate the token and the laws that govern such interactions. If this were the case, then the token of the mental event would be causally efficacious, but only as a purely physical event – only because the particles that instantiated it happened to bump into one another in just the right way to bring about the effect. It would not be *qua* mental event that the token brings about the effect, but only *qua* microphysical event.

This is a metaphysical problem about whether or not we can call mental events (using his terminology) causes in any significant way. If we cannot establish that mental events (or features of conscious awareness) are capable of being causes *qua* mental events, then the empirical demonstration that they are causally efficacious is spurious. Horgan's solution is to formulate a theory of quausation, of what it means for something to cause *qua* what it instantiates: c *qua* F causes e *qua* G, where c is the token instantiation of F, and e for G. His solution has some merits, and is, I think, very much on the right track. But his solution also has several important

shortcomings. He does not start with a notion of causation to modify in order to account for quausation, and thus ends up swamped in the effort to eliminate associations that are not causal through rather tortuous means. This problem has been addressed by a number of causal theories, and Horgan need not reinvent the wheel; he could have added onto an already developed view of causation. He does not provide a way to ascertain which features are causally relevant, instead assuming that we already know. Most problematically, he utilizes possible worlds as a means of evaluating counterfactuals, so that his claims that mental events are 'quausally' relevant end up depending on similarity judgments between worlds (p. 62), and what counts as 'too weird' a situation for two worlds to be similar. I'll focus on this last criticism.

Horgan claims that we should check nearby possible worlds that are relevantly similar to this one in order to find out if a particular property was quausal. If we find that these nearby worlds have the same events or properties as this one with respect to a particular causal event, then we can say that those events or properties are quausally relevant to the outcome – they are causally relevant *qua* the kind of thing of which they are instantiations. For instance, if we are asking if a decision to raise one's arm was quausally relevant to one's arm going up, we look around to similar possible worlds in which one's arm goes up, and see if that event was preceded by decisions to raise one's arm. If so, then the physical event which instantiated the decision was causally efficacious *qua* decision to raise one's arm in the arm going up.

However, we can always think of possible worlds in which the laws are just different enough that very similar physical states obtain, yet decisions to raise one's arm do not. Horgan is forced to claim that these worlds simply aren't similar enough to our own in order to count for evaluation of quausal relevance, and to claim that this lack of similarity is or should be intuitively apparent. This begs the question. Not all philosophers (including, one suspects, those

who thought of these counterexamples in the first place) share those intuitions about what should count as relevantly similar. And regardless of what one's intuitions are regarding the closest possible worlds, this is a dubious way to find out about the causal structure of our world, particularly if one does not already know what that structure is. The requirements of shared possible world intuitions is shaky ground on which to build a theory of causal relevance. If we are trying to find out what factors are causally relevant because we don't already know, then we don't know what the closest worlds are to which we should look, or what the outcomes would be in those worlds.

Because of this shortcoming, and other others mentioned above, the view of counterfactual buffer zones around single instances of causation offered in this chapter is a preferable way to solve the problem of quausation. My view does not have to worry about noncausal associations being quausally relevant, since I start with already established causes, in particular with causal relationships between variables, and with a methodology for determining whether or not causation holds between two variables. By looking at variables of which the token of causation is an instance, I provide a means by which to ascertain which properties or features are causally relevant ones, rather than by relying on intuitions. And by utilizing general or type causation as the means by which counterfactuals are evaluated, I avoid possible worlds altogether.

There will be counterfactuals that could be evaluated when using a possible world approach that cannot be evaluated on my approach, since there may not be variables that we can use as a means of evaluation. This is, I argue, a *strength* of my approach over the possible worlds approach, since the possible worlds approach offers a false kind of specificity. Using possible worlds and similarity relations between them, we could presumably answer questions about what

would or would not have affected Rice's becoming Secretary of State. We could see if going to that one party where she met that one influential person was in fact causally relevant, by looking to similar worlds and seeing if she went to the party and met the person there. But this actually gives us no information: we gain nothing from looking at possible worlds that we didn't already put into those possible worlds in the first place. We don't actually know what would have happened in those possible worlds, and any answer we provide is essentially a combination of intuition and guessing, lacking empirical justification. We should want to mark a difference between counterfactuals that can be answered using genuine empirical information, and those that can only be answered with intuitions about what is "too weird" to be similar. The possible worlds method of counterfactual evaluation does not mark such a difference, but the counterfactual buffer zone approach does.

We can answer quausal questions about features of conscious awareness using counterfactual buffer zones in the following sort of way. Let's suppose we are considering a single event of causation where a subject's decision to raise her arm led to her arm going up. This is both a physical and a mental event, in Horgan's terminology; we want to know, in my terminology, whether or not the physical processes involved in the decision to raise the arm were causally efficacious *qua* decision to raise the arm, or merely by dint of their brutely physical, or CQT, features. We are essentially asking this: of the things that could have been different about the decision and the arm raising without affecting the causal relation between them, is the fact that the subject decided to raise her arm one of those irrelevant features? Could we have wiggled that feature of the situation and still had the arm go up? Comparatively, could we have wiggled features of the brute physical situation – such as Salmon/Dowe CQT causal interactions between some of the particles – and still had the same effect? The answer should be rather obvious, that

wiggling the decision to raise the arm also wiggles the arm going up. The class of instances in which people decide to raise their arm lead almost always to arms going up, whereas the class of instances of people not deciding to raise their arm (or, perhaps, deciding not to raise their arm), leads only very rarely and under unusual circumstances to arms going up. On the other hand, we can be sure that there are any number of microphysical features of the situation that could have differed, either slightly or in some cases greatly, from how they actually occurred without altering the effect.

From this, we see that the counterfactual buffer zone around this instance supports the claim that the decision to raise the arm was causally relevant in the arm going up, *qua* decision to raise the arm. It was not causally relevant simply due to its physical interactions, since any number of those physical interactions could have been different without altering the causal relationship. It was causally efficacious in the way that it was because of exactly those properties that also make it a member of the right variable for evaluating the counterfactual, namely those properties that make it a member of a variable involving conscious awareness. It was not merely by virtue of the physical properties that the decision to raise the arm led to the arm going up – there are any number of nonidentical physical situations which would also have led to the arm going up, as well as there being similar physical situations in which the arm did not go up. The difference between these situations can be marked by saying that some count as decisions to raise the arm, and some don't. Merely being physical is insufficient.

The conclusion of this section, then, is that even a single instance of causation involving conscious awareness involves that awareness in an integral fashion – it is not merely that a physical instance also happens to be one of conscious awareness, but that its being an instance of awareness contributes to the causal relationship. Because the counterfactual buffer zone is an

ontological feature of the token of causation, and that zone can be delineated with respect to conscious involvement, conscious awareness can be a cause *qua* conscious awareness.

5.3 Causally articulated downward causation

This section takes up a final metaphysical problem associated with the possibility of conscious awareness being genuinely causally efficacious. The problem is that if it were to be so, awareness must engage in some kind of downward causation, which is at best a contentious possibility, and at worst outright incoherent. The debate turns on whether or not 'higher level' features of the world like conscious awareness could causally affect 'lower level' features of the world, including lower level, brutely physical, features, especially given that it is these lower level physical events which form the supervenience base for higher level events. This is a different question than whether or not higher level variables could genuinely causally affect other higher level variables. There are few, if any, ways to assert that awareness has genuine causal efficacy in action, without thereby asserting the existence of some kind of downward causation. If downward causation were a legitimate metaphysical possibility, then it would become an empirical issue to demonstrate whether or not it was actually exercised in case such as conscious involvement in action. Establishing that there is the genuine possibility of downward causation would thus advance the discussion of the causal structure of agency: we could turn to see if we can produce actual instances of such downward causation in systems involving conscious agency. In this section, I will demonstrate how systems with certain characteristics will be potentially capable of engaging in downward causation. These

characteristics, which center around having a certain degree of complexity, are had by a variety of systems, including that of conscious human agents.

One of the obstacles to conclusively demonstrating that conscious aspects of agency are causally efficacious in action is that there is little consensus about the nature of the interlevel relationship in this case. Many authors are willing to label conscious aspects of agency, such as intentions, decisions, and voluntary attention, as higher level processes, on many different understandings of how levels are individuated. Neuronal processes, at the level of neurons and of grosser brain anatomy, are similarly generally agreed on as comprising a lower level. There cannot be a change in these higher level processes without there also having been a change in the lower level processes: this is a case of weak supervenience. But what exactly is the relationship between the higher level conscious processes and the lower level physical or nonconscious ones? What is the nature of the levels in question? This problem, the relationship between consciousness or conscious awareness and the physical processes on which it supervenes, is at the heart of the philosophical problem of consciousness (Chalmers 1996). As yet, science has no answer to the problem; the prospects in the near future are not good for answering precisely what the nature of the interlevel relationship is in this case. Work has been done (see especially Koch 2004) on the neural correlates of various features of conscious awareness. But speaking in terms of correlates is deliberately ambiguous about the relationship between the levels being considered.

This is a somewhat oversimplified but still useful characterization of the current situation regarding the causal efficacy of conscious awareness: it is unclear whether there is any substantive notion of downward causation with which one could argue for the higher level causation that conscious causal efficacy would have to have if it had any at all. It is *a fortiori*

unclear whether there is such a notion that is suitable for conscious causal efficacy, given that existing accounts of downward causation are usually developed with respect to specific interlevel relationships, most often emergence. We do not know which of multiple possible interlevel relationships must be accounted for.

In this section, I provide a theory of downward causation in causally articulated complex systems that answers both of these needs. I develop the theory in deliberately abstract terms. In any system that can be represented with at least two levels, one of which weakly supervenes on the other, and where there are at least two distinct entities or processes at each level, there is the possibility of upward and downward causation in addition to the noncausal interlevel relation. Having a theory of downward causation that is applicable to different interlevel relationships, and a theory that is compatible with a broadly interventionist framework, is generally useful. My motivation for the project, however, concerns its ability to shed light on the problem of the causal efficacy of conscious agency. I demonstrate how, in specific kinds of systems, there is a metaphysically viable sense of downward causation. Whether any specific system that meets these criteria exhibits downward causation is an empirical question that must be answered on a case-by-case basis. But the possibility of such causation is itself an important step.

Section 5.3.1 argues that we can make substantive claims about interlevel causation even in systems for which we lack a clear characterization of the interlevel relationships. In 5.3.2, I develop a general characterization of complex multi-level systems applicable for many means of level differentiation, and of the conditions that relata must meet in order to stand in a causal relationship. This is used to illustrate the notion of causal articulation in section 5.3.3, where I present the theory of downward causation in complex multi-level systems. Section 5.3.4 addresses a criticism of calling this downward causation, namely, that a rearrangement of system

or level boundaries eliminates the spurious appearance of downward causation. I argue that recourse to boundary rearrangement is not always available, and that relying on such gerrymandering of level and system boundaries is an ad hoc solution to a nonproblem, since the considerations that made downward causation metaphysically dubious have been eliminated. Section 5.3.5 concludes: the possibility of a genuinely causal role for conscious awareness in action can be secured by demonstrating how, as agents, we meet the criteria for causal articulation in a multi-level complex system, even though we are notoriously unable to firmly characterize the relationship between the levels in question.

5.3.1 Dealing with unspecified level differentiations

The philosophical discussion of downward causation (DC), the view that higher level causes can have direct influence on lower level effects, has often been at cross-purposes. The debate has been as much about conflicting intuitions as anything else (see especially Sperry 1975, 1986, 1969, and Klee 1984). In fact, some of these papers offer the *same* examples both in support of and against the possibility of downward causation (e.g. Sperry's example of an object rolling down a hill).

A common way to defend the notion of downward causation has been the invocation of emergence, whereby higher level entities or processes have substantive causal abilities distinct from and not reducible to those of the lower level entities or processes out of which the higher level emerges (Bedau 2002, Robinson 2005, Meyering 2000, Emmeche et al 2000). But this strategy of defending downward causation by reference to emergence has not clarified the issue: emergence turns out to be almost as difficult to agree on as DC. One would like to be able to

explicate downward causation, if there is such a thing, without relying on the notion of emergence.[81]

An uneasy stand-off in the debate concerning downward causation has been reached on some key issues. This takes the form of a requirement, in discussions of levels and multi-level systems, that one must specify precisely the method of individuating levels being used (especially Craver 2007, chapter 5). The understanding seems to be that if we sufficiently specify what the actual relationship is between the levels, we eliminate the need to address downward causation, because the question of causality itself simply won't appear: causal relationships will hold between entities or processes at a given level, but never between entities or processes at different levels (Craver and Bechtel (2007). Instead, the relationships between entities and processes at different levels will always be of a noncausal variety: for instance, emergence, constitution, realization, et cetera.[82] This turns the debate away from downward causation towards higher level, intralevel, causation. It turns the question into one about whether the apparent causal relationships between higher level entities or processes are genuinely causal, or merely epiphenomenal and wholly derivative from the causal abilities of the lower level entities or processes. The debate about the epiphenomenality of higher level causes doesn't address either downward or upwards causation.

---

[81] The connection to emergence has in some ways been counterproductive for discussion of downward causation. For instance, Jaegwon Kim (1999) concludes that there could be no such thing as downward causation based on the fact that he finds emergence to be metaphysically problematic; he does not consider other interlevel relationship that could support downward causation.

[82] There are also levels such as analysis, but I will not be addressing these, as I take downward causation to be a relationship within physical systems themselves, and not simply a consequence of how we analyze them. While there is a close relationship between levels of analysis or explanation and levels in physical systems, the two are not identical.

But in some regards, the pendulum has swung too far in requiring fully specified levels. True, failure to specify the kinds of levels under consideration makes it all too easy to make claims about interlevel causation that closer inspection reveals as vague and confused. However, this critical injunction to clarify claims by specifying the interlevel relationship has given rise to a general sense that, when speaking of levels, we *only* make sensible claims when we confine them to a definite, single characterization of levels, usually deployed in a local context with reference to a single example or handful of examples. This, I argue, unduly limits the scope and strength of claims we can legitimately make in the context of interlevel relations. While some claims about multi-level systems must specify the kind of levels in question, there are claims that can be justified in terms of any type of level. Insofar as we can represent some system abstractly as a complex system with multiple levels, there will be claims that are true of all such systems by dint of these abstract structural features.

Besides the desideratum of a theory of downward causation that applies more widely than just to emergent interlevel relationships, there is the further desideratum of a theory of downward causation that coheres with existing methods of causal analysis. One of the most interesting and fruitful methodologies of causation currently available is that of interventionism. This broad category includes the work by authors such as Glymour, Spirtes, and Scheines (2001), Pearl (2000), Woodward (2003) and others. This is a difference-making view: causes make a difference to their effects, and thus we can ascertain if something is a cause by 'wiggling' it (changing its value), and seeing if the value of the effect also changes. The advantage of a theory of downward causation that fits in with interventionism is that, having established the basic possibility of downward causation, we can then utilize this well-developed criteria for establishing specific causal claims in interventionism. If downward causation meets the criteria

for genuine causation, according to interventionism, then there is substantive reason to give credence to such claims of DC. Given the basic metaphysical viability of downward causation, interventionism can provide the framework in which to find concrete empirical examples of DC.

In this section, I develop a theory of DC that meets these desiderata: in complex, multi-level systems, there exists a causal articulation of subsystems or components that allows for a substantive notion of downward causation. This holds regardless of how the levels in the system are differentiated, and does not require invocation of emergence in order to justify the claim of causal influence from a higher to a lower level. I start with a quite abstract description, so as to include as many kinds of systems and level differentiations in the scope of the theory as possible. The theory of downward causation itself is thus general, and could be applied to any kind of higher level causation, including psychological, biological, ecological, and other sorts. But my motivation is how this theory clarifies the long-standing debate in conscious agency, where we agree that there is an interlevel relationship but do not know what it is.

5.3.2 Complex systems and causal relata

There is a vast array of ways to break a single system into higher and lower levels. Systems with logically distinct level breakdowns will differ from one another in important respects, and some claims about levels have to be couched in terms specific to the kind of levels in question. That said, there are also claims that will be true of any system with levels, regardless of how those levels are differentiated, by dint of the very fact that there are levels at all. It is this latter kind of claim I am interested in making. In order to understand such a claim, I'll provide a weak, broad characterization of complex systems and of levels in a system. This characterization is generic in that it doesn't depend on any specific way of individuating systems, nor need there be a set of

necessary and sufficient conditions by which to draw a strict boundary between a system and its environment. For most of the cases we are interested in, the existence of a system with sufficiently firm boundaries is already more of less given, enough that we can distinguish between the system, its components or subsystems, and its environment.[83]

The systems to which my claims apply must be multi-level. A multi-level system is one that has at least two distinguishable levels, one of which can be understood as higher and the other as lower.[84] The broadest way to describe the relationship between these levels is in terms of a weak form of supervenience: there cannot be a change in the higher level without there also being a change in the lower level.[85] Using this definition allows us to collectively consider all the interlevel relations mentioned above and simply designate levels as higher or lower.

A multi-level system in this sense is any system where a change in the higher level must be accompanied by a change in the lower level. This segues into complex systems. Besides having at least two levels, the last criterion for a system needed to make my argument go through is that there be more than one entity or process at each of the higher and lower levels. This is

---

[83] Unless, as we'll see in section 5.3.4, we are trying to rearrange system boundaries for the express purpose of eliminating downward causation, a change in system boundaries will not impact the possibility of downward causation occurring in that system.

[84] To clarify: the system itself could be thought of as a level. In this case, there would then need to be at least two further levels below that of the whole system. I am using level terminology such that the system itself is not a level, but internally encompasses levels.

[85] This is in line with the notion of supervenience addressed by, for instance, Kim (1984). Since I am confining my discussion to single systems, the distinction between weak and strong supervenience is not relevant. Other authors (Stalnaker 1996) have defined supervenience differently, but these notions are both sometimes problematic, and most relevantly, stronger than is needed for downward causation to hold. As such, there is no reason to adopt extra problems if there is no pay-off.

intended to be as basic and inclusive a characterization of a complex system as possible: so long as there are multiple parts on at least two levels, the system is sufficiently complex.[86]

This is the broadest characterization, then, of a multi-level complex system: there must be at least two levels, one of which (designated as higher) cannot change without the other (designated as lower) also changing; and at each level there must be at least more than one entity or process. If a system has at least two such levels with multiple entities or processes at each level, then it qualifies as a multi-level complex system and the following case for downward causation applies to it.

We turn now to a basic condition on any causal relata. A commonly made argument against downward causation is that it conflates some noncausal relationship with causation. In many cases, this is true—some proponents of downward causation have proffered examples that label as causal a relationship that is better characterized as supervenient. Constitution is an example of such a noncausal supervenient relation. Consider Sperry's example of a wheel rolling down the hill and the molecules that make up the wheel (Sperry, 1969). He argues that the wheel, as a higher level entity, causes the molecules of the wheel to be where they are because of where the cart goes: the higher level entity causes the lower level entities to move. While Sperry takes himself to be defending downward causation, this is a clear case of confusing constitution with causation. The movement of the wheel does not cause the movement of the molecules. Rather, the movement of the wheel *is constituted by* the movement of the molecules.

---

[86] Again, most other characterizations of complexity will also meet this criteria, since it is deliberately weak and broad. There may be an exception: Mitchell's (2003) constitutive complexity certainly falls into this broad category, but her notion of dynamic complexity does not.

While relationships of constitution are not appropriately labeled downward causation, neither is it upwards causation: in cases of constitution, the lower level entities or processes do not cause the higher level ones, they constitute them. Daniel Wegner has been taken to task for making this particular mistake, in an insightful piece by van Duijn and Bem (2006). Wegner represents the relationship between neuronal processes and conscious awareness as one where lower level neuronal processes cause higher level awareness. van Duijn and Bem argue that the lower level constitutes the higher, rather than causing it. While one may quibble with labeling this relationship constitutive, their negative point is clear: the existence of this other relationship between neuronal entities and conscious awareness that supervenes on them is not causal.

It is a long-recognized condition on attributing causal relationships that the relata of said relation must be logically independent in a way that does not obtain in the cases of entities or processes in a supervenient relationship (see especially Dardis 1993, also Lewis 2000 and Andersen 2006). The underlying idea is that insofar as relata have some kind of logical relationship, such as constitution, emergence, or realization, then this precludes there also being a causal relationship between the same relata.

This requirement makes sense for causal claims. In the interventionist literature on causation, for instance, there must be the at-least-in-principle possibility of intervening on the effect without also intervening on the cause. This is not possible when the lower and higher levels stand in a relationship like that of constitution: it is logically impossible to intervene on the higher level entity or process without thereby, and by dint of, intervening on the lower level entities or processes out of which it is constituted.

This is made intuitively clear in an example. Consider two different variables: Americans who voted in the last election, and female Americans aged 25-40 who voted in the last election.

It doesn't make sense to ask about a causal relationship between these two, because they overlap: the latter is a subset of the former. This means that every time we have an instance of a female American between 25 and 40 who voted in the last election, we also have an instance of an American who voted in the last election. To say that the broader variable causes the narrower would be tantamount to saying that, for each individual who counts as a female American between 25 and 40, their going to vote on each instance was also and at the same time a cause of their going to vote at the same instance. Whatever else we want to say about causation, instances of a cause cannot be at the same time identical with instances of the effect. The logical relationships that preclude causal relationships are all also weakly supervenient relationships. Where there cannot be a change in the higher level without there also being a change in the lower level on which it supervenes, there exists the kind of logical relationship between relata that precludes there also being a causal relationship.

One further constraint is needed before spelling out the details of downward causation in complex multi-level systems. While we are considering the way systems can be represented in terms of level relationships, it must be emphasized that it is the entities or processes at a given level that are the candidates for causal relata. It is not levels themselves. To treat levels as possible causal relata would be an inappropriate reification, and an unnecessary one, too. Levels are a way of classifying features of entities and processes, and indicating generalizations that might be made about the sorts of relationships that hold between things at different levels. What is capable of standing in various relationships to one another, such as causation and constituency, are entities and processes. An entity at a higher level is constituted by some entities or processes at a lower level, but not *by* that lower level, or by all the entities at that lower level. Two levels

are only in a relationship of, say, constitution because of the relationship that exists between the entities and processes; levels are entirely dependent on the things that are at levels.

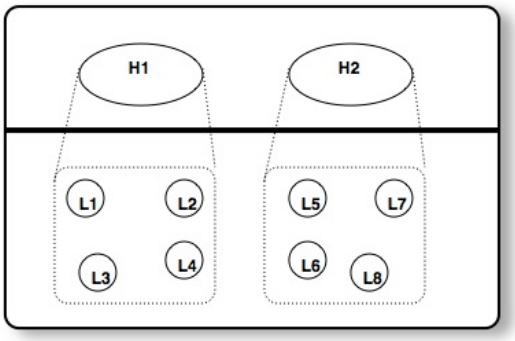
These features create the possibility of downward causation by providing constraints on relata that must be met in order for there to even be the possibility of upward or downward causation. One of the primary intuitions against the metaphysical viability of downward causation is that it violates precisely the first condition discussed above. The intuition is that to claim downward causation, we must attribute causal relationships to entities or processes that already stand in some kind of generically supervenient relationship (see section 5.3.4). Thus, in order to argue for a substantive notion of downward causation, we must meet this constraint: the causal relata of downward causation cannot also and at the same time stand in a supervenient relationship. Further, when considering downward causation, we must do so in terms of entities and processes at different levels, not in terms of the levels themselves as relata.

5.3.3 Causal articulation in complex systems

In this section we see how complex multi-level systems allow for causal relationships between levels while respecting the constraint on logical independence between causal relata. The key to downward causation is a kind of causal articulation that is possible in these kinds of systems. I'll start with a diagram and then bring out the idea of causal articulation by contrasting a vicious and innocuous sense of self-causation.

Figure 6: Complex multi-level system

Dotted lines indicate supervenient relationships between entities and processes at higher and lower levels.



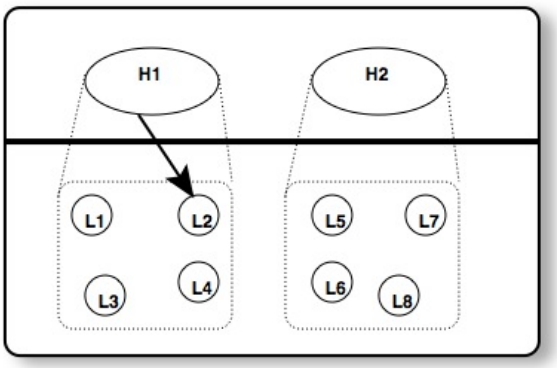
In Figure 6, there is a system (box) with at least two internal levels (divided by black line), and there are at least two things at each level, as per the requirements on multi-level complex systems in the previous section. I have not specified the interlevel relationship, because the argument is intended to be generic.

Consider the question of whether or not this system could causally influence itself. There are two distinct kinds of self-causation one could attribute to this system, the first vicious, the second innocuous. In the vicious version of self-causation, the relata violate the condition that causal relata not already be in a supervenient relationship. In the innocuous version, a higher level entity or process has a causal effect on lower level ones, while also meeting the conditions on causal relata.

First to vicious self-causation. In Figure 7, an arrow representing causal influence from H1 to L2 has been added. This is downward causation, insofar as it is from the higher level entity or process and affects the lower level entity or process, and it is vicious insofar as the influence by H1 is on the supervenience base of H1.

Figure 7: Vicious self-causation

Solid line indicates causal relation from higher level entity or process to entity or process in its own supervenience base.



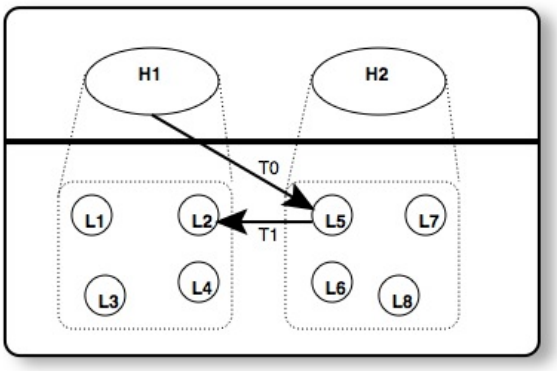
By asserting a causal relationship between H1 and component or subsystem which makes up H1, the condition that causal relata be logically independent is not met. H1's causing of L2 would have to count as both cause and effect in the change in L2. In other words, there is no way for H1 to causally influence L2 except by means that must already include L2: we have to call L2 both a cause and its effect. In vicious self-causation, a higher level cause is construed as causally influencing its own supervenience base.

However, we have the materials to avoid higher level relata causing their own supervenience bases. We can now invoke two ideas: that the relata in causation are entities and processes, not the levels themselves; and that at each of the two levels there are multiple entities and processes. When these conditions are met, the possibility of causal articulation within a system is generated, and this causal articulation allows for innocuous self-causation.

In innocuous self-causation, the cause is a higher level entity or process, and the effect is a lower level one, but a lower level one that *is not in the supervenience base of the cause*. In figure 8, an arrow representing causal influence is added from H1 to L5. This is downward causation, a higher level entity or process affecting a lower level one, and it is innocuous insofar as there is no logical relation between the cause and effect.

Figure 8: Innocuous self-causation

Solid line indicates causal relation at time T0 from higher level entity or process to entity or process at a lower level but not in its own supervenience base, and a causal relation from that lower level entity or process at some later time T1 to the supervenience base of the higher level entity or process.



Note also that the arrow from H1 to L5 is time indexed: at some moment (or, really, any length of time if need be), H1 causally influences L5. At some later time (even if only just slightly later), L5 is then represented as causally influencing L2. And this influence on L2 also constitutes an influence on H1. The innocuousness of this type of self-causation has to do with the mediation of influence of H1 on its own supervenience base via a lower level entity or process that is not part of that base. H1 does not directly causally affect its own supervenience base, but it can do so indirectly, by affecting lower level entities and processes which are distinct from its supervenience base, which in turn affect H1's supervenience base. Synchronically, a higher level entity or process can *directly* influence lower level entities not in its own supervenience base. Diachronically, it can *indirectly* influence its own supervenience base, thereby influencing itself.

This example displays what I mean by causal articulation. Complex, multi-level systems have subsystems or components that can interact with one another causally; these subsystems in turn have subsystems or components, which can also interact causally. There is no intrinsic reason why levels should constrain causal relations, so long as those causal relations are between

independent relata. The existence of multiple distinct but interacting subsystems or components renders a system causally heterogeneous, able to internally affect itself in a way unavailable to causally homogeneous systems.

We can now clearly state the features of systems which, in conjunction with an emphasis on entities and processes as causal relata, rather than the levels themselves, allow for causal articulation of parts in the system: i) at least two levels; ii) more than a single entity or process at each level; and iii) the condition that nothing directly influence its own supervenience base. Whenever a complex system meets these conditions, there is the possibility of downward causation. This possibility cannot be ruled out *a priori* based on metaphysical considerations, nor will every system that meets these conditions necessarily engage in downward causation. Whether such a system engages in downward causation remains an open empirical question to be answered on a case-by-case basis.

5.3.4 Gerrymandering system and level boundaries

The account of downward causation presented so far has been pitched at an abstract level, without reference to any specific system or in terms of any specific relationship between levels. There are two major concerns that need to be addressed about this account. Both of these concerns revolve around the idea that the downward causation detailed in the last section is only apparent, and that the appearance of downward causation can be eliminated by a better arrangement of either the levels in the system, or of the boundaries of the system itself. I will address each of these in turn, and argue that while it is possible in some cases to gerrymander either the levels or the system boundaries to eliminate causation that crosses levels, it is always ad hoc to do this for the sole purpose of avoiding downward causation. Furthermore, in many

cases it is not possible to do this kind of gerrymandering, since we have independent reasons for using a particular breakdown of levels or a particular system boundary, so that the necessary rearrangement would be counterproductive to analysis. This section concludes by uncovering the fundamentalist assumption which underlies both of these concerns: that the only 'real' causes are lower level, and that even if the appearance of downward causation cannot be eliminated, it is still only appearance. I have already addressed this issue elsewhere. In this section, I do not rebut this concern entirely, but confine it to a very limited space in which it remains a concern only as part of a metaphysical position that makes no possible difference to reality.

I will start by discussing the possibility of rearranging system boundaries in order to eliminate cross-level causation. This is a tactic used by, among others, Robert Klee (1984). His basic solution is that whenever we find something that appears to be downward causation between levels in a system, we should instead consider the causal relata as two different systems interacting with one another. When we find apparent examples of downward causation, he says, we are ignoring the differences "between determinative connections between levels in a system, and determinative connections between two independently functioning systems" (1984, 58). Anytime one of these 'determinative connections' could count as downward causation, this is an indication that we should instead break the system up into two distinct systems that causally interact, thereby eliminating the appearance of downward causation.

First, it is key to note that the motivation for Klee to offer this way out of downward causation is the worry that the relationships between levels are being mistaken for causal relationships. He thinks that the appearance of downward causation is a consequence of committing what is, on any account, a fallacy: conflating the logical relationships between levels—in his article, specifically that of emergence—with genuinely causal relationships. And

in fact, this conflation has been made by advocates of downward causation, and if downward causation were only possible in this way, Klee would be correct in his denial of it. However, his motivating worry does not address the kind of system elaborated above, where there are distinct entities or processes at each level, such that a claim of downward causation can be made in explicit distinction to the relationship between levels.

Second, having addressed the concern regarding conflation of interlevel and causal relations, there is no particular reason why we should seek to eliminate downward causation, by rearranging systems or in any other way. If downward causation does not violate logical constraints on causal relata, then we lack a reason to avoid it. Positing a rearrangement of systems solely for the purpose of avoiding downward causation is ad hoc. We have no reason to think it metaphysically unviable, and so no pressing reason to eliminate it. If we make the minimal presumption that there was some reason for the original system boundaries, then we need a reason to alter those boundaries, and in the case of causally articulated complex systems, simply wanting to avoid downward causation is not itself a reason: it is just gerrymandering.

This leads to the third point regarding the rearrangement of system boundaries: not only is it ad hoc, but it may be counterproductive, and we cannot assume it is always available as an option. There will be cases in which overriding considerations point toward one way of dividing up systems. The fact that this division leads to downward causation cannot be helped, since there were compelling reasons to use the original system individuation. It is not hard to find examples of this. In population biology, for instance, when considering issues like sexual reproduction and competition, there is something distinctly robust about the treatment of individual animals as entities at one level, in contrast to populations at a higher level, or genomes or other units at a lower level. If cross-level causation appears in these cases, it may not be possible to simply

change what counts as a single system. In cases where there are prior, empirically based reasons for already using one kind of system individuation, there is already an existing argument against changing that system individuation. This means that no general claim can be made for changing system individuations in all cases where we encounter apparent downward causation. Any such claim would have to proceed on a case by case basis, to ensure that there aren't previous reasons which block such re-orientation of system boundaries, and in cases where there are such reasons, the claim that downward causation can be eliminated by system rearrangement will simply fail.

Similar considerations apply to using a rearrangement of levels in order to eliminate the appearance of downward causation. Recent work by Carl Craver and William Bechtel (2007) advocates this kind of approach, where they resist downward causation between levels in complex systems, but offer instead a weakened kind of influence within the context of mechanisms as the interlevel relationship. Mechanisms are treated as higher level wholes, and components of mechanisms—entities and their organized activities—are the lower level.

> In our view, the phrase 'top- down causation' is often used to describe a perfectly coherent and familiar relationship between the activities of wholes and the behaviors of their components, but the relationship is not a causal relationship. Likewise, the phrase 'bottom-up causation' does not, properly speaking, pick out a causal relationship. Rather, in unobjectionable cases both phrases describe mechanistically mediated effects. Mechanistically mediated effects are hybrids of constitutive and causal relations in a mechanism, where the constitutive relations are interlevel, and the causal relations are exclusively intralevel. (Craver and Bechtel 2007)

Craver and Bechtel thus countenance the existence of higher level causes as genuinely causal, rather than as derivatively causal and dependent on lower level or fundamental causes. But, they hold that this causal efficacy of higher level entities can only be directed at other higher level entities, never at a lower level ones. What others have labeled top-down causation, they claim, is always noncausal. They share a very similar concern to Klee: the logical relationships between levels preclude causal relationships, and it is a fallacy to say that the relationship between a

whole and its parts is causal. This motivating issue about the distinction between logical and causal relationships is, as I've addressed with respect to Klee, both a good one, yet also out of place in the context of causally articulated complex systems. It is possible to both respect this distinction and still leave the possibility for direct causal relationships between levels.

There is a more basic disagreement between their view and the view espoused in this section, however. Much of the work of demonstrating that causation never crosses levels is done by their definition of a level: "In levels of mechanisms… an item X is at a lower level than an item S if and only if X is a component in the mechanism for some activity w of S" (Craver and Bechtel 2006, 548). The reason there can be no interlevel causation in their view is because there is nothing else at a level than a single higher level entity (the mechanism), and the components on which it constitutively supervenes. A basic requirement on complex systems, in order to be causally articulated in the appropriate way, is that there be more than a single entity or process on both the higher and the lower levels, and that there be two identifiable levels within the system itself. Craver and Bechtel's level individuations avoid downward causation, but only for cases which didn't qualify in the first place. They haven't addressed causally articulated systems.

As such, there is agreement that a higher level entity cannot directly causally influence the lower level entities and processes on which it supervenes. The disagreement is about whether or not it is a legitimate technique for avoiding downward causation that we always constrict ourselves to only considering mechanisms as the level differentiation, and considering only one mechanism at a time. The issue is to see if it is always possible or justified to take a case of apparent downward causation in a causally articulated system, and rearrange the level boundaries so that the only level comparisons which can be made are those between a single entity or process and the lower level entities or processes on which it supervenes.

Craver and Bechtel's proposal to rearrange levels ends up in the same bind that faced the system-rearrangement proposal of Klee: the option is not always open, and when it is open, such rearrangement is ad hoc. Consider a case where the mechanism is the entire system, the higher level, and it is constituted by several lower level components. There could be no downward causation in such a system, since there is nothing else for the higher level to affect except for its own supervenience base. However, if each of those components are in turn composed of smaller components, then there is the possibility for a mid-level component to directly influence a lowest level component. In other words, any time that components of mechanisms are in turn mechanisms with their own components, there is the possibility of downward causation. eliminating DC in a way consistent with Craver and Bechtel's position leads straight into a dilemma. Their solution is that we ought not consider all three of these levels at the same time: there is one mechanistic relationship, and a single set of levels, between the original mechanisms and its components, and a distinct, non-comparable mechanism and set of levels between those components and the components which constitute them. No downward causation would be possible, in that case.

But the awkwardness of such a stipulation should be clear. Aside from wanting to avoid downward causation, no reason to stipulate that we must ignore the original mechanism and its relationship to its components as we consider those components and the next lower level of components. Even if we could isolate these three levels into groups of two, there is no reason to do so. So long as we keep interlevel supervenience relationships separate from interlevel causal relationships, we can consider all three of these levels simultaneously, if it is useful to do so. It is ad hoc to disallow *a priori* the chance to simultaneously consider a mechanism, its components,

and the components of its components. But as soon as we allow for this kind of consideration, the door is open for downward causation.

Craver and Bechtel are somewhat unusual in treating higher level causes as genuinely causal at higher levels. Klee, in contrast, presents his case against downward causation via emergence as an argument for micro-determinism: only those causes at the lowest levels, namely that of particles and their interactions as described by physics, have genuine causal capacities. Klee represents the mainstream view of microphysicalism or causal fundamentalism (see chapter 4) in this regard. Downward causation as I have elucidated it doesn't address this: one could still argue that the real causal force of the apparent higher level cause resides at the lower level.

I have addressed causal fundamentalism elsewhere, and will not reiterate those arguments here. However, I add to them by showing how, even if one does not accept my other arguments, there is a narrow range in which this is a genuine concern. Like the perniciousness of strong versions of skepticism, the view that all 'real' causation is lower level is difficult to banish completely. But, we can outline the conditions under which causal fundamentalism is actually problematic for causal claims: namely, as an untestable article of metaphysical faith. We could hold a different metaphysical view, as equivalently untestable as causal fundamentalism; or, if we take our causal claims to be primarily empirical, we needn't be concerned with fundamentalism because there is an enormous variety of empirically verified causes that are not fundamental.

First, consider how the view that I have advocated fits in with interventionism. Higher level variables can be considered legitimate causes of lower level variables if those variables meet the same criteria for causation that any other variable must. The condition that a higher level cause cannot directly influence its own supervenience base can be re-expressed in terms of

instances of those variables, as we've seen in chapter 4. Two variables cannot stand in a causal relationship if there are any instances that count as instances of both variables: if a single event occurs that could count as both the cause and as the effect. Other kinds of logical relationships (as we've seen) preclude the possibility of also being in a causal relationship. Besides not already standing in a logical relationship, the variables must meet the basic interventionist criteria (see chapter 3).

The view of downward causation in causally articulated complex systems presented here is quite amenable to the interventionist approach. If wiggling a higher level variable leads to a change in a lower level variables, and the higher does not supervene on the lower (instances of the former are never also instances of the latter), then the higher level variable causes the lower. There is nothing more to the matter. Any causal claim in interventionism is made relative to a set of variables under consideration. A defender of causal fundamentalism might argue that insofar as we claim to have a higher level cause which, when wiggled, changes a lower level effect, we have failed to include the right set of variables; if the appropriate lower level variables were included, we would find there to be no downward causation at all.

What I want to illustrate about this response by causal fundamentalism to an otherwise acceptable application of interventionist criteria to downward causation is that it is a purely metaphysical point. To advocate that there is never such a thing as downward causation, even if it meets other criteria for causation, is to make an in-principle claim about what exactly we would find if we had unlimited or perfect epistemic access to the world. It is one thing to claim that, *for a specific case*, relevant lower level variables have been left out. In such a case, a critic could actually provide the missing variable(s), and demonstrate how they eliminate the appearance of downward causation. But to make the broader point that there must always be

such a variable missing, it is no longer possible to provide the missing variable(s) and demonstrate how the appearance of downward causation goes away.

The broader point, an essential premise for causal fundamentalism, is that there must always be such a variable even though it is not actually provided. Such a claim is at root a metaphysical commitment that is not itself supported by evidence. Given complete epistemic access to the world, it is possible that we would always be able to find lower level variables that obviate the higher level ones, thus finding that there is no such thing as downward causation. But it is also possible that we wouldn't find this: that instead we would find irreducible instances of downward causation, where no additional lower level variables could adequately capture the causal efficacy attributed to a higher level variable. And, we most assuredly lack the perfect epistemic access needed to adjudicate the in-principle claim that it is always lower level causes that are genuinely efficacious.

Thus, we find that the claim by causal fundamentalism that there is always some further variable that would eliminate downward causation is a matter of metaphysical faith. If one does not share this faith, then one is not in any worse of a position vis-à-vis downward causation: neither view can be finally adjudicated with the sort of epistemic access we have. But, there is still the prima facie standing case in favor of downward causation.[87] Unless one wants to debunk the entirety of interventionism (which may be possible but would be a gargantuan task), we currently have more reason to think that there is the possibility of downward causation than we do to think there isn't. This constrains the range in which the metaphysical challenge of causal

[87] There are further responses left open for the causal fundamentalist, but again, these are addressed in chapter 4.

fundamentalism is problematic: for those of us who do not want to wager on perfect epistemic access, causal fundamentalism lacks teeth.

5.3.5 Section conclusion

I have developed the theory of downward causation in this section in deliberately abstract terms. In any system with at least two levels, one of which supervenes on the other, and where there are at least two distinct entities or processes at each level, there is the possibility of upward and downward causation in addition to the noncausal interlevel relation. Having a theory of downward causation that is applicable to different interlevel relationships, and a theory that is compatible with a broadly interventionist framework, is generally useful. My motivation for the project, however, concerns its ability to shed light on a particular problem: that of the causal efficacy of conscious agency.

This is a somewhat oversimplified but still useful characterization of the current situation regarding the causal efficacy of conscious awareness: it is unclear whether there is any substantive notion of downward causation with which one could argue for the higher level causation that conscious causal efficacy would have to have if it had any at all. It is *a fortiori* unclear whether there is such a notion that is suitable for conscious causal efficacy given that existing accounts of downward causation are usually given in regards to specific interlevel relationships (usually emergence or constitution) and we do not know what the interlevel relationship is in the case of conscious agency.

This theory of downward causation in causally articulated complex systems answers the need for a viable theory of downward causation in general, especially for one that can be applied in the debate about conscious agency, where we lack a clear specification of the interlevel

relationship. I have demonstrated how, in specific kinds of systems, there is a metaphysically innocuous sense of downward causation that is possible. Whether or not there is downward causation in any specific system that meets these criteria is an empirical questions that must be answered on a case by case basis.

This section primarily presents a possibility claim. However, it is an extremely useful possibility claim to establish: there is nothing philosophically suspect about downward causation, properly conceived; and we need not specify the interlevel relationship to see if a given system exhibits it. We can then turn to cognitive science to see if conscious agency can be represented so as to meet the criteria for downward causation, and if the higher level variables for conscious awareness are causally influential on lower level variables representing physical processes.

5.4 Conclusion

The goal of this and the previous chapter has been the establishment of the metaphysical legitimacy of higher level causes as genuinely causal, particularly for higher level causes that involve conscious awareness. There are several interrelated challenges to this thesis that I have addressed.

In chapter 4, I rebutted the claim that individual instances of causation are always lower level and only apparently higher level. What we are trying to explain, when we explain a single instance of causation, is not merely the specific collection of particle bumpings and conserved quantity transfers that connect the cause and the effect. We are trying to explain this particular occurrence of a phenomenon, where that phenomenon would still have occurred had the sequence of bumpings been somewhat different than they actually were. Part of the explanandum

is this counterfactual robustness of the phenomenon. This counterfactual robustness can be captured using higher level properties and causes, and is lost when we are constrained to use only lower level causes, since these add a spurious specificity to the explanation and render it counterfactually fragile.

I started chapter 5 by addressing the claim that the higher level variables I offered in chapter 3 were somehow spurious or mere placeholders for the 'really' causal variables, which would be at a lower level. This claim runs into problems when we try to cash it out in terms of replacement variables. If the lower level variables are simply replacements of the higher level ones, in the sense that they maintain precisely the same extension of instances, then the higher level and lower level variables are really just the same variable, described in two different ways. Unless the lower level description provides us something that the higher level doesn't, such as a better way to identify new instances, then we have no reason to privilege one way of expressing this variable over another. Higher level intensions will generally be more useful for identification of instances than lower level, gerrymandered, ones, which contradicts the intuition that the lower level is more 'real'. On the other hand, if the lower level variables don't preserve the same extension, then they aren't actually replacement variables for the original higher level ones. Instead, they are a different way to divide the phenomena up into variables. In this case, the onus is on the lower level advocate to defend why the parsing of instances in the new way is scientifically preferable to the old. Lower level variables simply aren't a metaphysically superior replacement for the higher level variables.

Yet a critic could still defend the metaphysical primacy of lower level or micro causation by saying that it is not *qua* higher level cause that these instances are causally efficacious, but only by dint of their microphysical instantiation. My response began from the view that single

instances of causation have a counterfactual buffer zone around them, such that changes within that buffer zone could have occurred without altering the causal relationship that did actually occur. Looking just at one instance of causation at a time, we have no way to evaluate counterfactuals with respect to these buffer zones, and thus it may appear as if the microphysical instantiation is doing all the causal work. But using general or variable causation, we have the means to evaluate these counterfactuals, by seeing what variables this instance of causation counts as an instance of. This allows us to use actual occurrences of causation to evaluate what could or could not have been different in a particular instance without altering the causal relationship. It also provides justification to say that it was not the actual microphysical instantiation on any given occasion that was really causally efficacious in bringing about the effect, since any number of details about that instantiation could have been different. Instead, it is the instance's membership in higher level or macrophysical variables that cannot be changed without changing the causal relationship; this membership provides grounds to say that it was by dint of being an instance of that variable that this cause brought about the effect.

Finally, I demonstrated that there is no particular metaphysical problem with downward causation from higher to lower levels within a system, so long as that system meets the requirements for causal articulation and the influence is not on the supervenience base of the higher level cause. We should not feel compelled to gerrymander our system boundaries or level differentiations in order to avoid downward causation.

Thus, not only are higher level causes like those involving conscious awareness metaphysically legitimate when used as variables, they are also metaphysically legitimate in single instances of causation, and in causation from higher to lower level entities or processes. We have no reason to insist that, against the evidence and against our best epistemic grasp of the

world, higher level causes are only apparent or that only the very small can be causally efficacious. This holds for awareness as well as for other kinds of higher level causes. We should accept that causes occur at many different size scales and degrees of complexity.

REFERENCES

Andersen, Holly (2006), "Two Causal Mistakes in Wegner's *Illusion of Conscious Will*", *[PSA 2006] Philosophy of Science Assoc. 20th Biennial Mtg (Vancouver): PSA 2006 Contributed Papers*. [philsci-archive.org]

Banich, Marie T (2004), *Cognitive Neuroscience and Neuropsychology*. Wadsworth Publishing.

Bargh, John (1994), "The Four Horsemen of Automaticity: Awareness, Intention, Efficiency, and Control in Social Cognition", in Thomas K. Srull and Robert S. Wyer, Jr., *The Handbook of Social Cognition*. Hillsdale, NJ: Lawrence Earlbaum Associates, 1-40.

Bargh, John, and Tanya Chartrand (1999), "The Unbearable Automaticity of Being", *American Psychologist* 54(7): 462-479.

Bargh, John, P.M. Gollwitzer, A. Lee-Chai, K. Barndollar, and R. Trotschel (2001), "The Automated Will: Nonconscious Activation and Pursuit of Behavioral Goals", *Journal of Personality and Social Psychology* 81(6): 1014-1027.

Bayne, Tim, and Neil Levy (2006), "The Feeling of Doing: Deconstructing the Phenomenology of Agency", in Natalie Sebanz and Wolfgang Prinz (eds.), *Disorders of Volition*. Cambridge, MA: The MIT Press.

Bechtel, William, and Robert C. Richardson (1993), *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Princeton, NJ: Princeton University Press.

Bedau, Mark (2002), "Downward Causation and the Autonomy of Weak Emergence", *Principia* **6**: 5-50.

Beebee, Helen (2006), *Hume on Causation*. New York: Routledge.

Bickle, John (2003), *Philosophy and Neuroscience: A Ruthlessly Reductive Account*. Dordrecht: Kluwer Academic Publishers.

Bogen, James (2004), "Analyzing Causality: The Opposite of Counterfactual is Actual", *International Studies in the Philosophy of Science* 18(1): 3-26

Breitmeyer, Bruno (1985), "Problems with the Psychophysics of Intention", *Behavioral and Brain Sciences* 8(4): 539-540.

Callender, Craig (1999), "Reducing Thermodynamics to Statistical Mechanics: The Case of Entropy", *The Journal of Philosophy* 96(7): 348-373.

Campbell, John (2007), "An Interventionist Approach to Causation in Psychology", in Alison Gopnik and Laura Schulz (eds.), *Causal Learning: Psychology, Philosophy and Computation*. Oxford University Press, 58-66.

Campbell, John (2008), "Causation in Psychiatry", in Kenneth Kendler and Josef Parnas (eds.), *Philosophical Issues in Psychiatry: Explanation, Phenomenology, Nosology*. Baltimore, MA: Johns Hopkins University Press.

Carruthers, Peter (2007), "The Illusion of Conscious Will", *Synthese* 96: 197-213.

Cartwright, Nancy (1979), "Causal Laws and Effective Strategies", *Nous* 13(4): 419-437.

Cartwright, Nancy (1983), *How the Laws of Physics Lie*. Oxford: Clarendon Press.

Cartwright, Nancy (1999), *The Dappled World: A Study of he Boundaries of Science*. Cambridge University Press.

Cartwright, Nancy (2001), "Modularity: It Can - And Generally Does - Fail", in Maria Carla Galavotti, Domenico Costantini, and Patrick Suppes (eds.), in *Stochastic Causality*. Stanford, CA: CSLI Publications, 65-85.

Cartwright, Nancy (2002), "Against Modularity, the Causal Markov Condition, and Any Link Between the Two: Comments on Hausman and Woodward", *British Journal for the Philosophy of Science* 53(3): 411-453.

Chalmers, David (1996), *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Choudhury, Suparna, and Sarah-Jayne Blakemore (2006), "Intentions, Actions, and the Self", in Susan Pockett, William P. Banks and Shaun Gallagher (eds.), *Does Consciousness Cause Behavior?* Cambridge, MA: The MIT Press.

Cole, Jonathan (2007), "The Phenomenology of Agency and Intention in the Face of Paralysis and Insentience", *Phenomenology and the Cognitive Sciences* 6: 309-325.

Craver, Carl; and William Bechtel (2007), "Top-Down Causation without Top-Down Causes", *Biology and Philosophy* 22(4): 547-563.

Craver, Carl (2007), *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford University Press.

Cunnington, Ross, C. Windischberger, L. Deecke, and E. Moser (2002), "The Preparation and Execution of Self-Initiated and Externally-Triggered Movement: A Study of Event-Related fMRI", *Neuroimage* 15: 373-385.

Dardis, Anthony (1993), "Sunburn: Independence Conditions on Causal Relevance", *Philosophy and Phenomenological Research* 53(3): 577-598.

Dardis, Anthony (2008), *Mental Causation*. New York, NY: Columbia University Press.

Davidson, Donald 1980, "Intending", reprinted in *Essays on Actions and Events*. Oxford University Press, 1984, 83-102.

Descartes, Rene (1996), *Meditations on First Philosophy*, translated by John Cottingham. Cambridge University Press.

Devine, Patricia G., M.J. Monteith, J.R. Zuwerink, and A.J. Elliot (1991), "Prejudice with and without Compunction", *Journal of Personality and Social Psychology* 60(6): 817-830.

Dowe, Phil (2000), *Physical Causation*. Cambridge University Press.

Drain, Sean, and Anthony Greenwald (1998), "Replicable Unconscious Semantic Priming", *Journal of Experimental Psychology* 127(3): 286-303.

Eberhardt, Frederick, and Richard Scheines (2007), "Interventions and Causal Inference", *Philosophy of Science* 74: 981–995

Emmeche, Claus, Simo Køppe, and Frederik Stjernfelt (2000), "Levels, Emergence, and Three Versions of Downward Causation", in P.B. Andersen, C. Emmeche, N.O. Finnemann, and P.V. Christiansen (eds.), *Downward Causation: Minds, Bodies, and Matter*. Aarhus: Aarhus University Press, 13-34.

Farrer, Chloe, N. Franck, N. Georgieff, C.D. Frith, J. Decety, and M. Jeannerod (2003), "Modulating the Experience of Agency: A Positron Emission Tomography Study." *Neuroimage* 18:324-333.

Farrer, Chloe, and C.D. Frith (2002), "Experiencing Oneself vs. Another Person as Being the Cause of an Action: The Neural Correlates of the Experience of Agency", *Neuroimage* 15:596-603

Fodor, Jerry (1974), "Special Sciences (or: The Disunity of Science as a Working Hypothesis", *Synthese* 28(2):97-115.

Forster, K.I. (1981), "Priming and the Effects of Sentence and Lexical Contexts on Naming Time: Evidence for Autonomous Lexical Processing", *The Quarterly Journal of Experimental Psychology* 33(4):465-495.

Frith, Chris (2002), "Attention to Action and Awareness of Other Minds", *Consciousness and Cognition* 11(4):481-487.

Gallagher, Shaun (2000), "Self-Reference and Schizophrenia: A Cognitive Model of Immunity to Error through Misidentification", in Dan Zahavi (ed.) *Exploring the Self: Philosophical and Psychopathological Perspectives on Self-experience.* Amsterdam and Philadelphia: John Benjamins, 203-239.

Gallagher, Shaun (2006), "Where's the Action? Epiphenomenalism and the Problem of Free Will", in Susan Pockett, William P. Banks and Shaun Gallagher (eds.), *Does Consciousness Cause Behavior?* Cambridge, MA: The MIT Press, 109-124.

Gallagher, Shaun (2007), "The Natural Philosophy of Agency", *Philosophy Compass* 2: 347-357.

Gazzaniga, Michael (1998), *The Mind's Past.* Berkeley, CA: University of California Press.

Gazzaniga, Michael (2000), "Cerebral Specialization and Interhemispheric Communication: Does the Corpus Callosum Enable the Human Condition?", *Brain* 123:1293-1326.

Gilden, Lloyd, H.G. Vaughn, and L.D. Costa (1966), "Summated Human EEG Potentials with Voluntary Movement", *Electroencephalography and clinical neurophysiology* 20(5):435-438.

Gjelsvik, Olav (1990), "On the Location of Actions and Tryings: Criticism of an Internalist View", *Erkenntnis* 33(1):39-56.

Glymour, Clark (2007) "When is a Brain like the Planet?", *Philosophy of Science* 74: 330-347.

Gomes, Gilberto (1998), "The Timing of Conscious Experience: A Critical Review and Reinterpretation of Libet's Research", *Consciousness and Cognition* 7(4):559-595.

Gomes, Gilberto (2002), "The Interpretation of Libet's Results on the Timing of Conscious Events: A Commentary", *Consciousness and Cognition* 11(2):144-161.

Greenwald, Anthony G., and Mahzarin R. Banaji (1995), "Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes", *Psychological Review* 102(1):4-27.

Hacking, Ian (1982), "Experimentation and Scientific Realism", *Philosophical Topics* 13:154-172.

Haggard, Patrick, and Martin Eimer (1999), "On the Relation Between Brain Potentials and the Awareness of Voluntary Movements", *Experimental Brain Research* 126: 128-133.

Haggard, Patrick, Chris Newman, and Elena Magno (1999), "On the Perceived Time of Voluntary Actions", *British Journal of Psychology* 90:291-303.

Haggard, Patrick, Sam Clark, and Jeri Kalogeras (2002), "Voluntary Action and Conscious Awareness", *Nature Neuroscience* 18 March 2002, DOI: 10.1038/nn827.

Haggard, Patrick (2003), "Conscious Awareness of Intention and of Action", in Johannes Roessler and Naomi Eilan (eds.), *Agency and Self-Awareness*. Oxford University Press, 111-127.

Haggard, Patrick, and Jonathan Cole (2007), "Intention, Attention and the Temporal Experience of Action", *Consciousness and Cognition* 16(2):211-220.

Hall, Ned (2004), "Two Concepts of Causation", in John Collins, Ned Hall, and L.A. Paul (eds.), *Causation and Counterfactuals*. Cambridge, MA: The MIT Press, 225-276.

Hardcastle, Valerie Gray (1998), "On the Matter of Minds and Mental Causation", *Philosophy and Phenomenological Research* 58(1):1-25.

Horgan, Terence (1989), "Mental Quausation", *Philosophical Perspectives*, Vol. 3, Philosophy of Mind and Action Theory, 47-76.

Horgan, Terence (1993), "From Supervenience to Superdupervenience: Meeting the Demands of a Material World", *Mind* 102(408):555-586.

Horgan, Terence, John L. Tienson, and George Graham (2003), "The Phenomenology of First Person Agency", in Swen Walter and Heinz-Dieter Heckmann (eds.), *Physicalism and Mental Causation: The Metaphysics of Mind and Action*. Charlottesville, VA: Imprint Academic, 323-40.

Hornsby, Jennifer (1980), *Actions*. London: Routledge and Kegan Paul.

Jack, Anthony and Tim Shallice (2001), "Introspective Physicalism as an Approach to the Science of Consciousness", *Cognition* 79: 161-196

Jackson, Frank and Philip Pettit (1990), "Program Explanation", *Analysis* 50(2):107-117.

Joordens, Steve and Suzanna Becker (1997), "The Long and Short of Semantic Priming Effects in Lexical Decision", *Journal of Experimental Psychology: Learning, Memory, and Cognition* 23(5):1083-105.

Joordens, Steve, Marc van Duijn, and Thomas Spalek (2002), "When Timing the Mind One Should also Mind the Timing: Biases in the Measurement of Voluntary Actions", *Consciousness and Cognition* 11:231-240.

Kim, Jaegwon (1984), "Concepts of Supervenience", *Philosophy and Phenomenological Research* 45(2):153-176.

Kim, Jaegwon (1998), *Mind in a Physical World*. Cambridge, MA: The MIT Press.

Kim, Jaegwon (1999), "Making Sense of Emergence", *Philosophical Studies* 95:3-36.

Kitcher, Philip (1984), "1953 and all That. A Tale of Two Sciences", *The Philosophical Review* 93(3):335-373.

Kitcher, Philip (1991), “Explanatory Unification”, in Richard Boyd, Philip Gaspar, and J.D. Trout (eds.), *The Philosophy of Science*. Cambridge, MA: The MIT Press.

Klee, Robert (1984), “Micro-Determinism and Concepts of Emergence”, *Philosophy of Science* 51(1):44-63.

Koch, Christof (2004), *The Quest for Consciousness: A Neurobiological Approach*. Englewood, CO: Roberts and Company Publishers.

Kornhuber, Hans-Helmut and Luder Deeke (1965), “Cerebral Potential Changes in Voluntary and Passive Movements in Man: Readiness Potential and Reafferent Potential”, *Pflügers Archiv für die Gesamte Physiologie*, 284:1-17.

Lau, Hakwan, Robert Rogers, Patrick Haggard, and Richard Passingham (2004), “Attention to Intention”, *Science* 303:1208.

Lewis, David (2000), “Causation as Influence”, *The Journal of Philosophy*, 97(4):182-197.

Libet, Benjamin (1965), “Cortical Activation in Conscious and Unconscious Experience”, *Perspectives in Biology and Medicine* 9:77-8.

Libet, Benjamin, W.W. Alberts, E.W. Wright Jr, E. Feinstein (1967), “Responses of Human Somatosensory Cortex to Stimuli Below Threshold for Conscious Sensation”, *Science* 158:1597-1600.

Libet, Benjamin (1985), “Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action”, *Behavioral and Brain Sciences* 8:529-566.

Libet, Benjamin (1989), “The Timing of a Subjective Experience”, *Behavioral and Brain Sciences* 12:183-85.

Libet, Benjamin (2003), “Can Conscious Experience Affect Brain Activity?” *Journal of Consciousness Studies* 10(12):24-28.

Libet, Benjamin, E.W. Wright, B. Feinstein, D.K. Pearl (1979), “Subjective Referral of the Timing for a Conscious Sensory Experience”, *Brain* 102:193-224

Machamer, Peter, Lindley Darden, and Carl Craver (2000), “Thinking about Mechanisms”, *Philosophy of Science* 67(1)1-25.

McDowell, John (1994), “The Content of Perceptual Experience”, *The Philosophical Quarterly* 44(175):190-205.

McLaughlin, Brian (1989), “Type Epiphenomenalism, Type Dualism, and the Causal Priority of the Physical”, *Philosophical Perspectives* 3:109-135.

McLaughlin, Brian and Karen Bennett (2008), "Supervenience", Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Fall 2008 Edition)*, URL = <http://plato.stanford.edu/archives/fall2008/entries/supervenience/>.

Mele, Alfred (1992), *Springs of Action: Understanding Intentional Behavior*. Oxford University Press.

Mele, Alfred (2003), *Motivation and Agency*. Oxford University Press.

Mele, Alfred (2007), "Decisions, Intentions, Urges, and Free Will: Why Libet Has Not Shown What He Says He Has", in Joseph Keith Campbell, Michael O'Rourke, Harry S. Silverstein (eds.), *Causation and Explanation*. Cambridge, MA: The MIT Press, 241-264.

Mele, Alfred (2008), "Proximal Intentions, Intention-Reports, and Vetoing", *Philosophical Psychology* 21(1):1-14.

Menzies, Peter (2003), "The Causal Efficacy of Mental States," in Swen Walter and Heinz-Dieter Heckmann (eds.), *Physicalism and Mental Causation: The Metaphysics of Mind and Action*. Charlottesville, VA: Imprint Academic, 195-223.

Menzies, Peter, and Huw Price (1993), "Causation as a Secondary Quality", *British Journal for the Philosophy of Science* 44(2):187-20

Metzinger, Thomas (2000), "The Subjectivity of Subjective Experience: A Representationist Analysis of the First-Person Perspective", in Thomas Metzinger (ed.), *Neural Correlates of Consciousness*. Cambridge, MA: :The MIT Press.

Metzinger, Thomas (2006), "Conscious Volition and Mental Representation: Toward a More Fine-Grained Analysis", in Natalie Sebanz and Wolfgang Prinz (eds.), *Disorders of Volition.* Cambridge, MA: The MIT Press.

Meyering, Theo (2000), "Physicalism and Downward Causation in Psychology and the Special Sciences" *Inquiry* 43(2):181-202.

Mill, J.S. (1862), *A System of Logic, Ratiocinative and Inductive*. London: John W. Parker.

Mitchell, Sandra (1997), "Pragmatic Laws", *Philosophy of Science*, Vol. 64, Supplement. Proceedings of the 1996 Biennial Meetings of the Philosophy of Science Association. Part II: Symposia Papers, S468-S479.

Mitchell, Sandra (2000), "Dimensions of Scientific Law", *Philosophy of Science* 67(2):242-265.

Mitchell, Sandra (2003), *Biological Complexity and Integrative Pluralism*. Cambridge University Press.

Mitchell, Sandra (2008a), "Exporting Causal Knowledge in Evolutionary and Developmental Biology", *Philosophy of Science* 75:697-706.

Mitchell, Sandra (2008b), "Explaining Complex Behavior", in Kenneth Kendler and Josef Parnas (eds.), *Philosophical Issues in Psychiatry: Explanation, Phenomenology, Nosology*, Baltimore, MA: Johns Hopkins University Press.

Moskowitz, Gordon B. (2002), "Promiscuous Effects of Temporary Goals on Attention," *Journal of Experimental Social Psychology* 38:397-404.

Mossel, Benjamin (2005), "Action, Control, and Sensations of Acting", *Philosophical Studies* 124:129-180.

Nahmias, Eddy (2005), "Agency, Authorship, and Illusion", *Consciousness and Cognition* 14:771-785.

Newell, Allen (1992), "SOAR as a Unified Theory of Cognition: Issues and Explanations", *Behavioral and Brain Sciences* 15:464-492.

Newman, Leonard and James Uleman (1989), "Spontaneous Trait Inference", in James Uleman and John Bargh, *Unintended Thought*. New York: Guilford Press, 155-188.

Nisbett, Richard and Timothy Wilson (1977), "Telling More Than We Can Know: Verbal Reports on Mental Processes", *Psychological Review* 84:231-259.

Norton, John (2003), "Causation as Folk Science", *Philosophers' Imprint* 3(4):1-22.

O'Shaughnessy, Brian (1973), "Trying (as the Mental 'Pineal Gland')", *Journal of Philosophy* 70:365-86.

O'Shaughnessy, Brian (2003), "The Epistemology of Physical Action", in Johannes Roessler and Naomi Eilan (eds.), *Agency and Self-Awareness*. Oxford University Press, 345-357.

Pearl, Judea (2000), *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Pettit, Phillip (1995a), "Causality at Higher Levels", in Dan Sperber, David Premack, and Ann James Premack (eds.), *Causal Cognition: A Multidisciplinary Debate*. Oxford University Press, 399-421.

Pettit, Phillip (1995b), "Microphysicalism, Dottism and Reduction" *Analysis* 55(3):141-146.

Pietroski, Paul (2000), *Causing Actions*. Oxford University Press.

Pockett, Susan (2002), "On Subjective Back-Referral and How Long It Takes to Become Conscious of a Stimulus: A Reinterpretation of Libet's Data", *Consciousness and Cognition* 11(2):144-161.

Reichenbach, Hans (1956), *The Direction of Time.* Berkeley, CA: University of California Press.

Robinson, William (2005), "Zooming in on Downward Causation", *Biology and Philosophy* 20:117-136.

Ross, Don, James Ladyman, and John Collier (2007), "Rainforest Realism and the Unity of Science", in James Ladyman, Don Ross, David Spurrett, and John Collier, *Every Thing Must Go: Metaphysics Naturalized*. Oxford University Press.

Ryle, Gilbert (1949), *The Concept of Mind*. University of Chicago Press.

Salmon, Wesley (1984), *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press.

Salmon, Wesley (1998), *Causality and Explanation*. Oxford University Press.

Sato, Atsushi, and Asako Yasuda (2005), "Illusion of Sense of Self-Agency: Discrepancy Between the Predicted and Actual Sensory Consequences of Actions Modulates the Sense of Self-Agency, but not the Sense of Self-Ownership", *Cognition* 94:241-255.

Schaffer, Jonathan (2003), "Is There a Fundamental Level?" *Nous* 37(3):498-517.

Shoemaker, Sydney (2001), "Realization and Mental Causation", in Carl Gillett and Barry Loewer (eds.), *Physicalism and its Discontents*. Cambridge University Press, 74-98.

Simon, Herbert A. (1962), "The Architecture of Complexity", *Proceedings of the American Philosophical Society* 106(6):467-482.

Smith, Michael (1983), "Actions, Attempts, and Internal Events", *Analysis* 43(3):142-146.

Sperry, Roger W. (1969), "A Modified Concept of Consciousness", *Psychological Review* 76:532-536.

Sperry, Roger W. (1975), "Mental Phenomena as Causal Determinants in Brain Functions", *Process Studies* 5(4):247-256.

Sperry, Roger W. (1986), "Macro- versus Micro-Determinism", *Philosophy of Science* 53(1):265-270.

Spirtes, Peter, Clark Glymour, and Richard Scheines (2001), *Causation, Prediction, and Search* MIT Press.

Stalnaker, Robert (1996), "Varieties of Supervenience", *Noûs* 30, Supplement:221-241.

Strayer, David L., Frank A. Drews, and William A. Johnston (2003), "Cell Phone-Induced Failures of Visual Attention During Simulated Driving", *Journal of Experimental Psychology: Applied* 9(1):23-32.

Suppes, Patrick (1970), *A Probabilistic Theory of Causality*. Amsterdam: North-Holland Publishing Company.

Taylor, Richard (1966), *Action and Purpose*. Prentice-Hall.

Thompson, Evan, and Francisco J. Varela (2001), "Radical Embodiment: Neural Dynamics and Consciousness", *Trends in Cognitive Sciences* 5(10):418-425.

Tsakiris, Manos, and Patrick Haggard (2003), "Awareness of Somatic Events Associated with a Voluntary Action", *Experimental Brain Research* 149:439-446.

Uleman, James S., and Gordon B. Moskowitz (1994), "Unintended Effects of Goals on Unintended Inferences", *Journal of Personality and Social Psychology* 66(3):490-501.

van Duijn, Marc and Sacha Bem (2005), "On the Alleged Illusion of Conscious Will", *Philosophical Psychology* 18(6):699-714.

Velmans, Max (2004), "Why Conscious Free Will both is and isn't an Illusion", *Behavioral and Brain Sciences* 27:677.

von Wright, Georg Henrik (1971), *Explanation and Understanding*. Cornell University Press.

Walter, Sven (2006), "Determinates, Determinables, and Causal Relevance", *The Canadian Journal of Philosophy* 37:217-243.

Wegner, Daniel (2002), *The Illusion of Conscious Will*. Cambridge, MA: The MIT Press.

Wegner, Daniel, and Thalia Wheatley (1999), "Apparent Mental Causation: Sources of the Experience of Will", *American Psychologist* 54:480-492.

Wilson, Jessica (2009), "Determination, Realization and Mental Causation", *Philosophical Studies* 145:149-169.

Woodward, James (2003), *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.

Woodward, James (2007), "Interventionist Theories of Causation in Psychological Perspective", in Alison Gopnik and Laura Schulz (eds.), *Causal Learning: Psychology, Philosophy, and Computation*. Oxford University Press, 19-36.

Woodward, James (2008), "Cause and Explanation in Psychiatry: An Interventionist Perspective" in K. Kendler and J. Parnas, (eds.) *Philosophical Issues in Psychiatry: Explanation, Phenomenology and Nosology*. Johns Hopkins University Press.

Wook-Yi, Sang (2003), "Reduction of Thermodynamics: A Few Problems", *Philosophy of Science* 70:1028-1038.

Zhu, Jing (2004), "Understanding Volition", *Philosophical Psychology* 17:247-273.