Intentionality as Methodology

by

Eric Hochstein

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Philosophy

Waterloo, Ontario, Canada, 2011

AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In this dissertation, I examine the role that intentional descriptions play in our scientific study of the mind. Behavioural scientists often use intentional language in their characterization of cognitive systems, making reference to "beliefs", "representations", or "states of information". What is the scientific value gained from employing such intentional terminology?

I begin the dissertation by contrasting intentional descriptions with mechanistic descriptions, as these are the descriptions most commonly used to provide explanations in the behavioural sciences. I then examine the way that intentional descriptions are employed in various scientific contexts. I conclude that while mechanistic descriptions characterize the underlying structure of systems, intentional descriptions allow us to generate predictions of systems while remaining agnostic as to their mechanistic underpinnings.

Having established this, I then argue that intentional descriptions share much in common with statistical models in the way they characterize systems. Given these similarities, I theorize that intentional descriptions are employed within scientific practice as a particular type of *phenomenological model*. Phenomenological models are used to study, characterize, and predict the phenomena produced by mechanistic systems without describing their underlying structure. I demonstrate why such models are integral to our scientific discovery, and understanding, of the mechanisms that make up the brain.

With my account on the table, I then look back at previous accounts of intentional language that philosophers have offered in the past. I highlight insights that each brought to our understanding of intentional language, and point out where each ultimately goes astray.

I conclude the dissertation by examining the ontological implications of my theory. I demonstrate that my account is compatible with versions of both realism, and anti-realism, regarding the existence of intentional states.

Acknowledgements

To start, I would like to thank my supervisor Chris Eliasmith for all his support, guidance, and encouragement. His confidence in my philosophical abilities, and in the quality of work, helped me to produce a better dissertation than I ever thought I was capable of. I am very grateful to have had him as my supervisor, and I am a better philosopher because of him.

I would also like to thank my committee: Tim Kenyon and Paul Thagard. Their help and feedback during the past few years has been invaluable. Both were always welcoming and happy to meet with me whenever I needed, and were extremely encouraging without ever pulling philosophical punches.

Next, I would like to thank the philosophy graduate students at the University of Waterloo. Over the past 5 years, many have become like family, offering a philosophical sounding board whenever I needed to work through ideas, and literally taking me in on occasion when I needed a place to stay. Their friendship has made my time at the University of Waterloo a joy.

In addition, there are a few people in particular that I would like to thank for being a huge influence on me philosophically, and whose comments and suggestions have helped to shape my philosophical views for the better. Of particular note are: Kurt Holukoff, Micheal McEwan, Alexander Winthers, Nora Boyd, and Nicolas Fillion.

Finally, I would like to thank all my family and friends. I would not be where I am today without them, their love, and their support. Their presence in my life is a big part of my successes and my joys. I am a better person because of them.

Table of Contents

Chapter 1
Introduction

The pioneering neuroscientist Santiago Ramón y Cajal famously claimed that "to know the brain […] is equivalent to ascertaining the material course of thought and will, to discovering the intimate history of life in its perpetual duel with external forces" (1937). In other words, with the age of substance dualism largely behind us, the quest for understanding the human mind has become a quest for understanding the human *brain*. We know that the brain is a massively complex system composed of neurological and physiological mechanisms, and so the question becomes: what is our best scientific method for learning about such a system?

Given the mechanistic nature of the brain, it seems reasonable to conclude that the best method for learning about the brain, and thus the mind, is to make sense of the physical mechanisms that constitute it. This mechanistic project of analyzing neurological phenomena, however, seems to be in sharp contrast with our more traditional means of making sense of the human mind: in terms of contentful mental states such as *beliefs, desires, wants, fears, hopes, dreams* and *intentions*. Unlike these contentful mental states, physical mechanisms appear to be devoid of content or purpose. They are merely physical objects obeying physical laws. As a result, this new mechanistic interpretation of the mind appears, at least *prima facie,* to be in conflict with the more traditional accounts. This conflict between the old and new way of understanding mental phenomena has led a number of philosophers to suggest that the traditional mental states account of the mind is simply rooted in a pre-theoretic *folk* understanding of the mind. This folk account is to be contrasted with the more *scientific* understanding of it in terms of physical mechanisms.

This distinction has given birth to a long running debate in the philosophy of mind regarding what role (if any) folk psychology currently has in our new scientific understanding of the mind. How does our previous account of contentful mental states fit into the more recent mechanistic story of the mind that has

emerged in the past century? Implicit in much of contemporary philosophy of mind is the assumption that, in order for the traditional mental state account to remain important to science (as opposed to being merely a heuristic for day-to-day purposes), we must either find a way to reduce contentful mental states to physical mechanisms, or provide a story about how such mental states emerge from physical mechanisms. In an important sense, I propose that neither of these options are correct.

The distinction between contentful mental states, and physiological mechanisms, is not necessarily a metaphysical one that requires an emergent or reductive story to connect mental states to physical mechanisms, but is instead an interpretative one. In other words, we can choose to interpret people *in terms of* physical mechanisms, or *in terms of* contentful mental states. This gets us to the crux of the issue: if we are systems made up of physical mechanisms, then what scientific value is there in adopting a mental state interpretation, as opposed to a mechanistic one?

If interpretations of systems in terms of contentful mental states are truly rooted in *folk* psychology, then this implies that whatever value this interpretation has is primarily limited to colloquial contexts. I will demonstrate that this is incorrect. I propose that our interpretation of systems in terms of contentful mental states acts as a particular sort of scientific tool that is commonly employed in the study of complex systems, and is essential if we want to learn about the physical mechanisms that make up the human brain.

In order to show this, it is first important to provide a better understanding of what differentiates the two kinds of interpretations (mental state/mechanistic). As I will show, it is *not* that one is part of "folk" psychology, while the other is part "scientific" psychology. Instead, it is that one describes systems in terms of states with intentionality, while the other describes systems in terms of causally interacting physical mechanisms. While mechanistic descriptions characterize systems in terms of the structured organization and interaction of their constitutive parts, intentional descriptions characterize systems in terms of having contentful states that are *about* other things (more on this below).

In this dissertation, I will argue that both types of descriptions (intentional and mechanistic) have essential pragmatic benefits in our scientific study of the mind. I propose that intentional descriptions ultimately behave within scientific practice as a kind of scientific model. Specifically, they function as a type of *phenomenological* model.[1] Phenomenological models primarily characterize or describe some phenomenon without attempting to decompose it into parts and operations for better understanding. Intentional models, in virtue of being phenomenological models, have profound *methodological* benefits to our scientific study of the mind. They allow us to generate predictions of systems when we do not understand their underlying mechanisms; they can be used to see patterns in behaviour that other sorts of scientific models (including mechanistic models) miss; and they allow us to see similarities in behaviour that exist across various mechanistic systems. All of these play an essential role in the discovery and understanding of the physical mechanisms that make up the brain. As a result, I propose that intentionality is best thought of as a feature of certain scientific models that are crucial in our discovery, and understanding, of unknown neurological mechanisms.

It is worth clarifying a few points regarding my use of the term "model" here. The term "model" is used in many different ways in science, but the sort that I will focus on in this dissertation is the sort that Carl Craver calls a "representational model". These are models that

> …scientists construct as more or less abstract descriptions of a real system. […] The skeletal account [of a representational model] is as follows. Take some feature (T) of a target system. T might be a static property of the system or it might be characterized as a mapping from inputs (or sets of inputs) onto outputs (or sets of outputs) implemented by a system. […] Modeling T involves constructing an algorithm or function (S) that generates a mapping from inputs onto outputs that is reasonably similar to T. The algorithms or procedures might be implemented in physical systems, written in computer programs, captured in mathematical equations, or sketched

[1] It should be noted that "phenomenological model" in this context has no direct connection to the philosophical domain of Phenomenology. Instead, I use the term as it is commonly employed in scientific domains like neuroscience and physics.

> in block and arrow diagrams. All that matters is that (i) for each input (or set of inputs) in T there is a corresponding input (or set of inputs) in S, (ii) for each output (or set of outputs) in T, there is a corresponding output (or set of outputs) in S, and (iii) for each input–output relation in T there is a corresponding input–output relation in S. (Craver 2006, pp. 356-357)

Given this definition, a *mechanistic* model is one that characterizes the input-output relation of T in terms of the physical interacting mechanisms that transform the input into the output (for more details, see Section 2.2). Meanwhile, a *phenomenological* model is one that attempts to characterize and/or predict the input-output relation in ways that remain completely agnostic as to the structural features of the system that bring about the output given the input (for more details, see Section 6.1.3.).

It should also be noted that many current philosophical debates regarding the metaphysical nature of scientific models will be largely ignored in this dissertation. For instance, the question of whether scientific models are direct descriptions of the world, or are abstract objects that mediate between our scientific theories and the world, is a question I will not address. While a great deal of philosophical debate currently surrounds these issues, they are tangential to my overall project. As a result, I will not delve into them.

To fully flesh out my project, it is best to begin in this chapter with an explanation of why the folk/scientific distinction as it is understood above does not appropriately capture the distinction between mental state descriptions, and mechanistic descriptions.[2] Instead, the relevant distinction is between descriptions of systems in terms of states with *intentionality*, and descriptions of systems in terms of physical mechanisms. Expanding on this idea, I then provide a clear set of criteria for distinguishing intentional descriptions of systems from other sorts of descriptions.

With this criteria for intentional language set, and with my proposed goal in mind, I will then briefly situate my position amongst others on the philosophical playing field. Once this general

[2] This is not to say that there is no useful distinction between folk and scientific psychology however (see below).

foundation is in place, I will lay out a road map for the rest of the dissertation, which will demonstrate the value and importance of intentional language to science.

1.1 Where the Folk and the Psychology Mix

As philosophers, we are cautioned early on about the dangers of trusting our uninformed intuitions. The way things seem to be can often be misleading, and without some rigorous means of keeping our intuitions in line, they will often lead us astray. Consider our pre-theoretical intuitions about physics. Our everyday intuitive understanding of physics tells us that heavier objects fall faster than lighter ones, and that a dropped object always falls straight down. Yet these accounts, despite seeming so intuitive, have proven false. This has led many to draw a distinction between folk notions of physics, and scientific notions of physics. The counter-intuitive scientific findings of quantum mechanics can be sharply contrast with our incorrect, yet highly intuitive, folk notions that, for example, everything that goes up must come down.

In a similar vein, there are many philosophers that draw a sharp distinction between folk and scientific accounts of psychology. Before there existed anything like a rigorous neuroscience or cognitive science, we colloquially explained and predicted each other's behaviour in terms of things like beliefs, desires, intentions, hopes, and fears. We understood that one runs from an assaulter because one *fears* for one's life, has a *desire* to escape, and a *belief* that one can escape by running. Yet, as scientific fields dedicated to the study of mind emerged, a different type of vocabulary came with it. Scientific disciplines like neuroscience were far more likely to predict and explain human behaviour by appealing to things like chemical interactions, patterns of activation, and physiological mechanisms, than they were in terms beliefs, desires, or fears. The discrepancy between these explanations was accounted for by many in terms of their belonging to different realms. Things like "beliefs", "desires", "intentions", "hopes" and "fears" were all considered to belong to the realm of the *folk*, and not the *scientific*, psychology. As a result, these

mentalistic terms became synonymous with "folk psychology". This is why philosophers like Stephen Stich claim that,

> In our everyday dealings with one another we invoke a variety of commonsense psychological terms including 'believe', 'remember', 'feel', 'think', 'desire', 'prefer', 'imagine', 'fear', and many others. The use of these terms is governed by a loose knit network of largely tacit principles, platitudes, and paradigms which constitute a sort of folk theory. Following recent practice, I will call this network *folk psychology.* (1983, p. 1)

This distinction between "folk" and "scientific" explanations in psychology may seem straightforward, but we must be cautious how we use these classifications. To demonstrate, let us return for a moment to folk physics. Consider that in folk physics, we often use concepts like *time* and *motion* in our explanations and predictions of things. From this, ought we to infer that such terms are only folk physical terms? That they ought to be relegated to the realm of the folk, and have no role whatsoever in scientific physics? That cannot be correct, since rigorous scientific physics does make reference to both time and motion.

What then makes one account a folk account, and the other a scientific one? It cannot be that one only employs folk terminology while the other employs scientific terminology, since the same terms can be found in both accounts. If we carry this over to the realm of psychology, the same lesson seems to apply. We cannot claim that terms like "beliefs" and "desires" are not part of scientific psychology *unless we can show that no scientific study of the mind ever employs them*. Yet, even if neuroscientists and some cognitive scientists do not, some cognitive scientists clearly do, as do behavioural psychologists, clinical psychologists, and developmental psychologists. To merely stipulate that they are not *really* doing scientific psychology because of the terms they employ is akin to telling physicists that they are not *really* doing scientific physics because they make reference to "time" and "motion".

As a result, we can conclude that it is not the *terms themselves* that make an account "folk" as opposed to "scientific".[3] What then is the contrast between folk and scientific theories of psychology? Perhaps the contrast is best understood in terms of conflicting explanations of the same phenomenon? If mechanistic neuroscientific explanations (which do not appear to make reference to classic mentalistic terms) and "folk psychological" explanations (which do) cannot be reduced from one to another, then we seem to have conflicting explanations of human behaviour. Given that neuroscience has proven to be a more successful project, we can relegate mentalistic terms to the realm of the folk.

But this story is not plausible either. We cannot use "folk" synonymously with "incorrect", since many rigorous scientific theories have proven to be false as well. Consider that Newtonian mechanics has been supplanted by quantum mechanics as the more successful project of physics. Do we imply from this that Newtonian physics is folk physics? Or more problematic still: if we suppose that quantum mechanics is supplanted by a better physical theory in the future, do we conclude from this that quantum mechanics was folk physics all along?

Ultimately, the distinction between folk accounts and scientific accounts is not one of terminology and incorrect theories, but of the role that terms play *in scientific practice*. Concepts are "folk" when they are not embedded in scientific theories and practices, and they are scientific when they are. As Bas Van Fraassen points out, "to ask that [explanations] be scientific is only to demand that they rely on scientific theories and experimentation, not old wives' tales" (Van Fraassen 1980, pp. 129-130). If we apply this lesson to psychology, then we must acknowledge that the conclusions reached by behavioural and developmental psychologists *are* based on scientific theories and experimentations (they certainly do not base their findings on old wives tales), despite their common employment of so-called folk psychological concepts like "beliefs", "desires", and "intentions". What is important to note from this

[3] One might point out that "time" and "motion" *mean different things* when invoked in folk contexts as opposed to scientific ones. This may be so, but the lesson should equally apply to psychology. If psychologists *do* make

is that classic mentalistic terms are not folk psychological in and of themselves, but only when embedded in folk (i.e., non-scientific) explanations or contexts.

1.2 From *Folk* Accounts to *Intentional* Accounts

Given the reasons above, I propose not to use the folk/scientific distinction when contrasting mentalistic vocabulary (like "beliefs") with mechanistic vocabulary (like "the spiking of neurons"), since the divisions of one do not necessarily reflect the divisions of the other. That being said, there *is* a distinction between mentalistic and mechanistic vocabulary. Philosophers are correct in calling to our attention the fact that descriptions in terms of "beliefs", "desires", "hopes", and "fears" are importantly different from descriptions in terms of physical objects and their causal interactions. Specifically: beliefs, desires, hopes and fears are all *about* or *directed towards* other things. Put another way, mentalistic descriptions interpret systems as having states with *intentionality*.

The term "intentionality", re-introduced to modern psychology and philosophy by Franz Brentano, is used to refer to this "aboutness" of our mental phenomena. According to Brentano:

> Every mental phenomenon is characterized by what the scholastics of the Middle Ages called the intentional (also mental) inexistence of an object, and what we could call, although not in entirely unambiguous terms, the reference to a content, a direction upon an object (by which we are not to understand a reality in this case), or an immanent objectivity. Each one includes something as object within itself, although not always in the same way. In presentation, something is presented, in judgment something is affirmed or denied, in love loved, in hate hated, in desire desired, etc. (Brentano 1970/1874, p. 119-120)

---

reference to "beliefs" and "desires" (among other mental states), then we might conclude that they mean different things than their folk counterparts, but not that belief-desire talk is itself only folk psychological.

With this in mind, I propose that to describe someone using mentalistic vocabulary like "believes", "desires", "hopes", or "fears" is to describe them using an *intentional language*.[4] But what exactly distinguishes intentional language from other sorts of language? I propose the following list to help clarify this distinction:

1) Intentional language characterizes systems in terms of states with aboutness.

The most obvious and straightforward criterion is that they describe systems in terms of having internal states that are about other things. Keep in mind that for our present purposes, this condition does not presuppose that a system "really" has content or not. It merely claims that describing a system using intentional language involves talking about the system *as though* it has internal states that are directed towards other things.

2) Intentional language can often be structured as propositional attitudes.

One of the defining features of intentional language, as originally noted by Bertrand Russell, is that intentional descriptions can often be structured as though they pick out a relation between a person and a proposition. For example, my belief that it is raining outside can be understood as a relation between myself and the proposition "it is raining outside". For this reasons, intentional descriptions are considered by many to be *propositional attitudes* (i.e., attitudes we have about propositions)

There are a few points worth noting here. First, this feature of intentional language does not commit one to any metaphysical story about the nature of propositions. One need not be a realist about

[4] By "intentional language", I do not mean an alternative to *natural* language. Following Sellars (1956), I propose that science is often conducted *within* a given natural language. By intentional language, I mean only a particular sort of linguistic device (one which employs sentences with a particular structure, or terminology of a certain sort).

propositions to accept this point. In other words, I am not claiming that intentional states *really are* relations between individuals and propositions (whatever they may be). I am claiming only that attributing an intentional state to a person can often be conveyed in the form of a sentence that structurally resembles a relation between a person and a proposition.

Second, this feature of intentional language similarly does not commit one to the idea that intentional content must always be sentential. The fact that we can use a proposition to characterize the content of one's mental state does not mean one is *ipso facto* committed to some variation of the Language of Thought Hypothesis. We may have plenty of reasons to think that mental content is *not* sententially structured. The present point is simply that we can often *use* a proposition to characterize this content for our purposes.

3) Intentional content can *misrepresent*, or represent things that *do not exist*.

A third defining feature of intentional language is that any account of a system in terms of intentional states must allow for the possibility of error. It is always possible for an attributed intentional state to *mis*represent the world. For instance, I might believe it is raining outside even when it is not. In this sense, intentional states can be false. A related point is that it is also possible for an attributed intentional state to be about something that does not genuinely exist. I might, for instance, have a mental representation of Santa Claus even though no such person exists in the world.

4) Intentional language is often considered to be normative.

Philosophers such as Willard van Orman Quine (1960) and Daniel Dennett (1987) note that another defining characteristic of intentional language is that it is rooted in assumptions of rationality. In other

words, describing a system in terms of intentional states can only be informative *if we assume that the system is rational*. Making sense of a system's behaviour in terms of its intentional states implicitly involves making a claim regarding how the system *ought* to behave given the *rational interconnections* of its intentional states. If a person acts irrationally, on the other hand, it becomes virtually impossible to make sense of their behaviour in terms of attributed intentional content.

As a simple example, imagine someone who is irrational in virtue of explicitly believing two blatantly contradictory propositions ('p' and '-p'). If we attribute a belief in both propositions to the person, then we are no better off than attributing a belief in neither. In either case we are left wondering what it is that the person *really* believes regarding p. As Stich notes, "when a person is *that* different from us, we are inclined to think that there is just *no saying* what he really believes" (1983, p. 101, emphasis in text). And given this, there is no way to make sense of the system's behaviour in terms of this intentional content. As a result, intentional descriptions are deeply intertwined with assumptions of the rationality of the system.

5) Intentional language can display referential opacity

Many consider a relevant feature of intentional language to be that such descriptions are *referentially opaque*. With many sentences of natural language, you can exchange a proper name with a co-referring name without changing the truth-value of that sentence. For example: if it is true that "The Morning Star is the second closest planet to the sun", then it is also true that "The Evening Star is the second closest planet to the sun." Given that "Morning Star" and "Evening Star" both refer to the planet Venus, we can exchange one for the other *salva veritate*. One exception to this rule is when the proper name is part of a proposition embedded in a propositional attitude. For example, the sentence "Benjamin believes that the

Evening Star is the second closest planet to the sun" may be false even though the sentence "Benjamin believes that the Morning Star is the second closest planet to the sun" may be true.

There are a couple of things to note about this criterion. First, intentional descriptions are not the only types of descriptions in which referential opacity occurs. And second, many instances of intentional attribution do not necessarily lead to a change in truth-value when replacing co-referring names. There are many instances of intentional descriptions where this does occur, however, and this is a feature not found with many other types of descriptions (like mechanistic descriptions).

It is my intention to show that intentional language, understood in this way, behaves within scientific practice as a kind of phenomenological model. One that is an essential component in our study of the mind.

1.3 A Very Broad Look At The Philosophical Playing Field

Before delving into my own account, it is helpful to get a sense of the different philosophical views regarding intentional language that have been offered over the past few decades. While I propose that intentional language plays an integral role in our study of the mind, others have argued that it provides no real scientific value (Churchland 1981; Stitch 1983). According to this type of view, intentional language is rooted in an outdated, and soon to be displaced, theory of the mind.

On the other side of the philosophical spectrum are those who propose that intentional language characterizes a genuine metaphysical property of minds (their genuine aboutness), and as a result, they insist that a proper scientific account of the mind must explain how physical mechanisms can bring about the intentionality of the mental. John Searle (1980, 1992), for instance, proposes that intentionality is an emergent property of certain types of biological systems. Meanwhile, Jerry Fodor (1998) proposes that there are nomological laws which connect certain kinds of physical states to the things that they are about, whenever the appropriate functional conditions are met.

As I progress through this dissertation, I will demonstrate why assuming positions on either side of this spectrum, before investigating how intentional language is used within scientific practice, is ill-advised (see Section 2.4). Moreover, when we examine the way intentional language is genuinely used by scientists, we find that none of these accounts are vindicated by actual scientific practice (See Section 4.1, 4.2, and 4.3).

In contrast to the positions above, the philosopher Daniel Dennett proposes that we should evaluate intentional language based on the pragmatic benefits of its usage, as opposed to whether it correctly identifies some metaphysical phenomenon of aboutness. He suggests that intentional language acts as a kind of predictive framework that we employ for interpreting the behaviour of systems (Dennett 1971; 1987; 1991a; 2007). This framework, which he calls the "intentional stance", employs intentional descriptions as a means of predicting how different kinds of systems will behave. By assuming a system is rational, we can predict the sorts of things it will do based on the rational interconnection of the intentional states that we attribute to it.

Dennett proposes that we adopt one of three different stances whenever we predict the behaviour of systems. The intentional stance is one such stance. The second is what he calls the *physical stance*, which predicts systems by applying known physical laws to the component parts of the system. The third is the *design stance*, which predicts systems by appealing to their designed function.

Recall that the account I will argue for in this dissertation is that intentional language functions as a sort of phenomenological model. In this regard, the picture I paint will appear, at first glance, to be remarkably similar to the account offered by Dennett. Phenomenological models are commonly used to predictively model systems when we are unable to identify their underlying causal mechanisms. In a similar vein, Dennett proposes that we adopt the intentional stance in order to generate predictions of systems, often when we cannot identify their underlying mechanisms. And so it seems that the general point I wish to make has already been made.

Despite appearances, however, my aim in this dissertation is not to re-affirm Dennett's account. While the broad strokes of Dennett's theory (that intentional descriptions are used as a predictive tool) are on the right track, many of the details are not. Much of what Dennett assumes about the intentional stance, and its role in scientific practice, are deeply problematic. Dennett, for instance, proposes that the intentional stance is primarily used as a *heuristic* device in everyday contexts, but that its value to scientific practice is rather limited (1987, p. 350). Given their *normative* component, intentional descriptions presuppose (but provide no scientific explanation for) the rationality and intelligence of the systems being predicted. In this respect, he proposes that intentional descriptions are "vacuous as psychology" (Dennett 1971, p. 99).

I propose that intentional language is far more important to the sciences of the mind than Dennett realizes, and that his framework of "stances" does not reflect actual scientific methodology. The first thing to note is that Dennett`s distinction between predictive "stances" (intentional/design/physical) does not cut along the same lines as the distinction between different scientific models (phenomenological/mechanistic). Our use of mechanistic models, for instance, can fall under more than one of Dennett's predictive stances (it can act as the physical stance, and the design stance, depending on how abstractly we describe the mechanism); and there are many kinds of commonly used scientific models that do not seem to fall under *any* of Dennett's stances (like our use of statistical models). A shift from *stances* to *models* betters fits with actual scientific methodology, and allows us to better understand the way in which we employ different types of linguistic tools in scientific contexts.

Even more problematic is the fact that many of Dennett's claims regarding the benefits and drawbacks of his three stances are not borne out by actual scientific practice. He insists that the physical stance will, at least in principle, *always* work to predict the behaviour of systems. Yet there is no type of description in science that has such a virtue. Similarly, his reasons for denying intentional descriptions an important place in scientific practice would force him to likewise discard many commonly used scientific

models that are invaluable to scientific practice, and which would fall under his conception of the physical stance (see Section 9.5 for a more exhaustive review of Dennett's theory).

My goal in this dissertation is therefore not to re-affirm Dennett's position, but instead to develop a novel account of intentional language that will ultimately vindicate certain parts of Dennett's view, while rejecting others. In a similar fashion, I will demonstrate how the account I propose also vindicates aspects of other major philosophical accounts of intentional language (such as functionalism, eliminative materialism, anomalous monism, and the co-evolutionary research ideology), while highlighting the aspects of those theories that ought to be discarded. Once my own account is on the table, I will provide an in-depth analysis of these philosophical positions in order to demonstrate how my account stacks up against those that have come before (see Chapter 9). First, however, it is important to understand the road I plan to take in order to flesh out my account, and the motivation for it.

1.4 The Shape of Things to Come

What benefits, if any, does intentional language bring to our scientific understanding of the mind? The remainder of this dissertation is dedicated to providing a clear answer to this question. The following breakdown of chapters highlights the path I plan to take to get us to that answer:

If we are interested in the role that intentional language plays in our scientific explanations, it is important to begin with an analysis of scientific explanation more broadly. And so in Chapter 2, I discuss the nature of scientific explanation, and an important pragmatic component to it. I propose that different sorts of descriptions can count as scientific explanations depending on the context one is working in, and the particular sort of question one is interested in answering. With this in mind, I examine what it means to provide a scientific explanation in the sciences of the mind specifically (as opposed to scientific domains such as physics). I suggest that the type of description (or model) typically considered explanatory within the sciences of the mind are mechanistic descriptions. From there, I look at the question of whether an intentional account of mental states is in conflict with a mechanistic understanding

of the brain. Ultimately, I conclude that the answer depends on the ontological story that one adopts regarding the nature of intentionality. If, on the other hand, we are concerned with the value of intentional *language* to scientific practice, then we must try to leave our ontological commitments at the door, and look instead at the way in which such language is actually being used in scientific contexts. While many philosophers have traditionally proposed that intentional language is used to characterize some unique ontological phenomenon of the mind (its genuine aboutness), I propose that actual scientific practice may not reflect this.

Building on this idea, I demonstrate in Chapter 3 that we find intentional language being used by scientists in domains not dedicated to the study of the mind, lending credence to the idea that its use in science is not necessarily to identify some unique ontological property of minds. As a particular example, I look at the way in which biologists attribute *information* to genes. I demonstrate that this use of the term "information" cannot be understood in the more technical senses associated with communication theory, and instead is best understood as an intentional term.

With this established in Chapter 3, I will have examples of intentional language being used both within the sciences of the mind (such as in psychology) and without (such as in biology). In Chapter 4, I look at what, if any, scientific benefits connect these different uses of the language together. I look at various possibilities and conclude that intentional language does have a scientific benefit, and that this benefit is one of prediction. Moreover, that this predictive value is not contingent on whether intentional descriptions denote (or abstractly describe) sub-personal physical mechanisms working within systems. In other words, intentional descriptions do not predict in virtue of abstractly describing the functioning of certain mechanisms working within the system that generate its behaviour. While we do, on occasion, find that attributed intentional states correspond to particular mechanisms working within the system, it is not in virtue of this correspondence that intentional models are predictive.

In Chapter 5, I provide evidence for this claim with an example from psychology. I argue that many intentional concepts used in traditional psychology, such as "beliefs" and "intentions", do not correspond to, or describe the functioning of, any particular type of physical mechanism working in the brain. However, there is empirical evidence that psychological models which employ such concepts *are* genuinely predictive. As an example, I look at the Theory of Planned Behaviour, and its documented predictive successes in psychology. This, I propose, demonstrates that the predictive value of intentional language is not based on its characterization of the structure, or mechanisms, of the system being predicted.

The fact that intentional language predicts systems while remaining agnostic as to their mechanistic underpinnings is a feature that other sorts of scientific models share as well. Specifically, this is a defining characteristic of phenomenological models. Such models are defined by their ability to usefully characterize or predict systems while ignoring the system's underlying structure or mechanisms. A prime example of this type of model in science is a statistical model. In Chapter 6, I argue that intentional models are best thought of as being a type of phenomenological model, and I demonstrate this by highlighting the numerous similarities that exist between our use of intentional models and our use of statistical models in science. While the two models have different ideal contexts for their application, they are a similar sort of scientific tool; one that is an essential part of our study of the mind.

In Chapter 7, I look at possible differences between intentional models and statistical models that might threaten the inclusion of intentional models into the class of phenomenological models. I argue that the differences that exist between the two types of models are not sufficient to deny that intentional models work fruitfully within science as a type of phenomenological model.

Having established what sort of scientific tool intentional descriptions are, I then turn in Chapter 8 to broader issues regarding the role that intentional models (understood *as* phenomenological models) have in our study of the mind more generally. I argue that intentional models have a profound

methodological role in our study of the mind. They allow us to generate essential predictions in contexts where other sorts of predictive tools are not available, or are uninformative. They similarly allow us to see similarities that exist across different mechanistic systems. This information puts crucial constraints on what the mechanisms of the brain must be like, and in what contexts they function. In this respect, intentional models play a huge pragmatic role in our ability to learn about the mechanisms of the brain, and thus in generating a mechanistic explanation of the mind.

Similarly, this pragmatic benefit is in no way contingent on the reducibility of the objects in an intentional model (like "beliefs" or "mental representations") to the objects in a mechanistic one. The question of reduction is irrelevant to the question of whether these models have methodological value to science. What matters are the pragmatic benefits that employing these models bring to our scientific practices, and the way they inform our understanding of the brain (and thus the mind). As a result, very little about the importance of such models in science hangs on the truth or falsity of reductionism. In a similar vein, I propose that the often debated question of whether the objects and categories postulated by psychology can reduce the objects and categories of neuroscience is irrelevant to the question of whether psychology characterizes important features of human interaction and behaviour which informs and constrains our study of neurological mechanisms.

In addition, it may also be the case that the complexity of the phenomena under investigation (the vastly complex system that is the human brain) may be such that no one particular type of scientific model may be sufficient in isolation to represent all features of it required for all our scientific purposes. In which case, intentional models may be more than just a methodological tool for generating mechanistic explanations; they may also be essential for representing certain aspects of mechanistic systems (e.g., behavioural regularities) that other sorts of models are unable to capture or identify. Different scientific models distort, emphasize, or abstract out, different features of systems being represented, and so the way

intentional models represent systems may be informative in a way that other sorts of models (like mechanistic models) are not.

With my story of the role that intentional language plays in science finally on the table, I take a historical look back in Chapter 9 at the different views regarding intentional language that have been offered by philosophers in the past. I demonstrate that many of these accounts helped to identify important features of intentional language, even if the accounts themselves ultimately went astray.

Lastly, in Chapter 10, I conclude by briefly examining what the *ontological* implications of my account are. I demonstrate that my account does not necessarily commit me to any particular ontological or metaphysical story regarding intentional states. To emphasize this, I highlight a number of different *realist* and *anti-realist* positions that are compatible with the account I provide. Regardless of whether one chooses to include intentional states in one's ontology, the methodological value of intentional language is in no way threatened. The value of such language to science is substantial, regardless of whether we choose to deem the objects they postulate "real" or not.

Chapter 2
Explanations, Intentionality, and Ontology

In order to understand the place of intentional language in our scientific study of the mind, it is prudent to begin with some general claims regarding scientific explanation, and the sciences of the mind, more broadly. Assuming that we reject substance dualism, how is it that the sciences of the mind scientifically explain mental phenomenon? And how does intentional language relate to such explanations?

In this chapter, I will attempt to address these questions in order to set the stage for the project ahead. Addressing these questions will first require saying something briefly about the nature of scientific explanation, and its important pragmatic features. Once this is done, I will be able to turn to the sciences of the mind in particular, and determine what counts as an appropriate explanation in *those* domains specifically. Lastly, I will turn to the question of intentionality. Do intentional descriptions fit into a mechanistic framework? Why has intentionality been considered a problem for such mechanistic accounts in the past?

2.1 The Nature of Explanation

What exactly constitutes a scientific explanation? In the mid-20th century, Carl Hempel and Paul Oppenheim (1948) proposed that we can provide scientific explanations through the use of deductive arguments involving claims about known universal laws. This interpretation of explanation, called the deductive-nomological (DN) model (also sometimes called "covering-law" explanations), held a great deal of sway during the mid and late 20th-century.

While initially intuitive, the DN model has come increasingly under fire in recent years. Due to asymmetries in explanation, there are many cases which satisfy the DN model yet are nevertheless considered unexplanatory. To borrow an example from Bas Van Fraassen (1980, p. 104), suppose that a barometer falls exactly when there is a storm coming (and suppose we can identify a law-like relation that

guarantees this). Suppose we then construct a deductive argument at time *t* which concludes, based on the barometer, that a storm is coming. Even though this satisfies the DN requirement for an explanation, we do not think that the oncoming storm is *explained by* the barometer falling. On the contrary, the barometer's falling is explained by the oncoming storm. Such examples are numerous in science (Van Fraassen 1980; Salmon 1989, p. 47).

To make matters worse, there are also clear cases in which we provide scientific explanations even though we cannot identify universal physical laws (Woodward 2000, Bechtel 2008). In the life sciences, for example, we often provide explanations despite being unable to identify laws:

> One of the most jarring results of joining a naturalistic perspective to a focus on the life sciences is that in many parts of biology one seems to look in vain for what philosophy has commonly taken to be the principle explanatory tool of science, that is, laws. The few statements that have been called laws in biology, such as Mendel's laws, have often turned out to be incorrect or at best only approximately correct. [...] But that does not mean that biologists and psychologists are not developing explanations. If one investigates what biologists and psychologists seek and treat as sufficient for explanation, it often turns out to be mechanisms, not laws. (Bechtel 2008, p. 10)

In the life sciences, we simply do not find anything that resembles universal laws, and so we cannot appeal to such laws to generate explanations. Instead, we explain by appealing to physical mechanisms. To better understand why, consider Craver's claim that an explanation of a system's behaviour does not simply tell us (among other things) how a system in fact behaves, but also how it *would* behave in various counter-factual situations (2006, p. 358). When working with phenomena in physics, if we assume that laws of physics remain invariant, then we can predict how a system would behave in counter-factual situations by applying the known universal laws to the new conditions. When we switch to the life sciences, however, we do not find universal laws that we can appeal to in the same

way.[5] Thus to explain the counter-factual behaviour of systems in these domains requires understanding how the physical parts that make up the system interact with one another so as to produce its behaviour. This mechanistic understanding of the system allows us to *intervene* in its workings so as to determine how the system would behave if conditions were different (Craver 2006; Eliasmith 2010). In this respect, covering-law explanations allow us to identify the counter-factual behaviour of systems in scientific domains like physics, but it is mechanistic explanations that allow us to identify counter-factual behaviour of systems in scientific domains like chemistry or neuroscience. What this means is that the sorts of descriptions that can count as an explanation often depend on the domain of inquiry in which one is working.

The complications with scientific explanation do not end there however. While appeals to covering-laws can often be explanatory in physics, and appeals to mechanisms can often be explanatory in neuroscience, we should not conclude that an entire domain of inquiry uses only a single type of description for all explanatory purposes either. Consider the use of mechanistic explanations in neuroscience. Machamer et al. argue that mechanistic models are most commonly used as explanations in neuroscience, yet they are also careful to insist: "we do not claim that all scientists look for mechanisms or that all explanations are descriptions of mechanisms" (2000, p. 2). As Anthony Chemero and Michael Silberstein (2008) note, mechanistic descriptions and dynamical descriptions (i.e., descriptions of systems by way of Dynamic Systems Theory) can both provide explanations of mental phenomena *depending on the particular question we are interested in answering about the system*. As they put it:

[5] We similarly cannot (at least at present) use the universal laws of physics to explain phenomena in the life sciences. The sciences of the mind, for instance, are interested in explaining phenomena that exists in humans, but not rocks. The universal laws of physics, meanwhile, do not distinguish between humans and rocks. And so such descriptions are at the wrong level of abstraction to identify the differences we care about. Meanwhile, when we switch to neuroscience or psychology, we are better able to characterize the differences we care about, but we find no universal laws that we can use to typify them.

> On our view, dynamical and mechanistic explanation of the same complex system get at different but related features of said system described at different levels of abstraction and with different questions in mind. (Chemero & Silberstein 2008, p. 17)

And so what counts as an explanation can often depend on what sorts of questions one is interested in answering. This is why Van Fraassen suggests that an explanation is best thought of not simply as a particular type of description (be it a description in terms of covering-laws or mechanisms), but as an answer to a question about why something is the case:

> An explanation is not the same as a proposition, or an argument, or a list of propositions; it is an *answer* (Analogously, a son is not the same as a man, even if all sons are men, and every man is a son). An explanation is an answer to a why-question. So, a theory of explanation must a theory of why-questions. (1980, p. 134)

This being the case, the sort of descriptions that will count as an explanation will depend on the sorts of why-questions we are asking, and the sorts of things we are looking for in an answer. And we have little reason to think that all why-questions will be answered satisfactorily by appeals to laws, or to mechanisms. Put another way, a suitable answer to a why-question will depend on the class of alternatives (or the *contrast class*) one could provide as an answer in the appropriate context. To demonstrate, consider the following question:

> Why did Ben walk to the amusement park?

A satisfying answer to this question (and thus a satisfying explanation for why Ben walked to the amusement park) will depend on the context in which we ask the question. For example, we can construe the above question in the following ways:

(1) Why did *Ben* walk to the amusement park?

(2) Why did Ben *walk* to the amusement park?

(3) Why did Ben walk to the *amusement park*?

The class of possible answers that will be appropriate for (1) will not be the same as the ones that will be appropriate for (2) or (3). Question (1) is concerned with why Ben, *as opposed to someone else*, walked to the amusement park (thus the contrast class is the set of possible people that could have walked to the park in that situation). Question (2), on the other hand, is concerned with why Ben walked to the park, *as opposed to taking some other means of transportation* (thus the contrast class becomes the set of possible transportation options available to Ben). Finally, question (3) is concerned with why Ben walked to the amusement park, *as opposed to some other location* (making the contrast class the set of possible locations available to Ben). "The difference between these various requests is that they point to different contrasting alternatives. [...] In general, the contrast-class is not explicitly described because, *in context*, it is clear to all discussants what the intended alternatives are" (Van Fraassen 1980, pp. 127-128).

It should be noted that this pragmatic component to explanation is not merely a feature of colloquial explanations, but of explanations in science as well:

> It might be thought that when we request a *scientific* explanation, the relevance of possible hypotheses, and also the contrast class, are automatically determined. But this is not so, for both the physician and the motor mechanic are asked for a scientific explanation. The physician explains the fatality *qua* death of a human organism, and the mechanic explains it *qua* automobile crash fatality. To ask that their explanations be scientific is only to demand that they rely on scientific theories and experimentation, not old wives' tales. Since any explanation of an individual event must be an explanation of that event *qua* instance of a certain kind of event, nothing more can be asked. (Van Fraassen 1980, pp. 129-130)

Given this, appeals to physical laws will be explanatory *if we are working in the appropriate context and responding to the appropriate questions*. Meanwhile, appeals to mechanisms will be explanatory in different contexts and/or in response to different sorts of questions. Similarly, some contexts will involve scientific explanations that appeal neither to laws nor mechanisms. We might, for example, explain why a particular group of people have a particular contagious disease due to their geographic proximity to one another. Such a response would easily count as a scientific explanation despite making appeals neither to laws, nor to mechanisms.[6]

Of course, it is not surprising that appeals to laws are more often than not considered explanatory in physics, and similarly with mechanisms in biology or neuroscience. Given that a particular domain of scientific inquiry is often interested in finding answers to very specific types of why-questions, they will often be looking for answers of a particular sort (laws in physics, mechanisms in biology, etc). This is not universally true, but it is common (not every why-question in neuroscience is necessarily going to be a question that appeals to mechanisms, but many are given the focus of inquiry).

A detailed philosophical investigation of the nature of explanation is not the goal of this dissertation, and so I will not dwell on this topic. The point is merely to emphasize the complexities and pragmatic nature of explanations. Considering a particular type of description to be explanatory or

[6] One might argue that such a response *indirectly* makes appeals to laws and mechanisms, since it is by way of laws and/or mechanisms that geographical proximity allows the disease to spread from person to person. But this sort of response must be resisted, otherwise we risk trivializing mechanistic explanations as a whole. Consider Bechtel's claim that "if one investigates what biologists and psychologists seek and treat as sufficient for explanation, it often turns out to be mechanisms, not laws" (2008, p.10). However, given that it is by way of universal subatomic physical laws that the parts of a mechanism are able to interact the way they do, then we should equally be able to claim that all descriptions of mechanisms *indirectly* make appeals to laws for their explanations (in exactly the same way our geographic proximity account above indirectly appeals to laws and/or mechanisms for *its* explanations). Yet, how then are we to make sense of Bechtel's claim that biologists and psychologists explain by way of mechanisms as opposed to laws? It seems that mechanisms are explanatory because they answer the sort of questions we are interested in, irrespective of their relation to subatomic physical laws. Similarly, I propose that the geographic answer may answer the sort question we are interested in, irrespective of their relation to physical laws or mechanisms.

unexplanatory *simpliciter* is inappropriate. Instead, the question is whether a given description can be explanatory given the appropriate question and context.

2.2 Mechanistic Explanations and the Sciences of the Mind

As mentioned above, the type of model typically (if perhaps not universally) considered ideal for explanations in the life sciences are mechanistic models. These sorts of explanations allow us to intervene in systems so as to determine counter-factual behaviour, and they similarly provide us with explanations that are compatible with physicalism (i.e., they do not require positing any additional spooky substances to explain phenomena). With this in mind, the sciences of the mind similarly appeal to mechanistic models for explanations (Dennett 1994; Machamer et al., 2000; Bechtel 2005, 2008; Glennan 2005; Craver 2006; Craver & Bechtel 2006; Bechtel & Abrahamsen 2007; Thagard 2009; Eliasmith 2010; Zednik 2011). But what exactly is a mechanistic model?[7]

A mechanistic model explains some phenomenon by "identifying component parts and operations within a system and showing how they are organized to realize the phenomenon of interest" (Bechtel & Abrahamsen 2007). Typically, a mechanistic account of this sort has four major components: The phenomenon, the parts, the activities, and the organization.

The *phenomenon* of a mechanism can be thought of as the thing (regularity, process, capacity, state) that needs explaining. After all, "mechanisms are always mechanisms *of* a given phenomenon. They are the mechanisms *of* the things that they *do*" (Craver 2006, p. 368). A proper account of the phenomenon must include relevant information about when and how it appears. Thus, it must include

[7] A note of clarification: Talk of mechanisms is ubiquitous in science, however its use is not always consistent. We must therefore be clear on what we mean by a "mechanism". For example, in political science we might talk of the *mechanism* responsible for social change. Similarly, an economist might talk of the *mechanism* responsible for the rise in monetary inflation. In such cases, a description of a mechanism need not explicitly describe particular physical entities, and the specific ways they spatiotemporally interact. While this is a legitimate scientific use of the term "mechanism", it is going to be too vague for our purposes. We are interested in how particular physical objects (such as neurons) are interacting with one another to produce some phenomenon (like planning for one's future).

things like the precipitating conditions, inhibiting conditions, modulating conditions, non-standard conditions, and by-products of the phenomenon (Craver 2006, p. 368).

The *parts* of a mechanism are the relevant components within the system that interact with one another in order to produce the phenomenon. A complete mechanistic account must accurately determine what the appropriate parts of the system are. These parts must be real, and not merely fictional posits. This distinction between real objects and fictional posits can often be murky and difficult to define however. As Craver tells us, "there is no clear evidential threshold for saying when one is describing real components as opposed to fictional posits" (2006, p. 370). Craver proposes that, as a rough guideline, we can consider a part real when it exhibits a stable cluster of properties, can be detected using multiple independent devices, can intervene into other components and activities, and is plausible in the relevant circumstances (Craver 2006, pp. 370-371). It is worth noting here that Craver's conditions on being "real" need not be interpreted as a general set of criteria for one's ontology. While some philosophers may wish to go this route, it is certainly not something we must be committed to (see Section 10.2). Craver's criteria should be interpreted primarily as a guideline for what sorts of objects can function as the causally interacting physical parts of biological mechanisms. It need not be committed to anything else beyond that.

The *activities* of a mechanism (also known as the *causal aspect* of a mechanism) are the interactions and processes that go on between the component parts in order to produce the phenomenon. Consider the simple functioning of a mousetrap:

> Pressing the trigger [of the trap] *releases* the catch, *allowing* the spring to *launch* the impact bar. The verbs in this description of the mousetrap refer to the relevant causal relations among the component parts. (Craver & Bechtel 2006, p. 470)

Lastly, the *organizational* aspect of a mechanism is the relevant way in which the parts are situated spatially and temporally within the mechanism. This includes the "relative locations, shapes, sizes, orientations, connections, and boundaries of the mechanism's components" (Craver & Bechtel 2006, p. 470). The way in which the parts are structured and placed within the mechanism determines how the parts can interact with one another in order to produce the relevant phenomenon.

Given the above aspects of a mechanism, it is not uncommon for mechanisms to be made up of sub-mechanisms. In such cases, the parts making up the system satisfy the criteria for being mechanisms themselves, creating a hierarchy of levels:

> Many of the components of a mechanism are themselves mechanisms –they perform operations in virtue of their parts (now subparts of the original mechanism) performing operations of their own. This mereological relation gives rise to a clear sense of levels –parts are at a lower level than the mechanism they comprise. (Bechtel 2005, p. 315)

This type of multi-leveled mechanistic description is the sort of thing we are looking for if we want to explain how systems can realize mental activity within a physicalist framework. The proper understanding of these physical mechanisms will also explain why some things are endowed with mentality (they have the appropriate physical mechanisms), while others are not.

As the sciences of the mind progress, they get better at discovering the physical mechanisms responsible for different mental phenomena. By studying subjects with brain damage, for example, we can learn which physical parts and organizations of the brain are the ones responsible for which mental phenomena by seeing which mental capacities the person lacks due to the damage. Consider vision:

> The first clues as to which brain parts perform visual operations came from analyzing patients with visual deficits stemming from brain damage. Bartolomeo Panizza, who studied patients

> experiencing blindness after stroke-induced occipital lobe damage, proposed that the occipital lobe was the cortical center for vision. (Bechtel 2008, p. 91)

Similar such experiments have helped us better understand which physical mechanisms are at work in producing other mental phenomena like memory and language use. In this way, instead of building up to explaining mental phenomena by first understanding everything there is to know about physics, and then "working our way up" to neuroscience, we work by reverse engineering. We look at the sorts of things that have the mental phenomenon or ability in question, and what physical mechanisms are at work when such a phenomenon is present. We then work backwards to understand what these mechanisms are and how they work.

2.3 The Problem of Intentionality

Given that the sciences of the mind are mechanistic, it seems reasonable to conclude that a correct description of the brain in terms of physical mechanisms will provide us with a complete understanding of the mind. However, descriptions of mechanisms seem to leave no room for the phenomenon of intentionality. Physical parts causally interacting in space are not *about* anything; they merely are. John Searle has something similar in mind when he tells us that "Darwinian mechanisms and even biological functions themselves are entirely devoid of purpose or teleology" (Searle 1992, p. 52). In this regard, mechanistic descriptions seem, at least *prima facie*, insufficient to provide us with a physicalist explanation for *how* the phenomenon of intentionality is produced by the brain.

However, whether intentionality is a genuine problem for mechanistic explanations or not depends on our ontological story about intentionality. Is intentionality an ontological property inherent in certain physical states of the brain? Or is intentionality just something we linguistically ascribe to systems? Or is it something else entirely? Different ontological stories will be compatible with, or in conflict with, our mechanistic explanations. If intentionality is a robust ontological property of the mind

that mechanistic explanations are unable to ever account for (as Brentano himself thought), then mechanistic explanations seem insufficient to explain the mind. On the other hand, if intentionality is merely part of an outdated or unscientific way of interpreting the behaviour of people, then there is no tension between intentional accounts and mechanistic ones because there is no intentionality. And so our account of the value of intentional language to science seems to vary greatly depending on our ontological story of what intentionality actually *is*.

But which ontological story do we accept? We must be cautious in how we proceed here. Any ontological account of intentionality that we adopt from the outset has the potential to taint our perception of what the benefits of intentional language are to science. Thus the questions we should start with are not *ontological* in nature. Instead, they are:

- *Do we* use intentional language in science?
- If so, then in what contexts?
- What are the scientific benefits (if any) gained from *talking this way* about systems?

Once we determine if there are benefits in employing this linguistic tool, and what they are, then can we begin drawing ontological conclusions. To do otherwise is to invite confusion and generate problems. In the section that follows, I will demonstrate just how adopting an ontological stand on intentionality prior to understanding the role of intentional language in scientific discourse can be a roadblock to our scientific understanding of the mind.

2.4 The Problems with a Premature Ontology

Let us consider two ontological stories that have been adopted by various philosophers in the past. One is a specific type of ontological realism regarding intentionality, while the other is a type of anti-realism. I

will demonstrate how adopting either can negatively influence our assumptions about the way in which we use intentional language in science. Let us begin with the realist position.

2.4.1 Intentional realism

Suppose we try to explain the use of intentional language in science by appealing to a robust ontology of intentionality. Intentionality, we might claim, is some unique and mysterious (at least for the moment) metaphysical property that is attached to certain physical states, like brain states, but not others, like disorganized piles of rocks on a beach. Given such an ontological picture, intentional descriptions would identify very real intentional properties that certain physical states have. While intentional descriptions characterize certain physical states *in terms of* their genuine ontological property of intentionality (by identifying mental states such as beliefs, desires, and mental representations), mechanistic descriptions can only describe physical states in terms of the spatiotemporal interactions of their constitutive parts and not in terms of their intentionality. While we can know facts about neurological mechanisms through scientific investigations, we know we have intentional states like beliefs and desires because we directly experience them. Or as John Searle puts it, "it seems crazy to say that I never felt thirst or desire, that I never had a pain, or that I never actually had a belief, or that my beliefs and desires don't play any role in my behavior" (Searle 1992, p. 48).

If we adopt this seemingly intuitive ontological picture, then the proper scientific uses of intentional language would appear to be that they pick out all, and only, those physical states that possess this genuine ontological property of intentionality. Even if it were useful to talk of piles of rocks *as though* they had mental states with "aboutness", this would not be an appropriate *literal* use of the language. The language is used to capture important features of our mental states; features we have direct experience of.

The problem with this story is that it assumes we have direct experiential access to the appropriate classifications of our mental phenomena. The assumption is that the appropriate use of

intentional language is to capture the genuine intentional states that we experience. But why think that intentional language would capture the boundaries of these mental states appropriately? Why assume that intentional language carves our mental life at its joints? This sort of story relies on commonly-used linguistic categories ("beliefs", "desires", "mental representations", etc) to draw conclusions about the existence of objective ontological states/properties. However, it is far from obvious to claim that we know we have intentional states like beliefs *because we experience them*. On the contrary, we never experience beliefs *qua beliefs*. At best we experience *something*, but how we linguistically classify it is up to us. The above position assumes that we have direct experiential access to the proper linguistic classifications of our mental phenomena. But how do we know that a particular mental event should be characterized in terms of a mental state like a belief or a desire (a state with intentionality), as opposed to using some radically different type of classification system altogether? Mental events by themselves do not tell us what classification we *ought* to give them. *We* do the classifying. And classifications in terms of intentional terminology may turn out to be ill-fitting to the mental phenomena we are trying to capture.

So we are now inevitably left with the question: Do we use intentional language because there is a robust ontology of intentionality (because there *really are* intentional states, and our language is trying to capture that), or do we adopt a robust ontology of intentionality because we use intentional language (given that intentional language is pervasive and beneficial in our lives, we cannot help but interpret ourselves and others in terms of intentional states)? We describe systems using intentional language in scientific contexts all the time regardless of whether we think those systems "really" have the mental phenomena of intentionality, "really" do not have it, or whether we simply cannot tell one way or another. The assumption that the proper use of intentional language is ultimately determined by our robust ontology of intentionality may simply not be the reason that intentional language is beneficial to scientific practice.

Those who embrace a robust ontological view of intentionality are happy to grant that intentional language can be beneficial to science even if the system described does not genuinely have intentional states. However, they point out that we must be cautious not to take such uses of the language literally. The pervasiveness of intentional language in all different contexts and domains just means we must take care to distinguish the genuine cases of intentionality (the systems that can be literally described as having mental representations, beliefs, and intentions) from the as-if cases (the systems that are merely *usefully described* in terms of "mental representations", "beliefs", and "intentions"). And, ontologically speaking, it is the literal cases that are philosophically important. Or as Searle puts it:

> [There is a] distinction between the sort of facts corresponding to ascriptions of intrinsic intentionality and those corresponding to *as-if* metaphorical ascriptions of intentionality. There is nothing harmful, misleading, or philosophically mistaken about *as-if* metaphorical ascriptions. The only mistake is to take them literally. (1992, p. 82)

But even if we suppose that such a story is true, it assumes that the literal cases are distinguishable from the metaphorical ones. But this downplays the fact that intentional language may be a powerful linguistic tool with pragmatic benefits, and so its usage will be pervasive in scientific practice regardless of whether we can keep track of the appropriate ontological distinctions. In which case, given its widespread and important usage as a pragmatic tool, it may be nigh impossible for us to tell apart the ontological cases from the mere pragmatic ones in our linguistic usage. We may be unsure whether we are using the terms literally or not in a given context just so long as the language provides the pragmatic benefits we need. In this sense, keeping straight the literal from the metaphorical becomes exceedingly difficult since we may simply have no way to tell which cases meet our ontological standards and which do not. So even if Searle's story were correct, it would be a mistake to assume that we must know the correct ontological underpinnings of intentional language in order for us to use it in productive ways. As

Mark Wilson points out, there are "situations where speakers employ terminology according to properly productive strategies yet entertain incorrect pictures of their underpinnings" (Wilson 2006, p. 308).

Of course, our intentional realist could point to the fact that just because there are some cases that are difficult to discern as genuine or metaphorical, this does not mean that there are no clear-cut cases we can rely on to validate the literal/metaphorical distinction as cutting along ontological lines. Human beings are a clear-cut example of genuine intentionality, while cars and cell-phones are not. Even though we can describe objects like cell phones or cars as-if they had intentional states, we never fool ourselves into thinking that they literally do. There is a clear difference between our *real* intentionality (often referred to as "non-derived" intentionality) and the ascribed intentionality that we use to *talk* about things like cars ("derived" intentionality). This literal/metaphorical distinction of clear-cut cases seems to cut nicely across the ontological divide that the intentional realist wants. To deny that there is a difference between our genuine mental content and the "content" we ascribe to cars or cell phones seems almost a *reductio ad absurdum* of the position that wants to tear down this construal of the literal/metaphorical distinction.

But this sort of response is deeply confused as to the issue at hand. To deny that the appropriate literal use of intentional language must conform to our robust ontology of intentionality is in no way to deny that there are important differences between us and cars or cell phones. On the contrary, it is the fact that we *are* so different from such objects, and so much more complex, that we cannot help but think that one of the many ways in which we differ from them must be in terms of this thing called *intentionality*. It is this vast gulf in abilities, experiences, and capacities that leads us to draw a distinction between the "real" intentionality in us, and the mere "metaphorical" intentionality in other things. But the mistake is assuming that of the many countless differences between us and things like cell phones or cars, one of them must be *some unique and genuine ontological property of intentionality*, and that *we* have it and *they* do not. But we can deny this without denying that there are important and powerful differences

between our interactions with the world, and a car's (or its lack thereof). Just as we can say we are alive and cars are not, even if we deny that we have an *élan vital* when the car does not.

This robust ontological story of intentionality slants our assumption of how intentional language works by implicitly suggesting that that language first developed as means of describing genuine cases of intentionality, and was then adapted to be used in metaphorical ways (to describe objects that do not have intentionality). Under this interpretation, we first learned to talk about *people* as having intentional states like beliefs and desires *because they really do*. Then, we learned to apply these concepts metaphorically to describe things that we know do not have it (cars, toasters, cell phones, etc). But why assume that this gets the order correct? Instead, we may simply learn to talk about all kinds of systems as having intentional states of some sort or another. We make no distinction between people and objects in terms of our intentional ascriptions. However, we are acutely aware of the fact that we are very different from other sorts of things. We are more complex, and have far more impressive capacities, behavior, and experiences than cars, toasters and cell phones. Thus, we conclude that when we talk about *ourselves* in terms of intentional states, then *we must mean something different* than when we talk about other things. This leads us to conclude that we must mean the terms literally when applied to us, and not when applied to other things. Yet, it might not be the ontology of intentional states that drives our use of intentional language. And assuming that it is blinds us to the scientific benefits of using such terminology.

Adopting this sort of robust ontological realism regarding intentionality from the outset can generate confusion in our scientific study of the mind by insisting on an explanation of a phenomenon that may be a mistaken interpretation of a linguistic tool. And if intentionality does, in fact, exist as a unique type of ontological phenomenon, it may be disconnected from the intentional language we use when we describe systems. In which case, intentional descriptions might mischaracterize the genuine mental occurrences of intentionality. As a result, we must get clear on how we use intentional language, and what its benefits are, before we start to adopt realist ontological interpretations of intentionality.

2.4.2 Intentional anti-realism

Using the ontology of intentionality as a guide for the usage of intentional language is equally as misleading and problematic if we start with an anti-realist stand on intentionality. If we conclude that intentionality does not exist as a robust ontological property or phenomenon, then the temptation is to assume that intentional terms do not refer to anything real, and so such language has no role in science.

Paul Churchland, for example, proposes that intentionality is not some metaphysical feature of mentality, but is instead a feature of folk psychological concepts (1981, p. 70). Moreover, he suggests that if these folk concepts cannot find a place within our best neurological theories, then such concepts have no scientific value and ought to be eliminated. The underlying assumption being that neuroscience (in virtue of being our best science of the mind) tells us about the genuine properties and states of the brain. Thus, if intentional states/properties cannot be reduced to neurological states/properties, then they do not pick out anything *real* about the system, and so they ought to be replaced by the superior account. But the relevant question is whether intentional language is actually used in fruitful and productive ways in science, and not whether such language fits into the account we feel is more ontologically justified.

We do not, for example, insist on having a clear ontological story about numbers before we allow scientists to use them in their investigations of the world. If we decide that numbers are not real, do we then *banish* them from science? Do we insist that mathematical models only be used in science until they can be replaced by the "real" descriptions of systems in terms of natural language (quantum mechanics would be in bad shape if that were the case)? Does it really matter whether they are real or not just so long as we use them in the way that undeniably aids scientific practice? Ontological questions about numbers can be left aside in our pursuit of science since their ontological status does not impact the beneficial and possibly ineliminable role that numbers play in scientific inquiry. So the case may be with intentional language as well.

There is also something deeply anti-naturalist about taking this eliminativist view of intentional language (while this may not be a concern for those who care little about naturalism, many who take eliminativist positions do so for naturalistic reasons). Specifically, if one believes that we ought to let science be our guide for ontology, then we must see what role intentional language *does* have in science. Otherwise, we are taking an ontological position (intentional states are not real) and then dictating to science what it ought to do based on our ontological convictions. But this seems to get things backwards if one has naturalist tendencies. Let us first see how intentional language is used in science, and if it is useful. Then, and only then, can we start talking about elimination.

2.5 Bringing It All Together

So where does all this leave us? Is intentionality a problem for mechanistic explanations of the mind? If we start with our ontology of intentionality, then the answer will be "yes" or "no" depending on the story we tell. If we take an ontological view that identifies intentionality as a unique ontological property that can never be studied from a "third-person point of view" (Searle 1992), then not only is the answer "yes", but in some sense, intentionality becomes non-naturalizable *a priori*. On the other hand, if you think that intentionality does not really exist and that intentional language should be abandoned in scientific practice as a result (Churchland 1981), then the answer is "no". However, I propose that we should not *start* with our ontology, and then stipulate the "appropriate" use of intentional terminology as a result, since this will bias our interpretation of intentional terms from the get go. Instead, we should see if and how intentional language is useful. And then determine what ontological story we want to tell about it as a result.

Chapter 3
Intentional Language Outside the Sciences of the Mind

In the previous chapter, I argued that we ought to leave our ontological commitments about intentionality at the door when studying the mind. Instead, we should focus on our use of intentional *language*. And while such language is most commonly associated with the study of the mind, it is not limited to this domain. It is not just psychologists that talk about systems in terms of their "aboutness." Such language permeates science at all different levels. The real question is whether other scientific domains only use intentional language in unnecessary and metaphorical ways, or whether such language plays any fundamental role in their practices and theories.

Some propose that intentional language cannot be anything more than unnecessary metaphor below the level of psychology (or possibly neuroscience depending on one's ontological story) since the phenomenon of intentionality only exists at that level (e.g. as properties of minds). It simply does not go deeper than that. As Jerry Fodor puts it:

> I propose that sooner or later the physicists will complete the catalogue they've been compiling of the ultimate and irreducible properties of things. When they do, the likes of *spin*, *charm*, and *charge* will perhaps appear upon their list. But *aboutness* surely won't; intentionality simply doesn't go that deep. (Fodor 1987, p. 97)

However, such an argument is primarily motivated by the ontological assumption that the genuine phenomenon, or property, of intentionality only exists at the level of minds. Since minds do not exist below the level of psychology or neuroscience, scientific domains dedicated to the study of lower-level phenomena can only use intentional language in metaphorical and non-essential ways. But as we saw last chapter, assuming an ontological story of what intentionality is *before* we look at how we use the language is a recipe for disaster. Therefore, if we leave these ontological issues aside, we are left with the

real question: do other scientific domains use intentional terminology in their scientific explanations and theories? In this chapter, I argue that the answer is yes. And moreover, that there are reasons not to consider these uses to be merely unnecessary metaphor. This will provide us with clues as to value that intentional language may have as a pragmatic scientific tool.

In order to demonstrate this, it is first important to examine our scientific use of the term "information". While the term is used in different ways both in, and out of, scientific contexts, one of the most common uses of the term is as an intentional term. This use of "information" ascribes intentional content to a system in much the same way one ascribes beliefs or knowledge to a system. I will demonstrate that scientific domains like genetics and molecular biology explicitly invoke this intentional notion of information in their study of systems, and that doing so has substantial scientific benefits. To begin, however, let us turn our attention to the concept of information more generally.

3.1 The Different Meanings of "Information"

What exactly do we mean when we talk of "information"? Depending on context, the term is often used in one of two ways: The first is the colloquial everyday sense of the term, while the second is a more technical sense defined by mathematical communication theory.[8] Our first step is therefore to determine exactly what differentiates these two senses, and which of the two is relevant to our question regarding the presence of intentional language in scientific domains.

3.1.1 Semantic information

The most common interpretation of information, often found in everyday usage, is in terms of intentional content. In other words, one can have a piece of information *about* one thing or another. I can, for instance, have information *that it is raining outside*. Similarly, a physics textbook contains pieces of

[8] It should be noted that there are multiple technical notions of "information" associated with communication theory. While I will concentrate on *Shannon* Information here, the claims made in this chapter will apply equally to all other technical notions as well.

information *about physics*. This conception of information is in-line with the sort of information one gets at an information booth. When we approach an information booth and ask for the location of the bathroom, we get *information about* the location of the bathroom.

Information of this sort is commonly referred to as *semantic* information, given its connection with the semantic properties of language. As Christopher Timpson explains:

> The everyday notion [of information] is a semantic and an epistemic concept linking centrally to the notions of knowledge, language and meaning; to that of a person (language user) who might inform or be informed. (forthcoming, p. 43)

Semantics, virtually by definition, is an intentional concept. When we use language, we take an arbitrary set of symbols, and use them to represent features of the world. This explains the way in which books, and written statements in general, can contain information. Recall that one of the defining characteristics of intentionality is that our intentional content can be false (See Section 1.2). In other words, beliefs can be correct or incorrect, representations can represent or misrepresent, etc. To quote an often-used philosophical motto: "no representation without misrepresentation" (Dennett 1987, p. 307). When it comes to semantic information, we similarly have cases of correct and incorrect information. One can be *informed* or *misinformed* based on whether the information correctly or incorrectly represents the world. I may be given incorrect information about the location of the bathroom from the person at the information booth for example. For this reason, the "possession of information or misinformation is just as Intentional a notion as that of belief" (Dennett 1971, p. 90). Given this, I will use the term "information^{i}" when referring to this kind of information (to emphasize its status as an intentional term).

3.1.2 Shannon information

While information^{i} is by far the most common interpretation of the term "information", it is not the only one. The concept of "information" as characterized by Claude Shannon (1948), for example, is different

enough from the intentional (or semantic) concept that a different term would not have been altogether inappropriate in order to avoid confusion. Unlike information^{i}, Shannon Information (hereafter information^{s}) is a part of communication theory, and is used to mathematically characterize correlations between objects or events. As a result, issues of intentionality and semantics are unrelated to information^{s}.[9] Edward Collin Cherry succinctly summarizes this point when he says that "we are not concerned with the meaning or truth of messages [when dealing with information^{s}]; semantics lies outside the scope of mathematical information theory." (1951, p. 383)

While information^{i} is concerned with content or reference (for example: information about *the location of the bathroom*), information^{s} is (roughly speaking) a measure of the quantity of information^{i} that can be passed between two sources (a transmitter and a receiver). Or, more precisely, with the ability of a receiver to reproduce a message that originated with the transmitter.

In this regard, information^{s} says nothing about the content or "aboutness" of any given message being transmitted from a transmitter to a receiver, and is concerned only with the act of transmission itself. As Shannon (1948) himself put it, the intentional concerns "are irrelevant to the engineering problems" of effectively transmitting messages between two locations. Thus, to say that an object contains information^{s} is not to attribute a representation —or intentional content of any sort— to the system in the way that attributing information^{i} to it would be. Instead, all that matters for information^{s} is that there exists a reliable correlation between two events:

> In this sense, any process at all in which there is a reliable correlation between two states can be described in terms of information. This is the sense in which dark clouds carry information about bad weather, and tree rings carry information about the age of the tree. (Godfrey-Smith 2004, p. 276)

[9] Some philosophers have attempted to construct an ontological account of information^{i} *out of* information^{s} (most

Another important difference between the two types of information is that information^{s} does not allow for error. As discussed above, information^{i} allows for the possibility of being *mis*informed. One can receive information from an information booth that is inaccurate. Not so with information^{s}. Information^{s} mathematically characterizes statistical correlations between events, and so cannot be wrong or incorrect. As a result, information^{s} fails a necessary condition for being an intentional term:

> People who are trying to distinguish genuine semantic properties from information in the Shannon sense often point to the capacity for error, and the ability to represent nonfactual situations, as marks of semantic phenomena. In the simple clouds-and-rain case, there is no sense in which the clouds could *misrepresent* the weather. The correlation between clouds and rain might fail to hold in some particular case, but that does not imply that the clouds said something *false*. If I tell you a lie in this chapter, however, my words have indeed been used to say something false. I can also use these words in describe situation that I know does not obtain, such as my having won the lottery. These are not features of information in the mere Shannon sense. (Godfrey-Smith 2004, p. 279)

With this broad sketch of the differences between information^{i} and information^{s}, we can now turn to the question of whether intentional terminology has a role in scientific domains concerned with studying phenomena below the level of the mind. While information^{s} is uncontroversially used in these domains (right down to the level of quantum physics), what about information^{i}? If so, then perhaps it can provide us with a clue to as to why it is used in these contexts.

3.2 Information and Biology

Talk of "information" in molecular biology and genetics is pervasive (Sterelny et al. 1996; Sterelny & Griffiths 1999; Smith 2000; Godfrey-Smith 2000, 2004; Sarkar 2004). But how exactly do biologists use the term? Do biologists only employ technical, and non-intentional, senses of information, or do they ever

---

notably Dretske 1981), but this project this still grants the *non*-intentional status of information^{s}.

refer to information[i]? In what follows, I will demonstrate that there is compelling evidence that biologists explicitly invoke an intentional notion of information in their theories of genetics. Before demonstrating this, however, I will demonstrate why the information used in these biological contexts cannot be information[s] (or other technical notions). Following that, I will argue that information[i] is the best candidate for the way in which the term is used. Lastly, I will look at some objections to this idea, and show how they can be overcome.

3.2.1 Shannon information in biology

The first thing to note is that information[s] does have an important role in biological theory. As Peter Godfrey-Smith points out, "a lot of discussion in contemporary biology is facilitated by this conceptual framework" (2004, p. 278). But while information[s] may have a role to play in biology, this cannot be the sense in which biologists use the term when they speak of information encoded in genes. A far more robust concept of information is needed to capture such usage. As Sahotra Sarkar points out:

> When, for instance, it is said that the haemoglobin-S gene contains information for the sickle cell trait, communication-theoretic information cannot capture such usage. To take another example, the fact that the information contained in a certain gene may result in polydactyly (having an extra finger) in humans also cannot be accommodated by communication-theoretic information. (Sarkar 2004, p. 260)

Consider the way in which biologists speak of genes encoding information about phenotypic traits. If this referred to information[s], then the claim would only be that there exists a correlation between genotypic and phenotypic traits. This is true. However, it is not particularly controversial or informative. The application of communication theory to the relationship between genes and phenotypic traits would thus be superfluous, and would do little work for us theoretically. As Sarkar puts it: "The trappings of

Shannon's model of a communication system are extraneously added to a relatively straightforward point about genetic and environmental correlations, and do no cognitive work" (2004, p. 265).

A more serious problem is that if genes only encode information^{S}, then there would be no reason to give special attention to the information in *genes* regarding phenotypic traits. Put another way, if genes only had information about phenotypic traits because of a correlation between the two, then environmental conditions would equally encode such information. As such, there is no *privileged* sense in which genes carry phenotypic information:

> The standard apparatus [of communication theory] defines information as the covariation between a signal and its source. Holding environmental factors constant, genotypes covary with phenotypes. But other factors causally relevant to development also carry predictive information. The plant on which the butterfly eggs are laid covaries with developmental outcomes. So genes predict phenotypes, but they are not alone in doing so. (Sterelny 2000, p. 196)

Yet, biologists often propose something much stronger than just covariation between genotypic and phenotypic traits when they speak of encoding information. Instead, they propose that the "gene is not merely correlated with a trait: that trait explains why that gene has its form" (Sterelny 2000, p. 197). The information in genes regarding phenotypic traits is more robust than mere information^{S}. This provides a reason to think that biology requires a richer concept of information than is provided by communication theory.

Another reason to view biological information as more robust is that biological information allows for error. As we discussed earlier, it is incoherent to talk of information^{S} being in error. Communication theory only characterizes the existing correlations between events. In this sense, the information from the transmitter cannot be *misread* or *misinterpreted* by the receiver. This is not the case with biological information: "Strikingly, genetic information is often described [in terms of] misrepresentations" (Sterelny & Griffiths 1999, p. 104). Genes are thought to determine phenotypic traits,

yet organisms can develop different phenotypic traits when certain atypical conditions are present. In which case, it is common to speak of the information encoded in genes as being *misinterpreted* by the receiver due to interference from these conditions (Sterelny & Griffiths 1999, pp. 104-105; Smith 2000, p. 193). Consider Sterelny's claim that "when a human genotype results in a phenotype with dwarfed arms, the information in it has been **misread**" (Sterelny 2000, p. 197 —my emphasis).

A third reason to differentiate biological information from information^{s} is that biological information allows for specificity, while communication theory does not. Information^{s} never specifies the *content* of a message being transmitted between a transmitter and a receiver. It merely mathematically characterizes the act of transmission. In this regard, information^{s} lacks specificity by not telling us what the specific content of a message is. Meanwhile, specificity is one of the defining features of biological information:

> The main problem is that, at best, communication-theoretic information provides a measure of the amount of information in a message but does not provide an account of the content of a message, its specificity, what makes it *that* message. [...] Capturing *specificity* is critical to genetic information. Specificity was one of the major themes of twentieth-century biology. During the first three decades of that century, it became clear that the molecular interactions that occurred within living organisms were highly "specific" in the sense that particular molecules interacted with exactly one, at most a very few, reagents. Enzymes acted specifically on their substrates. Mammals produced antibodies that were highly specific to antigens. In genetics, the ultimate exemplar of specificity was the "one gene-one enzyme" hypothesis of the 1940s, which served as one of the most important theoretical principles of early molecular biology. (Sarkar 2004, p. 260)

Here again we see that the concept of information at work is far richer than the one employed by communication theory. The specificity of messages is not accounted for by communication theory, and so not relevant to information^{s}. Meanwhile, it plays an essential role in biological information. In fact, accounting for specificity is what led to the idea that genetic information was contained within DNA in

the form of a *genetic code*: "The code explained the specificity one gene-one enzyme relationship elegantly: different DNA sequences encoded different proteins as can be determined by looking up the genetic table." (Sarkar 2004, p. 261). This comparison between biological states and symbolic code provides the first clue that biological information may have more in common with linguistic information than communication theoretic information. In fact, the analogy between genetic structures (as code-like) and linguistic structures is more than merely a passing similarity:

> Routinely, talk of information is intertwined with linguistic metaphor, from both natural and artificial languages: there is a genetic *code*, because a triplet of DNA (or RNA) nucleotides codes for each amino acid residue in proteins (polypeptide chains); there are alternative *reading frames* —DNA is *transcribed* into RNA, RNA is *translated* into protein, RNA is *edited*, and so on. (Sarkar 2004, p. 260)

But are the similarities between linguistic information and biological information merely superficial? How deep does this analogy between genetic code and language go?

3.2.2 Intentional information in biology

If the information stored in the genetic code is sufficiently like the information stored in linguistic structures, then we seem to be left with a semantic interpretation of biological information. But how seriously are we to take the analogy between genetic code and language? Is it merely a useful heuristic analogy, or is it something more substantial? There is reason to think that it is indeed something more substantial:

> Since the discovery of the DNA double helix in 1953, many biologists have employed language as a useful metaphor to describe certain aspects of molecular biological phenomena. But recently it was postulated that language is more than just a metaphor and that linguistics provides a fundamental principle to account for the structure and function of the cell. (Ji 1999, p. 411)

This idea that the similarities between human language and the genetic code are much deeper than mere unnecessary metaphor, and play a substantial role in our understanding of biology (and biological information), has been argued for by a number of biologists and philosophers (Garcia-Bellido 1984; Sereno 1991; Ji 1997, 1998, 1999; Sterelny & Griffiths 1999; Smith 2000). To see why, consider some of the defining features of language. First, language is symbolic. Second, language is combinatorial. Third, the symbols we use in language are arbitrary (there is no *causal* relationship between the form of the words we use and the things they represent). When we look to the genetic code, we find these features as well:

> The main reason why the informational framework became central to the new molecular biology of the 1950s and 1960s was the characterization of the relationship between DNA and proteins as a universal genetic code. [...] Three factors make the informational interpretational of this relationship illuminating: (a) the relationship can be viewed as a symbolic one, with each DNA triplet being a symbol for an amino acid residue; (b) the relationship is combinatorial, with different combinations of nucleotides potentially specifying different residues; and most importantly, (c) the relationship is arbitrary in an important sense. Functional considerations may explain some features of the code –why, for instance, an arbitrary mutation tends to take a hydrophilic amino-acid residue to another such residue– but it does not explain why the code is specifically what it is. The physical mechanisms of translation do not help either. The genetic code is *arbitrary*. Along with specificity, this arbitrariness is what makes an information account of genetics useful. (Sarkar 2004, p. 266)

Just as there is no causal connection between the *form* of the words that we use and their *meanings*, so too is it the case that "in molecular biology, inducers and repressors are 'symbolic': [...] there is no necessary connection between their form (chemical composition) and meaning (genes switched on or off)" (Smith 2000, p. 185). Similarly, consider John Maynard Smith's claim that "linguists

would argue that only a symbolic language can convey an indefinitely large number of meanings. I think that it is the symbolic nature of molecular biology that makes possible an indefinitely large number of biological forms" (2000, p. 185). These are not the only similarities between genetic code and language either. As Sungchul Ji points out:

> Both human and cell languages can be treated as a 6-tupe {*L, W, S, G, P, M*}, where *L* is the alphabet (i.e., a set of basic symbols called *protosemata*), *W* is the vocabulary or lexicon (i.e. a set of words), *S* is an arbitrary set of sentences, *G* is a set of the rules governing the formation of sentences from words (the *first articulation*) as well as the formation of words from letters (the *second articulation*), *P* is a set of physical mechanisms realizing and implementing a language, and finally *M* is a set of objects (both symbolic and material) or processes referred to by the words and sentences. (1999, pp. 411-412)

Given this close relationship between genetic code and language, it has been argued that the only way to make sense of biological information is as semantic or intentional information: "The idea that genes have meaning in something like the way that human thought and language have meaning is lurking in the background of many discussions of genetic information" (Sterelny & Griffiths 1999, p. 104). This suggests that biological information may, in fact, be information[i].

Consider that one of the defining features of information[i] is that it allows for misrepresentations (while technical notions of information do not). In the case of biological information, we do find cases of misrepresentation: "In biology, mis-representation is possible because there is both an evolved structure carrying the information, and an evolved structure that receives it" (Smith 2000, p. 193). Allowing for misrepresentations in biological information pushes the concept into the realm of the intentional:

> Any talk of genes being misinterpreted, or of the information in the genes being ignored or unused, is a shift from the purely causal notion of information toward something like the intentional notion. So one way to make sense of the idea that some developmental pathways are

> programmed while others are misreading of the program is to suppose that genes contain intentional information rather than causal information: information that remains the same when the channel conditions change. (Sterelny & Griffiths 1999, pp. 104-105)

Smith echoes this point when he argues that biological information satisfies the criteria for being intentional, and that non-intentional notions of information (like information^{s}) simply cannot account for the way the term is used in biological contexts:

> In colloquial speak, the word 'information' is used in two different contexts. It may be used without semantic implications; for example, we may say that the form of a cloud provides information about whether it will rain. In such cases, no one would think that the cloud had the shape it did because it provided information. In contrast, a weather forecast contains information about whether it will rain, and it has the form it does because it conveys that information. The difference can be expressed by saying that forecast has intentionality, whereas the cloud does not. The notion of information as it is used in biology is of the former kind; it implies intentionality. It is for this reason that we speak of genes carrying information during development, and of environmental fluctuations not doing so. (Smith 2000, p. 193)

The way in which we use "information" in genetics displays the signs of being an intentional term. This is why Sterelny & Griffiths propose that "intentional information seems like a better candidate [than non-intentional information] for the sense in which genes carry developmental information and nothing else does" (1999, p. 104).

Whether or not information^{i} really has a critical role to play in biological theory is not without its controversy however. While the analogy between genetic code and language is undoubtedly strong, some suggest that we must be cautious how we approach it (Sarkar 1996, 2004: Griffiths 2001; Godfrey-Smith 2004). I will now examine some of these worries in detail and determine whether they provide a substantial threat to this picture.

3.2.3 Is intentional information important to biology?

I present here two common arguments raised against the value of information[i] to biology. First, that the information-framework may have a legitimate role in biology, but that this account cannot be making use of information[i]. Second, that information-talk itself may in fact have no real substantial role to play in biology. I will demonstrate that neither of these arguments are conclusive, and that we still have substantial reasons for viewing information[i] as a key part of biology.

Let us begin with argument that "information" cannot be meant intentionally in the context of genetics. It has been argued that even if information-talk is an important part of biological theory, we have good reason to think that it is not the semantic kind. One reason being that we simply do not have a handle on what the ontological properties of semantic information actually are, and so no good way to know when something has them:

> When we leave the precise Shannon sense of information, we encounter an unruly collection of different concepts. We encounter the larger and controversial domain of *semantic* properties – properties that involve, representation, reference, truth, coding, and so on. Despite a massive effort by philosophers and others over many years (especially the past 100 years), I think we do not have a very good handle on this set of phenomena. (Godfrey-Smith 2004, p. 278)

Given this, there are serious worries with the assumption that biological information is information[i]. Namely: that we have no way of knowing whether the information in biology has the properties associated with genuine intentional information or not. We need some way of determining whether biological information is genuinely intentional before we start treating the term as an intentional one. Yet, there is no empirical test for determining the presence of intentionality:

> So do genes encode information for phenotypic traits in something more than the Shannon sense? Answering this question is made awkward by the absence of a good philosophical theory of

> semantic properties, the absence of a good *test* that we could apply to genes. Most philosophers will agree that we do not have a clear diagnostic question to ask. (Godfrey-Smith 2004, p. 279)

This inability to verify the existence or inexistence of information[i] in a system casts doubt on whether such an intentional concept has any place in biological theories. With this in mind, Sarkar argues that we have no reason to think that biologists need to account for some ontological property of intentionality in their study of biological systems. As he puts it:

> There is no reason to suppose that any concept of biological information must be "semantic" in the sense that philosophers use that term. Biological interactions, at this level, are about the rate and accuracy of macromolecular interactions. They are not about meaning, intentionality, and the like; any demand that such notions be explicated in an account of biological information is no more than a signifier for a philosophical agenda inherited from manifestly nonbiological contexts, in particular from philosophy of language and mind. It only raises spurious problems for the philosophy of biology. (2004, p. 262)

The thing to note about these sorts of objections is that they fall prey to the exact problem that was discussed in Chapter 2. Specifically, that the only way to justify the use of an intentional concept in science is by first providing an *ontological account* of intentionality as a unique property or phenomenon. Only then we can determine what things genuinely have it, and whether we are permitted to describe them in that way. Yet, as we saw, this way of approaching intentional language is more of an obstacle to scientific progress than a part of it. Sarkar's assumption that semantic information has no role in biology because biologists should not have to develop an ontology of intentionality gets things backwards. If biologists can and do use intentional language, then this may be a reason why we ought to *re-think* our ontology of intentionality. Or at the very least, it provides us with good reason for thinking that the benefits of intentional attributions may be distinct from whatever ontological phenomenon intentionality turns out to be.

The relevant question is not whether we can generate a test to determine if the ontological property or phenomenon of intentionality is floating somewhere in our genes. The relevant question is whether or not biologists use "information" in analogous ways to the way we use other intentional concepts. If they do, then this is a reason to view the term as an intentional term. And as we saw in Section 3.2.2, there are good reasons for thinking this. The question of what ontological inferences we ought to draw from this are still up for debate. It does not, however, commit us to the idea that intentionality necessarily exists as some ontological property of genes that we must test for. At the moment, we are concerned only with whether scientists make use of intentional language in their theories. And so these ontologically-based objections do not provide evidence that they do not, nor that it is not beneficial to do so.[10]

Let us now turn our attention to the argument that information-talk in *any* form has no real substantial role to play in biological theory. This objection is based on the idea that "genetic information is no more than a metaphor masquerading as a theoretical concept" (Sarkar 2004, p. 266). Those who endorse this position suggest that the benefits provided by information-talk are at best extremely limited, and often greatly misleading:

[10] One might argue that Godfrey-Smith's point is not an ontologically-based one, but instead a conceptual one. In other words, given that we do not have a good handle on the *concept* of semantic information, we have no rigorous way of determining whether we are applying the term properly in the case of biology and genetics. Thus until we do, we cannot know whether the term has a substantial role to play in biology. The problem with this sort of argument is if conceptual uncertainty is a reason to deny a concept a place in scientific discourse, then the uncertainty surrounding the status of biological information in general ought to likewise force us to deny it a place in biology. Yet Godfrey-Smith argues that the information-framework provides substantial benefits to biology (2004). Quite often the meaning of a concept in a scientific theory is determined by the role it plays in the completed theory. Thus to insist on a clear meaning of the term *before* the theory is permitted to use it seems to be in conflict with the way theoretical terms are introduced and used in science.

But this interpretation of Godfrey-Smith's point (as highlighting a *conceptual* problem as opposed to an ontological one) is hard to support regardless. He argues, for example, that we have no diagnostic test we can apply to genes in order to determine if they have information[i]. However, if the problem is with the uncertainty of the concept itself, then the lack of such a test would be irrelevant. To insist on a test that can be applied to genes presupposes that intentionality or semantics is a property *of genes* that we can test for. Yet, if the problem is a confusion with the *concept* of information[i], then no test *of the genes* will be illuminating to this problem. Instead, the problem is a lack of diagnostic test for the *biologists* regarding their *use* of the concept. This more ontological

> Even the most charitable attitude toward the use of "information" in genetics can only provide a defense of its use in the 1960s, in the context of prokaryotic genetics (i.e., the genetics of organisms without compartmentalized nuclei in their cells). Once "the unexpected complexity of eukaryotic genetics" (Watson, Tooze, and Kurtz, 1983, ch. 7) –that is, the genetics of organisms with compartmentalized nuclei in their cells– has to be accommodated, the loose use of "information" inherited from prokaryotic genetics is at least misleading. (Sarkar 2004, p. 261)

Sarkar is suggesting that the more we observe complex biological phenomena, and develop increasingly intricate theories of genetics, the more our concept of "information" seems to fail to capture, or account for, the relevant features of genes. Information-talk therefore appears to fail at being predictive or explanatory in any rigorous scientific sense. For example, consider that information-talk is often thought to be useful in predicting amino acid sequences. Sarkar (1996) points out that this way of talking fails to be predictive in a number of cases due to complications such as variations from the universal code. These sorts of problems suggest that "there is good reason to believe that such talk of information in genetics may be unnecessary" (Sarkar 2004, p. 266).

The problem with this objection is that the failure of the information-framework is not nearly so apparent or obvious as it is made to seem. Smith, for example, strongly disagrees with Sarkar, and argues that "the concept of [information] played a central role in the growth of molecular genetics" (Smith 2000, p. 192). He similarly argues that Sarkar ignores the substantial successes that information-talk has brought to a rigorous study of biology:

> I think that Sarkar is over-eager to point to the failures of the information analogy and to play down its successes. For example, he does not explain that the discovery of the relationship between DNA and protein –as a triplet code in which the correct 'reading frame' is maintained by

---

interpretation is further supported by his claim that we have an "absence of a good philosophical theory of semantic **properties**." (my emphasis).

> accurately counting off in threes, and whose meaning can be destroyed by a 'frame shift' mutation also arose from the coding analogy. (2000, p. 184)

Take, as another example, the predictive value of the information-framework. While Sarkar suggests that the use of information-talk fails to effectively predict amino acid sequences, Smith points out that this is simply untrue:

> As a second example, Sarkar's argument that the code does not enable one to predict amino acid sequences (because of complications such as introns, variations from the universal code, etc) is seriously misleading: biologists do it all the time. (Smith 2000, p. 184)

In response, Sarkar grants that the attribution of information *has* had substantial benefits to biology in past. And not only that, but that the success of this way of talking is indeed something that must be explained: "I do not deny that the informational framework for molecular genetics has a certain perspicuity that its critics must explain and incorporate into their own putative alternatives" (Sarkar 2000, p. 208).[11] He merely questions whether or not it will "*continue* to be of explanatory value in contemporary biology" (2000, p. 208, emphasis in text). He suggests that current and future developments in genetics will make the informational-framework obsolete in biology. Of course, whether or not this will turn out to be the case is hardly obvious. And, interestingly, Sarkar eventually concedes in one of his later writings (2004) that information-talk might actually play an indispensible role in biology.[12] He says:

[11] For my purposes, this is all that really needs to be granted. Do biologists talk this way? Yes. Is it beneficial? *Yes*. In fact, if one is to argue against this way of talking in science, they must explain *why* this way of talking has proven beneficial in the past, and *if* other ways of talking can perform the same tasks.

[12] His argument in that paper actually shifts from one against information-talk *tout court* to one against interpreting biological information as information[i]. His argument, however, is based entirely on the claim that we have no reason to think that genes contain the *ontological property* of intentionality. And this, as we've seen, is an insufficient reason to deny "information" the status of being an intentional *term*.

> Routinely, talk of information is intertwined with linguistic metaphors, from both natural and artificial languages. [...] The use of such talk is so pervasive that it almost seems impossible that, short of pathological convolution, the experimental results of genetics can even be communicated without these resources. (2004, p. 266)

What then can we conclude about information in biology? While the use of the information-framework is not without controversy, there is a great deal of evidence that it has had substantial success as a part of our biological theories. There is also strong evidence that "information" in these contexts is a distinctly *intentional* term, and is not employed merely as a kind of unnecessary metaphor. This provides us with good reason to think that intentional language plays an important scientific role in biology and genetics.

The metaphysical and ontological assumptions concerning intentionality as a mental phenomenon often convince people that intentional language belongs primarily at the level of psychology (and possibly neuroscience). If intentional descriptions are instead interpreted first and foremost as a type of linguistic tool, then it becomes a legitimate question which scientific domains the tool is useful *for*. What I have demonstrated here is that intentional descriptions may have substantial uses in scientific domains like biology. This brings us to the key question: what exactly do intentional descriptions, as a linguistic tool, tell us about systems? How do intentional descriptions in biology relate to those in neuroscience or psychology? It is these question we will turn to in the next chapter.

Chapter 4
What Does Intentional Language Tell Us?

Last chapter, I argued that intentional language can be found in scientific domains other than the sciences of the mind (such as biology).[13] So what is it that ties together the different uses of intentional language in the different scientific domains? In this chapter, I examine different possible answers to this question, and see if the example from last chapter can help illuminate the answer.

4.1 Nothing

The first option is that *nothing* ties them together. Intentional language may ultimately have no scientific value whatsoever. As such, attempting to find a common pragmatic value that connects the different scientific usages together may be for naught. There is no value to connect.

The problem with this sort of pessimistic view is that we seem to have empirical evidence to the contrary. Consider the example from the last chapter. The attribution of information[i] to genes in biology is more than just a useless metaphor or unnecessary linguistic tendency; such attributions have been immensely valuable to genetics (see Section 3.2.3 for details). As another example, consider our use of intentional language in neuroscience. David Marr (1982) famously argues that understanding neurological mechanisms in terms of the information[i] they contain has been an essential part of neuroscientific practice (p. 19). Or consider cognitive science. One of the founding principles of this discipline is that we can explain and predict the behaviour of systems by attributing to them *representations* and *rules* that operate over them.

To insist that intentional language has no value to science would seem to conflict with much of actual scientific practice. Of course, one might insist that the value of intentional language to science can

[13] There is also evidence (albeit far more controversial) that intentional notions of "information" even exists at level of physics (see, for instance: Wheeler 1990, and Zeilinger 2005)

be *better* achieved by other means (Churchland 1981; Stich 1983), however this is a different issue altogether. The question of whether intentional language is indispensible, or whether other accounts are superior, is an issues I will turn to later (See Sections 8.1 and 8.2). At present, the issue is merely whether intentional descriptions have scientific value, and the evidence certainly suggests that they do. So it is a legitimate question to ask what this value is.

4.2 Nomological Laws

One possible answer for what ties together the different uses of intentional language is that they are all used to characterize the same sort of law-like relation that connects certain physical states (like brain states) to the things that they are about. This is a position famously argued for by Fodor. As he puts it:

> …what bestows content on mental representations is something about their causal-cum-nomological relations to the things that fall under them: for example, what bestows upon a mental representation the content *dog* is something about its tokenings being caused by dogs. (Fodor 1998, p. 12)

Similarly, according to this view, when I misrepresent the world (for instance, when a cat causes a dog-representation in me), then there is also a nomic relation between the cat and my dog-representation. However, *that* nomic relation is parasitic on the relation that connects *dogs* to my dog-representation. In essence, a cat can only sometimes cause a dog-representation in me because *dogs* cause dog-representations in me (and cats can be relevantly similar to dogs in certain contexts). This view would suggest that all intentional language (when used appropriately) is beneficial to science in virtue of characterizing a particular sort of causal nomic relation (constitutive of meaning) that connects certain types of physical states to the things they represent.

The problem with this idea is that it is extremely difficult to make sense of the notion of information[i] discussed last chapter in this way. In other words, it is highly dubious that information[i] is a

physical state within our genes that has a particular sort of causal nomic relation (constitutive of meaning) between it and the phenotypic traits it is "about". The information within our genes is not best understood as a physical state whose tokenings are caused by the presence of phenotypic traits. And so the benefits of intentional language in such cases seem not to be connected to the presence of this sort of nomic law.

Of course, one might insist that the use of intentional language in the case of biology is inappropriate or metaphorical. However, this does not change the fact that the language *is* intentional, and that it is beneficial to scientific practice. Similarly, there are reasons to doubt whether the benefits of intentional language in other domains are tied to this idea of nomic laws either, even in domains like psychology and neuroscience.

The suggestion that there are particular sorts of metaphysical representational laws (or nomic relations) that connect certain physical states to objects in the world, can only be scientifically useful to us if we have some principled means of determining or identifying these relations. Otherwise, how can we tell whether our fruitful scientific use of intentional language in any way corresponds with the appropriate nomic laws or not? To demonstrate, consider the following example described by Daniel Dennett:

> Consider a standard soft-drink vending machine, designed and built in the United States, and equipped with a transducer device for accepting and rejecting US quarters. Let's call such a device a two-bitser. Normally, when a quarter is inserted into a two-bitser, the two-bitser goes into a state, call it *Q*, which "means" (note the scare-quotes) "I perceive/accept a genuine U.S. quarter now." Such two-bitsers are quite clever and sophisticated, but hardly foolproof. They do "make mistakes" (more scare-quotes). That is, unmetaphorically, sometimes they go into state *Q* when a slug or other foreign object is inserted in them, and sometimes they reject perfectly legal quarters –they fail to go into state *Q* when they are *supposed to*. (Dennett 1987, p. 290)

Now imagine I say of the two-bitser that it has a "representation" of a quarter when, upon receiving my inserted quarter, it goes into state *Q*. Does this case involve the appropriate nomic relations connecting

state *Q* to the quarter in such a way as to be a legitimate scientific use of intentional language? Or is it merely a case of inappropriate, or derived (i.e., metaphorical) intentionality? According to Fodor, the vending machine lacks the appropriate nomic relations to be a literal case of intentional language:

> "That sort of case is irrelevant," Fodor retorted instantly, "because after all, John Searle is right about one thing; he's right about artifacts like that. They don't have any intrinsic or original intentionality –only derived intentionality." (Dennett 1987, p. 288)

But how can we tell? Could we not say that when the vending machine is in state *Q*, there is a nomic connection between it and the quarter it represents? Similarly, when we "fool" the machine with a slug, it goes into state *Q* because there is an asymmetrical dependency relation between the machine's representation of (certain kinds of) slugs, and its representation of quarters. In which case, our description of the vending machine would be an appropriate use of literal intentional language.

Perhaps one might protest that we need not posit this sort of nomic relation between state *Q* and the quarter in order to explain the vending machine's behaviour. We can account for the correlation between state *Q* and the quarter entirely by way of physical mechanisms (without the need to appeal to any metaphysical representational laws). But this sort of argument assumes that nomic relations are brute metaphysical laws that are built into the fabric of the universe. However, Fodor explicitly denies this:

> I now add the considerably less tendentious assumption that if there are such meaning-making laws, they surely couldn't be basic. Or to put it another way, if there is a nomic connection between *doghood* and *cause-of-DOG-tokeninghood*, then there must be a causal process whose operation mediates and sustains this connection. Or, to put it another way, if informational semantics is right about the metaphysics of meaning, there must be mechanisms *in virtue of which* certain mental (-cum-neural) structures 'resonate' to *doghood* and *Tuesdayhood*. […] Mechanisms of semantic access are what sustain our ability to think *about* things. (Fodor 1998, p. 75)

But if this is the case, then why can we not claim that the two-bitser is the "causal process whose operation mediates and sustains" the connection between state *Q* and the quarter? The appropriate nomic relation connecting state *Q* to the quarter is the direct product of that mechanism. And so again, we would have a clear-cut case of intentionality and not mere derived or metaphorical intentionality. Clearly more is needed to determine what sorts of nomic relations are off the table as sufficient for appropriate scientific intentional descriptions. But Fodor has no story about what such nomic laws are like. As Daniel Hutto points out, Fodor "fails to give a scientifically respectable explanation of the dependency relationship" (Hutto 1999, p. 48). And so how can we tell if the scientifically fruitful uses of intentional language in psychology and neuroscience are truly those that characterize this nomic relation?

If we can, and do, use intentional language in scientifically fruitful ways that do *not* conform to the sorts of nomic laws Fodor suggests (as in the biology case), and we have no principled means of telling whether these nomic laws exist in any other contexts, then what reason do we have for thinking that this nomic relation is the foundation for our fruitful uses of intentional language in science? The evidence we have about the usage of intentional language in science seems to either contradict this position, or be such that we can never confirm it. Such an account, therefore, has little benefit for us scientifically.[14]

[14] A similar problem applies to Fred Dretske's attempt to develop an account of information[i] in terms of information[s] (1981). Like Fodor, Dretske also claims that intentional language characterizes a kind of nomic relation. He proposes that this nomic relation can be built up out of the sort of relations characterized by communication theory. However, Dretske's account faces all the same troubles that Fodor's does. First, it seems unable to account for examples such as the application of information[i] in biology discussed last chapter. Second, Dretske claims that not all nomic relations arising out of information[s] are sufficient to count as intentional, but he provides no story of how to tell which nomic relations are the relevant ones. Consider Dennett's two-bitser example. According to Dretske, such a simplistic machine would "lack something that is essential" for intentionality (1985, p. 23). But why think so? We can easily characterize the relation between state *Q* of the two-bitser, and the quarter it represents, in terms of information[s]. So why assume that this case is not one that captures the appropriate nomic relations for intentionality? Without a more definitive set of criteria, Dretske's account has the same faults that plagued Fodor's account.

4.3 Distinct Biological Property

Another possible answer for what connects the different uses of intentional language in science is its ability to characterize some unique property of biological matter. This idea, suggested by Searle (1980, 1992), proposes that intentionality is some distinctive property of biology, and that the appropriate use of intentional language in science is one that identifies this property. In this sense, the case of information[i] in biology really might be an appropriate use of intentional language (given the biological nature of our genes). But using such language to characterize *computers*, as engineers often do, would be a mistaken (or at least metaphorical) use of the language. While Searle grants that talking about non-biological systems as having intentional states may still be extremely beneficial to science, he insists that the appropriate literal use of the language is to characterize this unique biological property.

The problem with this account of intentional language is that it faces the same sorts of problems that our account of nomic laws faced: we have no scientific means of determining *what* this unique property of biology is, or how it is generated. We can, and clearly do, use intentional language in science to fruitfully characterize non-biological systems all the time, as we commonly see in work on artificial intelligence. So what is the value of talking this way in those contexts? Clearly it is not its ability to characterize some particular biological property. Or consider Dennett's two-bitser example again. Does the vending machine lack the appropriate biological property necessary to literally have a representation? If so, how can we tell? Searle provides no guidance as to what this property is, how we can find it, or why only biological matter might have it.

In this respect, Searle's account bears a striking similarity to the *vital force* theory of life that was abandoned long ago. To demonstrate, imagine a vitalist; someone who believes that life —*real honest to goodness life*— is a metaphysical force that inhabits certain systems. Now imagine that our vitalist fully grants that many scientists can, and do, still talk of systems that lack the vital force as being "alive" in ways that are productive and fruitful to scientific practice. He merely insists that the appropriate literal

use of the term "life" (and the real scientific value of life-talk) is to characterize this metaphysical force. However, our vitalist has no idea what this vital force is, and grants that we have no current scientific means of ever telling which systems have it and which do not, nor any explanation for why.

Under these circumstances, the vitalist's conception of "life" is simply unhelpful to us scientifically. It provides no value at all. Moreover, if our fruitful and productive scientific use of the term "life" does not cut along the lines that the vitalist insists that it should, then its value to science is clearly not what the vitalist insists that it is. I propose that the same holds true for Searle's account of intentionality. He insists that it is a metaphysical property that certain systems have, but has no story about what it is, and grants that we have no current scientific means of ever telling which systems have it and which do not, nor any explanation for why. He also admits that intentional language is used in fruitful scientific ways that do not conform to his account (the so-called "metaphorical" cases). As such, I propose that the value of intentional language to science is not what Searle insists it is.

4.4 Prediction and/or Abstract Mechanistic Descriptions

4.4.1 Intentional language and prediction

So what then do we use intentional language for? Perhaps we can find a clue in our example from last chapter. What are the benefits of talking about genes in terms of information[i]? According to Smith, it allows biologists to predict amino acid sequences. Not only that, but biologists use it to make such predictions "all the time" (Smith 2000, p. 184).

With this in mind, I propose that what ties together the different uses of intentional language in the various scientific domains is its ability to help us generate predictions of systems. But even if this is true in the case of biology, is it equally true in the sciences of the mind? Some philosophers have explicitly argued for this idea; the most well known example being Daniel Dennett (1971, 1987, 1991a, 1991b). Dennett proposes that our use of intentional language involves our taking a *stance* towards a

given system in order to make predictions of it (for a more detailed account of Dennett's position, see Section 9.5). Jerry Fodor similarly champions the predictive successes of intentional attributions in our scientific study of the mind (1987). Even Paul Churchland, who argues that intentional language ought to be eliminated from science due to its explanatory deficiencies, still grants that intentional language allows us to predict "the behavior of other persons with a facility and success that is remarkable" (1981, p. 68). We also know that it is common practice in neuroscience and cognitive science to make predictions by interpreting systems in terms of *representations* and *information*[i]. And so there certainly seems to be evidence for the predictive value of intentional language. But is the value of intentional language really its predictive power, or is its predictive power just a by-product of a more important feature of the language: namely, its ability to abstractly describe the functioning of mechanisms working within the system?

4.4.2 Do we use intentional language as abstract mechanistic descriptions?

According to some philosophers and cognitive scientists, the reason for the predictive successes of intentional models is because the ascribed intentional states correspond to sub-personal physical mechanisms operating inside the system that are causal in its behaviour, and so its predictive success comes from its mechanistic interpretation of the system. This intuition can be deceptively bolstered by the fact that many different scientific accounts of the mind argue that at least some intentional states do correspond to physical sub-personal mechanisms. For instance, connectionists propose that we can understand (at least some) intentional states as being nodes in a neural net, and it is this neural net that generates behaviour. Or consider the *neural engineering framework* (Eliasmith & Anderson 2003). According to this framework, at least some intentional states can be understood as patterns of activation in the brain that are best modeled as vectors in a multidimensional state-space. Or take a very different sort of story: Fodor proposes that intentional attributions correspond to mechanisms in the brain that are at a higher-level than those described by neuroscience (for more details on Fodor's account, see Section 9.2). While these accounts do not claim that *only* intentional states which correspond to sub-personal

mechanisms will prove predictively fruitful, those who have his intuition may view these accounts as providing evidence for such a position. These accounts can give the impression that intentional attributions act as a kind of abstract mechanistic description of the system.

To further emphasize this point, consider once again Dennett's example of the soft-drink vending machine. When I say that the vending machine can represent a quarter, I am proposing that the vending machine has the ability to distinguish quarters from non-quarters, and can make inferences (to output a can of pop, to ask for more change, etc) based on its having this representation. Now, in this case, its ability to do this is in virtue of a particular "sub-personal" mechanism working within the vending machine (the two-bitser). It is this sub-personal mechanism that explains the discriminatory behaviour of the system, and thus allows us to predict it by attributing to it the relevant representation. And so under this account, intentional descriptions are useful to science because they are really just roundabout ways of providing abstract mechanistic accounts of systems, and predicting based on those accounts. In other words, some propose that the real benefit of intentional language is its ability to tell us about the functioning of sub-personal mechanisms. We can then use this information to form predictions.

4.4.3 Prediction does not require abstract mechanistic descriptions

I propose that this idea is ultimately misguided, and does not fit with our actual use of intentional language to make scientific predictions. This is not to say that we do not sometimes find correlations between the attribution of a relevant intentional state, and a sub-personal mechanism. And indeed, our understanding of systems is greatly enhanced in cases where we *do* find such correlations. However, such instances are not what validates the use of intentional language in science. Our use of intentional language to make fruitful predictions in science often does not require this correlation. In other words, it is not in virtue of corresponding to sub-personal mechanisms that intentional attributions are scientifically relevant.

This is reflected in our actual scientific practices in two important ways. First, the existence (or inexistence) of such a correlation can often not be relevant to our scientific purposes when employing intentional language. And second, there is empirical evidence that intentional attributions which do not conform to physical mechanisms are used to make successful predictions in science.

Let us examine the first issue. I propose that even if there is no underlying mechanism that corresponds to the intentional attribution, the predictive benefits of employing such terminology are often what we care about. Imagine that we are developmental psychologists interested in learning about the ability of infants to grasp items out of their reach. To conduct this experiment, we present the child with a desired object. Then, we interpret the child's future actions *given her goal of* trying to reach the object. Now, do we suppose that the scientific value of attributing a "goal" to her in such a context is entirely determined be whether there is a particular goal-state explicitly encoded in the brain? Suppose there is only a direct connection between the visual stimulus and the motor behaviour? In such a circumstance, there may be no particular mechanism in the brain that corresponds to an explicitly represented goal that the infant has, and which is causal in her behaviour. Would this lack of a correlation between the intentional ascription, and a particular physical mechanism, imply that such an intentional attribution would cease to be useful in this context? I propose not.

In such an experiment, we are not treating the intentional attribution as a kind of abstract mechanistic description. We are uninterested in the question of whether goal-states are neurological mechanisms, since it is not what we are trying to learn about the infant. All that matters is that the child *behaves in accordance with such a goal* so that we can study her grasping ability. Depending on the why-question being asked by the psychology, it may simply not matter whether there is a particular mechanism correlated with the attribution of the goal or not. And so the value of the intentional term is not its ability, or inability, to correlate with some particular causal mechanism. Instead, its value is in its ability to help us predict the general behaviour of the child (her pursuit of the desired object) so that we can observe

other behaviours and capacities. In this respect, not all fruitful and important scientific uses of intentional language need to correspond to particular physical mechanisms, just so long as they are predictively valuable.

But even if we happen to be uninterested in mechanistic correspondence when using intentional language in science, it may be a contingent fact about intentional language that the only time it ever *is* predictive is when there *is* such a correspondence. In other words, the predictive value of intentional language may be derived from its abstract mechanistic account of the system. And this brings us to our second issue: the matter of whether intentional attributions can be predictive in the sciences of the mind when they do not correlate with any particular sub-personal mechanisms is ultimately an empirical one. And there is empirical evidence that such accounts genuinely are predictive in these contexts. This means that the predictive value of intentional models is not contingent on them being abstract structural descriptions of systems. In the following chapter, I provide clear evidence of this by looking at the Theory of Planned Behaviour in psychology. This will provide us with clues as to the sort of predictive tool intentional models can be, and how they fit into our methodological study of the mind.

Chapter 5
Prediction Without Mechanistic Correspondence

Do intentional descriptions allow us to generate accurate predictions in the sciences of the mind when there is no known correspondence between posited intentional states and sub-personal mechanisms working in the brain? In this chapter, I demonstrate that they can, and I provide a clear empirical example to demonstrate this. However, before providing this example, it is important to stress that many of the neurological mechanisms responsible for behaviour are still largely unknown. And so it is difficult to say with absolute certainty which intentional attributions will or will not eventually end up corresponding to particular mechanisms. I therefore begin with some suggestions regarding what kinds of intentional states are likely (given what we currently know) not to correlate with specific neurological mechanisms.

5.1 Traditional Psychological Concepts

The sorts of intentional states that have been considered most controversial in terms of their correspondence to physical mechanisms have been those traditionally associated with "folk" psychology. These include beliefs, intentions, desires, hopes, fears, and regrets (among others). In Section 1.1, I highlighted the problem with associating such concepts exclusively with folk psychology[15], but for our current purposes what matters is not whether they are folk concepts. Instead, what matters is whether such concepts correspond to the functioning of particular physical mechanisms, since it is the predictive success of non-mechanistic intentional descriptions that we care about (regardless of whether they are folk or not). To avoid associating such concepts exclusively with "folk" psychology, I will instead refer to such concepts as belonging to *traditional* psychology. In this sense, concepts like "beliefs" and "intentions" are traditional psychological concepts.

[15] Concepts are not folk or scientific by themselves. The *context in which they are used* makes them folk or scientific. And concepts like "beliefs", "intentions", and "desires" are used in both folk, and scientific, contexts.

So do concepts like "beliefs" and "intentions" abstractly describe the functioning of particular neurological mechanisms? What would it mean for some physical mechanism to act as a belief or an intention, exactly? Answering these questions requires a better understanding of what terms like "beliefs" and "intentions" mean. And following Paul Churchland, I propose that the semantics of these terms are ultimately determined by the network of laws and generalizations in which they fit. As Churchland puts it, the "semantics of the terms in our familiar mentalistic vocabulary is to be understood in the same manner as the semantics of theoretical terms generally: the meaning of any theoretical term is fixed or constituted by the network of laws in which it figures" (1981, p. 61). But what are the laws and generalities that determine the meanings of these mental state terms? Ultimately, these generalities can be understood as the expected interactions that go on between these mental states to produce behaviour. As Braddon-Mitchell and Jackson put it:

> The fact that people tend to move in such a way that what they desire is satisfied if what they believe is true is more than an interesting truth. It is in part constitutive of our understanding of belief and desire. (2007, p. 53)

The mechanistic legitimacy of beliefs, intentions, and desires is therefore dependent on whether we can find mechanisms within the human brain or body that behave and interact in a way that mimics these generalizations.

There are multiple reasons for thinking that no such mechanisms exist. First, attempts to find physical mechanisms within the brain that behave appropriately to warrant being called "beliefs" or "intentions" have not been successful to date (it is for this reason that Churchland purposes that we eliminate these terms from our scientific usage).[16] Second, we use such psychological concepts to

[16] Although it should be noted that this could be in the process of changing. Recent neurological research suggests that, in particular contexts, certain propositional attitudes from traditional psychology really *might* find some sort of mechanistic grounding (Harris, Sheth & Cohen 2008; Andersen & Cui 2009). Harris, Sheth & Cohen (2008), for

describe the behaviour of everything from fish, to birds, to reptiles, to insects and spiders, to clams, and to computers (Dennett 1987, p. 22). Given that these systems have radically different internal physical mechanisms generating their behaviour, the assumption that they all have explicit mechanisms that meet the generalities specified by traditional psychology is highly questionable. Third, we used such psychological concepts long before we knew anything about the internal mechanisms generating behaviour. And so the insistence that these attributions happened to, unbeknownst to us, always correspond to some physical mechanism is implausible.

Of course, it should be noted that these reasons are not enough to *rule out* the possibility that there are mechanisms that correspond to such mental states. In fact, some have suggested that the explanatory and predictive successes of such psychological concepts provide us with evidence that they *do* correspond to physical mechanisms. Fodor, for instance, goes this route (Fodor 1987). For Fodor, these are not neurological mechanisms, however, but higher-level cognitive mechanisms that are merely being implemented by the lower-level neurological machinery. The human brain, under this interpretation, works as a sort of Turing Machine. And as is the case with Turing Machines, it can be implemented by different sorts of physical systems. This explains why radically different mechanistic systems can have beliefs and intentions (as well as other traditional psychological states). According to Fodor, this sort of account is not only plausible, but is the "only game in town" when it comes to explaining behaviour (1975). This would suggest that such concepts really *do* correspond to the behavior of physical mechanisms.

---

example, propose that in some situations, the having of a belief *can* be understood mechanistically. Despite appearing to contradict my claim above however, this research does not mechanistically vindicate the sorts of cases relevant for our purposes here. In other words, their research does not mechanistically ground the idea that we have belief-states interacting with intention-states (as well as other propositional attitudes) in order to generate behaviour in the way proposed by most traditional psychological models. Instead, their research only demonstrates that there may be a neurological mechanism by which people judge a given proposition to be true or false when it is presented to them. Therefore, their research (as it currently stands) does not provide a mechanistic grounding for the predictive power of the psychological models discussed in this chapter.

The problem with Fodor's account is that it runs into empirical problems (these are distinctly different from the ones discussed last chapter regarding the metaphysical nature of intentionality). First, ever since Fodor developed his account in the 1970s, many new games have come to town. Among them: connectionism, and the more recent neural engineering framework, which do not posit a correlation between traditional psychological states and mechanisms. And so even if Fodor could claim victory in the past by running unopposed, he cannot make such a claim anymore. Second, Fodor himself admits that we do not have any story about how the brain implements this higher-level cognitive mechanism. As he puts it, there is an "utter lack of neuroscience of Mentalese" (2008, p. 79). Worse still, he admits that we have no principled means of finding out how this system is implemented:

> I've heard it offered, as an argument against LOT, that no one so far has ever seen a neural token of an expression in Mentalese. But given that we have no idea how Mentalese (or anything much else) is implemented in the brain, how would one know if one did? (Fodor 2008, p. 79 — footnote)

And so we have no real way to confirm that this sort of account is true. Third, and most importantly, we have considerable evidence that the brain does not implement this sort of high-level cognitive mechanism that outputs traditionally defined psychological states in order to generate behaviour (see, for instance, Churchland 1980; Dennett 1987; Eliasmith & Anderson 2003; Craver 2007; and Bechtel 2008 to name only a few). As Dennett puts it:

> [Such cognitive models] seem to lead quite systematically down recognizable dead ends: hopelessly brittle, inefficient, and unversatile monstrosities of engineering that would scarcely guide an insect through life unscathed. (1987, p. 229)

In fact, numerous philosophers have explicitly argued that the prospect of ever finding mechanistic counterparts for the classically defined psychological states (whether as part of a higher-level cognitive mechanism or not) is simply not likely to be vindicated by our best neurological research (Churchland 1981; Stich 1983; Bickle 2003).

Of course, given that it is impossible to say with absolute certainty which intentional states will ultimately be found to correspond to physical mechanisms once we have developed a completed neuroscience, the best we can do is make an educated guess based on our current evidence. And given our current state of information, many of the propositional attitudes used by traditional psychology are good candidates for intentional states that appear to have no mechanistic counterparts. And so the question is: can the application of these intentional concepts be predictive of human behaviour? If so, it would mean that an intentional description, which is not merely an abstract mechanistic description, can be genuinely predictive. And it is to this point that I will now turn.

5.2 Reasons to Doubt the Predictive Value of Traditional Psychological Models

Before examining the empirical question of whether a model employing traditional psychological concepts can be predictive, let us first consider some reasons to be *sceptical* that it would have predictive power. For brevity, I refer to such a model from here on as a "traditional psychology model", or a "TPM". I use this term not to refer to models that only work exclusively with traditional psychology concepts, but instead to refer to models that employ *at least some* concepts that are found in traditional psychology. Given that any predictive model which meets this minimal requirement would still be making predictions based on the application of intentional concepts which do not abstractly describe mechanisms, it is irrelevant for our purposes whether a TPM is substantially more complex than something like a folk psychological model. Let us now consider some potential worries facing the predictive success of TPMs.

First, if the behaviour of a system is the direct result of the collection of mechanisms that constitute it, then a TPM, which ignores these mechanisms, does not appear to have a solid foundation on

which to base predictions. Such a model would not take into account the actual features of the system that generate its behaviour, and so has no basis from which to generate accurate predictions.

Second, we often seem to use traditional psychological descriptions merely as a means of *re-describing* our past actions in order to justify them. In this respect, such concepts are used as a *post hoc* justification of behaviour instead of a predictive device. As Elliot Ludvig puts it:

> …an explanation in terms of "wanting" and "knowing" is entirely post-hoc and has no predictive power whatsoever. How do we know what someone "wants" except by inferring it from what they do? We can ask the person for an introspective report, but […] first-person accounts of behavior are notoriously unreliable. (2003, p. 141)

Third, given that everyday talk in terms of mental states like beliefs and intentions is ubiquitous, it is easy to *assume* that we are using a TPM as a means of predicting systems. However, the neurological mechanism by which we make predictions of other people may in no way depend on the attribution of intentional states like beliefs and intentions to them. In other words, the application of a TPM may not be the means by which we make predictions of others; we simply assume it is given our tendency to make reference to such intentional states in everyday talk. Christopher Gauker (2009) highlights this problem explicitly:

> My own opinion is that there is no reason to believe [that people can successfully predict one another's behavior on the basis of attributions of belief and desire]. Yes, we can often successfully predict what other people will do. Sometimes we do it by straight induction (People look both ways before crossing a street). Sometimes we can predict what a person who has a certain skill will do as a consequence of having that skill (She's a good chess player; so she will take my rook). One of those skills is language. We can often predict what people will do on the basis of what they have told us (He will meet me at my office at 10 tomorrow, because that's the time we agreed on).

> What I don't see is any evidence that we can reliably predict what people will do in a way in which attributions of belief and desire play an ineliminable role.

In a similar argument, Adam Morton (1996) argues that a TPM is simply far too complex and unwieldy to successfully use to predict the behaviour of others. This is due to two reasons. The first reason is that predicting the behaviour of others requires more than just understanding them and their interactions in their environment. It requires understanding how their goals are intertwined with the goals of others, and this quickly makes predictions too difficult:

> In quite simple situations one can need knowledge of knowledge of knowledge of preferences: knowledge cubed. It is not hard to find quite simple 2-person n-act situations in which each person needs knowledge to the *n*th. These situations are not artificial ones: their structure is that of situations that people face every day of their lives.
>
> It's very hard to think in the face of such complexity. In particular, it is hard to see what step by step procedures will take you from the given data to a sensible decision about what to do in the light of what the other(s) may do. (Morton 1996, p. 130)

The second reason is that any set of intentional states we use to predict a system may yield different results depending on what other intentional states we attribute to the system. And it is extremely difficult to figure out which additional intentional states ought to be ruled out of our predictive account:

> The conclusion is that in order to predict actions you have to know more than a finite list of beliefs and desires. You have to know general features of the person's motivations that allow you to exclude whole classes of states that would be inconsistent with the beliefs and desires you attribute leading to the predicted action. (Morton 1996, pp. 128-129)

These together, according to Morton, give us good reason to deny that applying a TPM is a feasible means of predicting behaviour.

Lastly, TPMs attempt to predict people based on properties or states we ascribe to them (i.e., their having some set of intentional states). However, Ross & Nisbett (1991) argue that the environmental context in which a person is situated is a far better predictive indicator of behaviour than the individual traits that we ascribe to people. This gives us at least some reason to question the predictive success of TPMs. With all of these worries in mind, let us now examine an empirical case of a TPM.

5.3 Theory of Planned Behaviour

As an example of a TPM, let us consider the Theory of Planned Behaviour (Ajzen 1985, 1988, 1991). Before getting into the empirical details, let me stress a few preliminary points. First, I am not claiming that the Theory of Planned Behavior is the only successful TPM that exists in psychology. Second, I am not claiming that the theory of planned behaviour is a model of "folk" psychology (i.e., it is not the way in which we attribute mental states to others in folk contexts, abstracted away from scientific methodology and theory). The question of whether folk psychology is predictive is, for our current purposes, irrelevant. My intention is only to demonstrate that a model, which posits intentional states that we have good reason to think do not conform to the behaviour of neurological mechanisms, can be predictive. I provide here only a brief sketch of the theory of planned behaviour, since many of the details are not relevant for the general point I wish to make.

In order to predict behaviour, the theory of planned behaviour employs a psychological model that contains two major components: The agent's *intention* to behave, and their *perceived behavioural control*. According to Ajzen, "perceived behavioral control, together with behavioral intention, can be used directly to predict behavioral achievements" (1991, p. 184). To understand how, let us consider in more detail the two components of the model. Let us begin with intention:

> ...a central factor in the theory of planned behavior is the individual's *intention* to perform a given behavior. Intentions are assumed to capture the motivational factors that influence a behavior; they are indications of how hard people are willing to try, of how much of an effort they are planning to exert, in order to perform the behavior. As a general rule, the stronger the intention to engage in behavior, the more likely should be its performance. (Ajzen 1991, p. 181)

What is it that determines the strength of an intention? According to the theory, the strength of an intention is determined by two different types of *beliefs*. The first are *behavioural beliefs*, which determine the agent's attitude towards the behaviour. The more favourably the agent views the behaviour, the stronger the intention. The second are *normative beliefs*, which determine the agent's subjective norms (Ajzen 1991, p. 189). These norms reflect the social pressures placed on the agent to perform, or not perform, the act. The greater the social pressure to perform the act, the stronger the intention.

Let us now turn our attention to the second component of the theory of planned behaviour: *perceived behavioural control* (PBC). PBC is an agent's perception of their own ability to perform a given action, and is determined by the agent's beliefs regarding their level of personal control over completing the task (Ajzen calls this third type of belief *control beliefs*). Put simply, perceived behavioural control can be understood as "people's perception of the ease or difficulty of performing the behavior of interest" (1991, p. 183). While a complete lack of *actual* behavioural control (e.g., a lack of ability, resources, or opportunity to perform the task) will obviously bar the agent from completing a task, a partial decline in actual behavioural control can be overcome by the agent given the appropriate PBC. In other words, "perceived behavioral control can often be used as a substitute for a measure of actual control" (Ajzen 1991, p. 184). For instance, a loss in behavioural control (a lack of opportunities) can be compensated for if the agent expends more effort to find and/or create new opportunities. But an agent will only make such efforts if they think that they have a chance at succeeding. In this regard, an agent who has a higher perception of behavioural control will be more likely to perform a task they only have

partial behavioural control over than one which has a lower perception of behavioural control. The agent's willingness to try, or to exert effort to overcome obstacles, is dependent on their perception of their own ability to succeed at such endeavours.

This, of course, assumes that one's perceived behavioural control is not grossly out of alignment with one's actual behavioural control. It also means that certain kinds of behaviour (those whose performances are within the actual power of the agent) will be more predictable by the theory than behaviour which is outside their volitional control.

5.4 Is the Theory of Planned Behaviour Predictive?

According to Ajzen, the theory of planned behaviour "permits prediction and understanding of particular behaviors in specified contexts" (1991, p. 206). This conclusion has been supported by numerous empirical studies that have examined the predictive merits of the model (see: Ajzen & Driver 1992; Sparks 1994; Blue 1995; Connor & Sparks 1996; Godin & Kok 1996; Hausenblas et al. 1997; Armitage & Conner 2001). Hausenblas et al., for instance, claim that the theory of planned behaviour has "considerable utility in predicting and explaining exercise behavior" (1999, p. 46). Similarly, Armitage & Conner, in their meta-analysis of the efficacy of the theory of planned behaviour, conclude that the studies they examined "provide support for the efficacy of the TPB [theory of planned behaviour] as a predictor of intentions and behaviour" (2001, p. 489).

It is important to note that these claims of predictive success are dependent on certain conditions being met by the model regarding its proper application, and the contexts in which it is applied. According to Ajzen, there are three major conditions that must be met in order for the theory to predict behaviour:

> First, the measures of intention and of perceived behavioral control must correspond to or be compatible with the behavior that is to be predicted. That is, intentions and perceptions of control

> must be assessed in relation to the particular behavior of interest, and the specified context must be the same as that in which the behavior is to occur. For example, if the behavior to be predicted is "donating money to the Red Cross," then we must assess intentions "to donate money to the Red Cross" (not intentions "to donate money" in general nor intentions "to help the Red Cross"), as well as perceived control over "donating money to the Red Cross." The second condition for accurate behavioral prediction is that intentions and perceived behavioral control must remain stable in the interval between their assessment and observation of the behavior. Intervening events may produce changes in the intentions or in the perceptions of behavioral control, with the effect that the original measures of these variables no longer permit accurate prediction of behavior. The third requirement for predictive validity has to do with the accuracy of perceived behavioural control. As noted earlier, prediction of behavior from perceived behavioral control should improve to the extent that perceptions of behavioural control realistically reflect accurate control. (Ajzen 1991, p. 185)

The predictive success of the theory therefore depends on the sorts of behaviours one is modeling. Some behaviours are predicted with a high degree of success, such as choice in leisure activities (Aizen & Driver 1992) and voting choice (Watters 1989; Ajzen 1991, p. 187), with a documented success rate of .78 and .84 respectively. Meanwhile, other behaviours are not very well predicted by the theory, like losing weight (Ajzen 1991, p. 187), or getting an "A" in a course (Aizen & Madden 1986).[17] Then there are the behaviours that fall somewhere in between; where the predictive success of the theory is not very strong, but neither is it insignificant (Bozionelos & Bennett 1999). What does all this say about the model's predictiveness then? After all, it cannot predict huge swaths of human behaviour. So are these limited and contextual successes of the model sufficient to consider it genuinely predictive, or not? Ultimately, speaking of a model as though it is predictive *tout court* is misleading.

[17] Ajzen suggests that these sorts of behaviours do not meet the above conditions for proper application of the model (Ajzen 1991, p. 187). Students, for instance, may not realize what is involved in getting an "A" in a course they have never taken, and so may mis-represent the ease of getting an "A" (leading to a disconnect between perceived behavioural control and *actual* behavioural control). Similarly, getting an "A" may not be within their volitional control depending on the course.

To demonstrate why, consider the question: under what conditions is a model genuinely predictive? Is a genuinely predictive model only one that can predict *all* the behaviour produced by a given system? While the theory of planned behaviour is predictive of some kinds of behaviours, it fails to predict other kinds of behaviour, and is only moderately successful at predicting others. So is the limited success of the model a reason not to consider it genuinely predictive? This seems like an unrealistic standard to hold our models to. After all, we have no models in science that are predictive of all human behaviour. What, then, does this say about the models we use to make predictions when studying human behaviour in physiology and behavioural neuroscience? Does this imply that neuroscience and physiology have no predictive models at all given that neither domain can predict all of human behaviour? No. It simply means that each domain studies and predicts different *aspects* of human behaviour.

Similarly, the predictiveness of a model can often depend on how course-grained, or fine-grained, our description of the action being predicted is. I might create a model of cat behaviour that predicts (extremely well) that my cat will try to escape from the bathtub when given a bath, but which does not predict exactly how the cat will flail its limbs to escape, or in which direction it will flee. It may be extremely predictive of behaviour at one level of description, but not another. Does this mean that the model is predictive, or not? The only appropriate answer is: it depends. It depends on what behaviour one is interested in, and how detailed our description of the behaviour needs to be for our purposes. After all, a model can be extremely predictive of a small range of behaviours at a particular level of detail, and thus play an essential scientific role so long as we are working at that level and within that range.

But even if we grant all this, shouldn't a genuinely predictive model at least be able to predict this small range of behaviour consistently? If, for instance, we claim that a model is successful at predicting voter choice, does this imply that the model is always (or almost always) successful at predicting voter choice under any circumstance? This too seems like an unrealistic standard for us to hold our model to. We have no model that can account for the particular behaviour of a system in every possible context.

As an analogy, consider our use of mechanistic models to provide explanations. While mechanistic models are considered the dominant method for providing explanations in the life sciences (see Section 2.2 for details), a particular mechanistic model is only ever useful within particular contexts. As Bechtel points out:

> The behaviour an entity [e.g. a component part of a mechanism] exhibits is often dependent upon context and there is no reason to think that the account of an entity offered by any inquiry considers how it will behave under all conditions but only those which are the focus of inquiry. As engineers are well aware, how a component will behave when inserted into a particular kind of system often needs to be investigated empirically. (Bechtel, 2008, p. 22)

A mechanistic model of a system is always an account of that system given a certain context. When we change the context, it can change how the parts of the system interact with one another. In this regard, a given mechanistic model cannot be used to predict or explain the behaviour of that system in every context. A similar point is made by Craver:

> For example, one might provide a model of verb-tense generation that performs perfectly well when the brain and vocal cords are working properly, but that provides no insight into how the system will behave if something breaks or if the system is in extreme environmental conditions. (Craver, 2006, p. 357)

Thus it seems that if our standard for a good scientific model is context-independent success, then this is a standard that even our best explanatory models in the life sciences cannot satisfy. And so again this seems to put the bar too high. Otherwise, we would have *no* explanations whatsoever in the life sciences.

With this in mind, I propose that to be relevantly predictive for scientific purposes, a model must be able to predict (at a rate better than chance) the production of *certain kinds* of behaviour produced by

certain kinds of systems given *certain kinds* of contexts and constraints. And empirical studies of the theory of planned behaviour demonstrate that the model is indeed predictive of "behaviors in specified contexts" (Ajzen 1991, p. 206). Crucially, the theory's use of "intentions" and "beliefs" to form predictions demonstrates that it does not predict by characterizing the abstract functioning of particular neurological and physiological mechanisms operating within the system.[18] This provides direct empirical evidence for the fact that predictively successful intentional models do not make predictions in virtue of abstractly characterizing particular sub-personal mechanisms that generate behaviour.

5.5 Responses to Sceptical Worries

With this established, let us now re-visit the worries we had regarding the predictive success of TPMs. The first worry was that TPMs have no firm foundation on which to generate accurate predictions *without* characterizing the causal mechanisms within the system. But this assumes that the only sorts of scientific models that are predictive of mechanistic systems are ones that characterize causal mechanisms. And we know that this is untrue. There are many examples in science of models that are predictive, but which make predictions without any appeal to causal mechanisms. Statistical models and dynamical models are clear examples of this (Eliasmith 2010). And so we cannot assume that certain psychological models would fail to be predictive simply in virtue of not correctly characterizing the causal mechanisms of systems.

[18] It should be noted that since the development of the theory of planned behaviour, a number of criticisms have been raised against some of the model's reported predictive successes. Armitage & Connor (2001) argue that evaluations of the theory's predictiveness may have been too limited in scope and sampling (p. 475). Similarly, that a "tendency for authors to report only significant findings may have inflated the reported values" (p. 475). Despite this, they ultimately conclude that "in spite of these weaknesses, evidence from narrative and meta-analytic reviews suggest that the TPB [Theory of Planned Behaviour] is a useful model for predicting a wide range of behaviours and behavioural intentions" (p. 475).

Similarly, others have argued that certain components of the model may be problematic. Sparks et al. (1995), for instance, argue that "subjective norms" may not be genuinely helpful in predicting intentions. A similar criticism of subjective norms motivated Trafimow & Finlay (1996) to suggest that the model be modified to better account for social pressures. These types of criticisms tend to propose alterations to the model that make it more predictive. Interestingly, none of these alterations suggest the removal of "beliefs" or "intentions" as the foundation

The second worry was that intentional attributions like "beliefs" and "intentions" often seem to be *ad hoc*. We do not use them to predict future behaviour, but only to re-describe past behaviour. There is indeed some truth to this. We often do attribute intentional states to people based on what we observe of their past behaviour. The mistake is to then assume that we cannot, or do not, use these attributions to then predict future behaviour.

In order to generate a predictive model, we must first observe regularities or patterns in the past behaviour of the system. We then generate a model which attempts to capture, or characterize, these regularities. Once this model is constructed, we use it to try to predict how the system will behave in similar contexts in the future given the patterns observed. This is true of virtually all predictive models. Suppose, for instance, that I wanted to predict the behaviour of a system using a statistical model. In order to do so, I would not simply generate *random* statistical models and then see if any of them applied. I would construct the model based on my observations of the system's previous behaviours in similar contexts. According to Eliasmith, statistical models "focus on describing the regularities in the data" (2010, p. 315) that we collect about systems, which are then used to help predict future behaviours. But, of course, we cannot find the regularities in the data without first *gathering* this data. And we can only gather this data from observing past behavior.

And so the fact that we often do not know what someone believes, or wants, or intends unless we infer it from their past behaviour does not mean that such attributions are not predictive. It simply means that this sort of after-the-fact attribution is the first necessary step in constructing a predictive model. And this is true of all predictive models.

The third worry regarding the predictive success of TPMs was offered by Chris Gauker. He argues that the way in which we predict the behaviour of others need not in any way make reference to classic psychological states. The fact that traditional psychological idioms are so commonplace means it

---

of the model. And so such criticisms, if true, still leave us with a demonstrably predictive model that relies on

is easy to mistakenly think that they are relevant in predictions. The neurological mechanisms in our head that allow us to predict the behaviour of others likely do so without the attribution of psychological states like beliefs and intentions. So their place within our predictive accounts is, at least in principle, eliminable.

Implicit in this sort of argument are two major criticisms. First, it assumes that TPMs are likely not predictive because they do not resemble the method by which we neurologically predict the behaviour of others in day-to-day life.[19] Morton (1996) seems to argue for a similar position, claiming that TPMs are simply far too complex for us to use in daily life to make predictions. Second, it assumes that TPMs may not be useful predictive devices in science because they have not been shown to be indispensable.[20] I propose that both of these implicit arguments are irrelevant.

In regards to the first criticism, consider our use of scientific theories to make predictions. Imagine, for instance, that we are trying to predict the outcome of an experiment in quantum physics (more specifically, we are trying to predict what one of our instruments will read at the end of the experiment). The neurological mechanisms in our brain that are used to make predictions of physical objects in our environment (like scientific instruments) do not make these predictions by calculating quantum mechanical theories. Our brain does not, for instance, posit the interactions of subatomic particles to make predictions of anything in its environment. But from this we do not assume that the theories of quantum mechanics therefore cannot be used to predict the results of our scientific experiments. Similarly, the theories of quantum mechanics surely fall victim to the same criticisms that Morton levelled against TPMs: that such models are too difficult for our brain to apply in daily contexts

---

intentional states that have no mechanistic counterparts. Thus they do not damage the essential point of this chapter.

[19] Which can be seen in Gauker's claim: "The idea is that person A can observe person B's behavior, on that basis figure out that B has certain beliefs and desires, and on the basis of the attribution of those beliefs and desires, successfully predict what B will do. My own opinion is that there is no reason to believe this." (2009)

[20] As evidence, consider Gauker's claim that: "What I don't see is any evidence that we can reliably predict what people will do in a way in which attributions of belief and desire [and other classical psychological concepts] play an ineliminable role." (2009) And similarly: "Nor even have I ever heard of a single real-life example in which it

to predict systems. But that hardly means that such accounts cannot be used to make effective, and possibly essential, scientific predictions. The invention of natural language has provided us with a powerful resource for making predictions and solving problems in a way that we could not do with internal neurological resources alone:

> Indeed, it may be that the intellectual explosion in recent evolutionary time is due as much to this linguistically-enabled extension of cognition as to any independent development of in our inner cognitive resources. (Clark & Chalmers 1998, p. 18)

And so the fact that intentional attributions may not be the way by which we *neurologically* predict others in day-to-day life does not mean that such attributions cannot be predictive in scientific contexts.

In regards to the second criticism (i.e., the dispensability charge), science does not use models or tools only on the condition that they all prove to be indispensible. Similarly, what counts as being indispensable is a murky issue. Consider once again our use of statistical models to make predictions in the life sciences. Are statistical models truly indispensible to the life sciences? According to Eliasmith, statistical models are ideal for making predictions when we do not know or understand the causal mechanisms underlying the behaviour of a system (2010, p. 315). But suppose that a full understanding of these mechanisms would allow us to make predictions of the system by other means? If that were the case, would statistical models be indispensible, or not? Since we could always, in principle, use another model to make the same predictions when we understand the mechanisms of the system, then in some sense they are not indispensible (we do not *need* them to make predictions). However, the very benefit of statistical models is that they allow us to predict when we *do not know* the mechanisms underlying the

---

was at least quite plausible that one person successfully predicted the behavior of another [and] it was not evident that the same prediction could have been made in other ways." (2009)

system. And so these models may be essential tools in situations where our knowledge of the system is limited. The same lesson ought to be applied to TPMs. We often need to make predictions of systems when we do not have the luxury of quantifying over relevant features of the system necessary to generate models like statistical models or mechanistic models. In this regard, TPMs will be ideal in helping to see patterns in behaviour that may be required in order to help generate statistical or mechanistic models (see Section 6.2.6 for more details). This, I propose, is more than enough to validate the use of such models in science. Ultimately, the indispensability of TPMs is irrelevant to the question of whether the model is predictive, or whether it is useful in scientific contexts.

What about Morton's criticism that attributing a set of intentional states to a system is always compatible with an indeterminate number of future behaviours depending on what other intentional states the agent has? The problem with this criticism is that this is essentially just a charge of under-determination. But under-determination is problem for all predictive models. For instance, the same statistical model can be used to predict conflicting behaviours of the same system *depending on what additional information we incorporate into the model*. And in any given situation, we can never be absolutely certain that such additional information is not relevant. But this does not *ipso facto* mean that statistical models are therefore never successfully used to make predictions in scientific contexts.

This sort of criticism misconstrues the way in which predictive models are generated. As mentioned in section 5.3, predictive models are developed by observing regularities in the behaviour of the system, and then generating a model that allows us to characterize those regularities. Morton assumes that our reasons for attributing intentional states to systems is completely divorced from our observation of how that system has acted in similar contexts in the past. But this is often not the case when it comes to predictive TPMs. To demonstrate, recall the first condition on proper application for the theory of planned behaviour:

> First, the measures of intention and of perceived behavioral control must correspond to or be compatible with the behavior that is to be predicted. That is, intentions and perceptions of control must be assessed in relation to the particular behavior of interest, **and the specified context must be the same as that in which the behavior is to occur**. For example, if the behavior to be predicted is "donating money to the Red Cross," then we must assess intentions "to donate money to the Red Cross" (not intentions "to donate money" in general nor intentions "to help the Red Cross"), as well as perceived control over "donating money to the Red Cross." (Ajzen 1991, p. 185, my emphasis)

In order for the model to predict future behaviour, we must first have a sense of how the person will behave in a given situation or context. We then use this as a guideline for determining future behaviours.

Imagine we have two different TPMs that we can use to predict the behaviour of a system. One reliably fits with the pattern of behaviour displayed by the system in similar contexts in the past. The other, despite positing many of the same intentional states, does not. In this situation, we have clear pragmatic grounds for choosing one model over another, even if it does not overcome the under-determination problem (i.e., even though we can never rule out the possibility that the present situation will have unknown features that may interfere with our predictions). Put simply: Morton's objection highlights a problem for predictive models *in general*, and not for TPMs in particular.

Lastly, consider the objection that TPMs ignore relevant environmental conditions by making predictions based only on ascribed traits or properties of the system. The problem with this argument is that it assumes that a TPM will never, in principle, take environmental conditions into account. And this is simply false. Consider the theory of planned behaviour. Ajzen tells us that "perceived behavioral control can, and usually does, vary across situations and actions" (1991, p. 183). In other words, the attributions of beliefs, intentions, and perceived behavioural control, are all context sensitive. Instead of ignoring environmental conditions and focusing only on traits of the system to make predictions, the

theory of planned behaviour bases its attributions of intentional states on the context in which the system is placed. Different contexts change the set of beliefs the agent has, which in turn changes the perceived behavioural control, and the intention. As Armitage & Connor put it,

> ...in situations where (for example) attitudes are strong, or where normative influences are powerful, PBC may be less predictive of intentions. Thus, Ajzen (1991) argues that the magnitude of the PBC-intention relationship is dependent upon the type of behaviour and the nature of the situation. (2001, p. 472)

Put simply, a TPM can be constructed to take into account relevant environmental conditions. In fact, Ross & Nisbett even acknowledge that we are not oblivious to relevant environmental causes even when we are in folk contexts: "Such considerations are fairly obvious once they are mentioned, and the layperson, upon reflection, will generally concede their importance" (1991, p. 3).

So the suggestion that intentional attributions are only predictive when they abstractly describe the functioning of particular physical mechanisms has now been shown to be false. This means that intentional attributions can be genuinely predictive in scientific contexts, and that this is true irrespective of whether they correlate with the functioning of particular physical mechanisms or not.

So far, this position, which I have taken some time to argue for, appears to share much in common with Dennett's theory of the "Intentional Stance". And so it appears that I have taken a rather circuitous route to reach a conclusion that has already been vigorously defended. So what was the point? I've taken a circuitous route to a similar conclusion for an important reason: the route I've taken informs the sort of story I believe we ought to tell about intentional language. The analogy between statistical models and intentional models that I have hinted at in this chapter is not an accident. There are key similarities between the two types of models that will help to illuminate the role and benefits of intentional attributions in science. And the details of this story will help to demonstrate what Dennett gets

wrong (see Section 9.5 for details). First, however, a greater understanding of the similarities between intentional and statistical models is in order.

Chapter 6
Similarities Between Statistical Models and Intentional Models

We know that intentional models can predict mechanistic systems without appealing to the underlying mechanisms that generate their behaviour. This ability, however, is not unique to intentional models. As I mentioned briefly last chapter, statistical models have this ability as well. As a result, it is time to put the similarities between these two models front and centre in our analysis. In this chapter, I argue that the relevant similarities between intentional models and statistical models provide insights into the way intentional models are important to science, and how they fit into the mechanistic explanations of the life sciences.

In order to show this, I begin by choosing an example of a statistical model, and an example of an intentional model, to compare and contrast. Following that, I highlight the relevant similarities that the two models share, and show how each contributes to our study of the mind in virtue of their shared benefits. This will ultimately show that intentional language is valuable to science in virtue of functioning as a type of phenomenological model.

6.1 Examples

6.1.1 Intentional model

For our intentional model, consider the case of Stanley the Volkswagon discussed by Parisien & Thagard (2008). "Stanley" was the name given to the Volkswagon Touareg that won the 2005 DARPA Grand Challenge by its creators at the Stanford University Artificial Intelligence Laboratory. The DARPA Grand Challenge is a contest in which contestants must construct an artificial-intelligence-guided car that can navigate a complex set of obstacles in an environment.

According to Parisien & Thagard, we can interpret the behaviour of Stanley on the DARPA course by attributing to it a set of representations or intentional states. For example, by understanding

Stanley as *identifying* his own location, *inferring* what obstacles lie ahead of him, and then *deciding* (or *instructing himself*) to drive in the right direction and at a manageable speed (Parisien & Thagard 2008, p. 170). Or, more specifically:

> Data from all [Stanley's] sensors were integrated into a *drivability map*, which is a single model of the environment that marks each cell in a two-dimensional map as either unknown, drivable, or undrivable. This information, along with other variables for the general condition of the environment such as terrain slope, are used to set the driving direction and velocity of the vehicle, which in turn control the steering, throttle, and brake. (2008, p. 171)

Understanding Stanley's behavior in terms of his ability to represent terrain as "drivable", "undrivable", and "unknown", as well as represent objects in his path, allows us to understand the sorts of things he will do on the DARPA course.

This intentional model of Stanley offered by Parisien & Thagard also allows for *mis*representations, which is a key feature of intentional attributions. As they note,

> One prominent example comes from the way Stanley uses its laser rangefinders to judge the terrain directly in front of the car. A rotating laser sweeps the ground in an arc several meters ahead, and the rangefinder computes depth information along that arc. As the car moves forward, it pushes the arc like a broom, combining the information from multiple sweeps to create a three-dimensional map. However, this process depends on the car's stability, because when the car pitches forward over a bump, the laser rescans a previous area, and then skips far ahead. This puts the scan lines out of sequence, making the rangefinder perceive a large, impassable obstacle (Thrun et al. 2006). Consequently, Stanley would carry out often dangerous avoidance maneuvers for an obstacle that never existed. (2008, pp. 173-174)

With this example of an intentional model in mind, let us now turn our attention to a statistical model.

6.1.2 Statistical model

For our statistical model, consider what is often called "The Thompson Effect". This is the phenomenon in which objects appear to us to move faster or slower based on how greatly the object contrasts with its surrounding environment. A few years ago, Alan Stocker and Eero Simoncelli (2006) developed a statistical model that can accurately predict how fast an object will appear to be moving to an observer given the level of contrast between the object and its surrounding environment. They did this by altering the modern framework of statistical estimation used to traditionally model perception. This modern framework viewed an optimal observer in terms of two probability distributions:

> First, the variability of a set of measurements, *m*, is specified as a conditional probability distribution, $p(m|v)$, where *v* is the stimulus speed. The variability is due to a combination of external sources (e.g., photon noise) as well as internal sources (e.g., neural response variability). When considered as a function of *v* for a particular measurement, this conditional density is known as a likelihood function. The second component is a prior probability distribution, $p(v)$, which specifies the probability of encountering stimuli moving at any particular retinal speed. According to Bayes' rule, the product of these two components (when appropriately normalized) gives the posterior distribution. Common choices are the mean, or the mode. Biases in the perceived speed of low-contrast moving patterns arise intrinsically with this model, assuming a prior that favors low speeds: Lower contrast stimuli lead to noisier measurements, producing a broader likelihood function, which leads to a lower speed estimate. (Stocker & Simoncelli 2006, p. 578)

The problem with this traditional model is that it is extremely difficult to experimentally determine what the prior distribution and the likelihood function actually are. Similarly, there are constraints on the model that cast doubt on its ability to accurately represent the phenomenon in question (Stocker & Simoncelli 2006, p. 578). To compensate for this, Stocker & Simoncelli embed a Bayesian estimator into the traditional model. This estimator is calibrated based on the trial-by-trial responses of

subjects observed in a forced choice speed discrimination experiment. According to Stocker & Simoncelli:

> We were able to validate the ability of a Bayesian observer model to account for the data, and also to determine the prior distribution and internal noise level associated with the best-fitting Bayesian estimation model. (2006, p. 579)

By embedding the Bayesian estimator into the traditional model of perception, Stocker & Simoncelli were able to generate a predictive model that overcomes the shortcomings of previous statistical accounts of the phenomenon. Most importantly for our purposes, however, is the fact that their model does not provide any mechanistic story for what generates the Thompson Effect, but still does "a good job of predicting the subject's performance under a wide variety of motion estimation tasks" (Eliasmith 2010, p. 319). In this regard, statistical models like Stocker & Simoncelli's model primarily function within scientific practice as a type of phenomenological model.

6.1.3 Statistical models are phenomenological models

Phenomenological models are defined by their ability to characterize, and predict, the behaviour of systems without attempting to decompose them into parts and operations for better understanding. As Craver describes them:

> …all one requires of a [phenomenological] model is that it be *phenomenally adequate*. That is, the input–output mapping in S [e.g., a given algorithm, function, or account, that generates a mapping from inputs to outputs] should be sufficiently similar to the input–output mapping in T [e.g., the observed regularity in the actual input-output of the system] for one's needs. Few models are actually isomorphic with the phenomenon, given that models typically abstract away from the precise details of the system being modeled, that they typically are only approximate, and that they make simplifying assumptions in order to apply a particular formalism. The weaker standard that the input–output mapping in T should be homomorphic with the mapping in S can

> be easier or harder to satisfy depending on how much detail one includes about the target phenomenon and on how similar one expects the model and the phenomenon to be. The richer and more fine-grained one's characterization of the target phenomenon, the more the space of possible models for the phenomenon is constrained, and so the more challenging it is to build a phenomenally adequate model. (Craver 2006, p. 357)

Many consider such models to be largely unexplanatory in the sciences of the mind due to their focus on characterizing and predicting phenomena instead of developing a *mechanistic explanation* of their production (see Section 2.2). Stocker & Simoncelli's model, for example, may be very good at predicting the Thompson Effect, but it says nothing about the underlying mechanisms responsible for it. This is why Craver claims that:

> A model can be richly phenomenally adequate and non-explanatory. This is the take-home lesson of the several decades of attack on covering-law models of explanation at the hands of advocates of causal–mechanical models of explanation: merely subsuming a phenomenon under a set of generalizations or an abstract model does not suffice to explain it. (2006, p.357-359)

As noted in Section 2.1, however, it is still possible for phenomenological models to be explanatory in the behavioural sciences depending on the context, and the particular question being asked. It is for this very reason that some propose we need a more pluralistic understanding of explanation in order to better account for actual scientific practice in the behavioural sciences (see Longino 2006; Chemero & Silberstein 2008). What is important to note for our current purposes, however, is the fact that we use phenomenological models to perform different tasks than those we use mechanistic models for, regardless of whether each can be explanatory in the appropriate circumstances. While mechanistic models are used primarily to characterize the structure of systems, phenomenological models are used to

describe the phenomena produced by mechanisms, measure or calculate crucial quantities, make essential predictions, summarize data, and function as heuristics for designing experiments (Bogen, 2005, p. 401; Craver, 2006, p. 355).

It is also important to note that phenomenological models come in all different shapes and sizes, each useful for predicting or describing different sorts of phenomena. Consider, for instance, the application of dynamic systems theory versus the application of statistical models in the life sciences. First, a brief explanation of the role of dynamic systems theory in cognitive science:

> What came to be called *dynamical systems theory* (DST) enables investigators to visualize the change in the state of a system over time. The simplest case is a plot of the states traversed by a system through time, that is, the system's *trajectory* through *state space*. Each dimension of state space corresponds to one variable of the system, and each point in the space corresponds to one of the possible states of the system. (Bechtel, 2008, p. 187)

By applying the appropriate set of differential equations, we can predict the trajectory of the system through this state space. Dynamical models have been used in cognitive science to describe and predict cognitive phenomena like the production of speech (Port, 2003), and the movement of animals (Kelso, 1995). What is important to note about such models, however, is that they often characterize and predict the behaviour of cognitive systems without appealing to underlying causal mechanisms. Instead, it is "usually only observable behaviour [that] is mapped to the model" (Eliasmith, 2010, p. 319). In this respect, many dynamical models act as phenomenological models, providing no mechanistic explanation for the behaviour of the system. However, while such models are predictive of certain types of cognitive behaviour, they are far worse at predicting others (see Kirsh, 1991; Eliasmith, 1996; Bechtel, 2008, pp. 192-200).

Statistical models, meanwhile, act as different sorts of phenomenological models, useful for making different sorts of predictions. Unlike dynamical models, statistical models predict by describing

"the probability of various measureable states of the system given other known states of the system" (Eliasmith, 2010, p. 315). This allows such models to identify regularities in the data we collect about systems needed to generate certain predictions that cannot be identified with other sorts of models.

Given that different sorts of phenomenological models predict in different ways, one type of model may be more useful than another for predictions given the particular dataset we have available to us at any given time. We might, for instance, lack the relevant information needed to construct a dynamical model of a system, but not a statistical one. Or vice versa.

6.2 Similarities Between Statistical and Intentional Models

With a clear example of an intentional model and a statistical model in hand (and a clear understanding of how statistical models function as phenomenological models), let us now compare and contrast the important features of these models. Their relevant similarities will provide us with insights into the scientific value of intentional models, and their role in scientific practice.

6.2.1 Neither model directly describes the physical mechanisms of a system

Statistical models, being a species of phenomenological model, characterize systems without telling us about their underlying causal mechanisms. In this respect, they are not descriptions of mechanisms, and can often be compatible with multiple physical implementations of the system (so long as those mechanisms produce the same statistical properties). As Eliasmith notes:

> Statistical descriptions are highly implementation independent. Statistical models focus on describing the regularities in the data and hence are silent with respect to the particular physical implementation. In essence, these descriptions would not change if the implementation changes and statistical properties do not. (2010, p. 315)

Intentional models are similarly implementation independent. To talk of a system in terms of its *representations* or its *beliefs* does not tell us what the underlying mechanisms are that are working within

the system. Similarly, such accounts are often compatible with multiple physical implementations of the system (so long as those mechanisms produce the same behavioural regularities characterized by the intentional model).

Consider our intentional model of Stanley. We can attribute to him representations of his environment without having any idea how Stanley is implemented. We can use our intentional model of Stanley to understand that he will avoid rocks of a certain size in his path, and turn right when a dangerous slope is on his left *even if we have no idea what the mechanisms inside Stanley are.* Instead of mechanistically interpreting Stanley's behaviour, we make sense of his behaviour based on the intentional content that the sensors provide him with. Similarly, we can imagine Stanley displaying the same behaviour, yet being implemented in very different ways (with a different engine, set of computers, and sensors).

We obviously understand that there is a mechanistic story to tell regarding what the car and its sensors are made of and how they are structured. But to adopt an intentional interpretation of Stanley is to make sense of the way he will navigate the obstacles on the course given the intentional information he has, and not in terms of the interaction of physical parts that generate his behaviour.

6.2.2 Both models can generate predictions of systems despite a lack of mechanistic data

We use both statistical and intentional models to form predictions of systems whose mechanisms we cannot identify. Stocker & Simoncelli's statistical model can predict the Thompson Effect despite a lack of information regarding the actual mechanisms generating it. In fact, one of the very benefits of statistical models is that they allow us to generate predictions of systems with unknown mechanisms. As Eliasmith puts it:

> The natural physical phenomena for statistical descriptions are complex phenomena with unknown mechanisms. Complex systems, in virtue of their complexity, often have many

> unknown or undescribed interactions between system components. As a result, known initial conditions often map to a wide variety of subsequent states. Statistical models are ideal for describing systems of this kind when prediction is of the utmost importance (e.g. in data analysis). (2010, p. 315)

We often explicitly use statistical models for the purposes of modeling systems that we cannot explain or predict mechanistically. The same is true of intentional models. We often use intentional models to predict the behaviour of systems whose mechanisms we do not understand. As Dennett points out, we can use intentional models to predict the behaviour of other people despite "knowing next to nothing about what actually happens inside people's skulls" (1987, p. 48).[21]

Suppose that a group of engineers were to stumble across the 2005 DARPA Challenge by accident, and observed Stanley navigate the challenges of the course. They are determined to understand how Stanley works, but cannot get a hold of him in order to reverse-engineer him. And so instead, they begin by trying to characterize for his behavior and his interactions in his environment. To do this, they generate an intentional model of Stanley. By witnessing the sorts of discriminations Stanley can make, and the way he navigates his environment, they determine that Stanley can represent objects of a certain size, and can identify certain types of terrain as being drivable or not. Similarly, they attribute decisions to Stanley ("he decides not to drive left because that part of the terrain is impassable") to account for his behavior in light of these representations. In this manner, they can predict that Stanley will likely turn left when he approaches the rock ahead of him given that he can represent the rock, and represent the fact that the area to the right of the rock is impassable (having observed Stanley make a similar discrimination in similar environmental conditions earlier in the Challenge).

Of course, this intentional account tells our engineers nothing about what mechanisms exactly are operating within Stanley. The intentional account by itself tells us nothing about what sorts of computers

[21] As the theory of planned behaviour discussed last chapter demonstrates.

are hooked up to Stanley, what sort of engine Stanley has, or how the sensors on the car work. In this way, the intentional model is just like a statistical model.

Similarly, just as with statistical models, the more fine-grained our account of the phenomena being modeled (in this case: Stanley's behaviour on the DARPA course), the more constraints it places on the predictive model we employ. So, for instance, on a first pass of predicting Stanley's behavior, we might attribute to him representations about "large rocks" in his path. Yet, the more we watch Stanley behave in his environment, the more we see that Stanley treats rocks the same way he does any other large object in his path. Thus we refine our intentional model, and attribute to him representations of "large objects" to better account for his behaviour in a wider range of situations. Or imagine the reverse case. Imagine we begin by attributing only the representation "large object" to Stanley, but then observe that he treats rocks in his path different from bales of hay. Given that bales of hay can be driven through without incident, Stanley does not try to avoid them. In that case, we might refine our intentional model to account for both representations of "rocks" and representations of "bales of hay". The lesson is that both statistical and intentional models can be used to predictively model systems when we cannot identify causal mechanisms, and that the more details we have of the system's behaviour in various contexts, the more predictive both models can become.

6.2.3 Both models can be generated from a detailed enough mechanistic model

If we know what the mechanisms are that underlie a system's behaviour, then we can use this information to help generate both a statistical model, and an intentional one.[22] Consider that statistical models make predictions based on the measureable probabilities of the various states of the system. If we can

[22] It should be noted that I am not claiming that an intentional model can always be generated from *any* mechanistic model. Not *all* mechanistic systems will be usefully modelled intentionally. However, if we have a system that can be modelled effectively with an intentional model, then knowing the mechanisms of that system will help us to generate an intentional model of it.

understand how the relevant mechanisms of the system operate in various contexts, then we can use this information to determine what the probabilities of the various states of the system are likely to be.

Similarly, the more we understand the mechanistic underpinnings of a system, the better we can understand the sorts of discriminations the system can make, and under what conditions. This information makes it substantially easier to determine what sorts of content attributions will yield predictions. As a clear example of this, consider the way in which Parisien & Thagard intentionally model Stanley:

> After a brief review of the hardware that Stanley used to interact with its environment, we discuss the software that enabled it to identify relevant features of the world and to plan an effective course using dynamic Bayesian networks and machine learning algorithms. We then describe **how these techniques enabled Stanley to represent the world**. (2008, p. 170, my emphasis)

The more we understand the hardware and software underlying Stanley's behaviour, the more we can understand how Stanley is causally interacting in the world. This allows us to then generate an intentional model of Stanley which can highlight important regularities in his behaviour.

6.2.4 Both models are used to help us learn about unknown causal mechanisms

If we are trying to learn about the unknown mechanisms of a system, one good way to do this is by understanding the behavioural regularities produced by these mechanisms. Statistical and Intentional models are useful for exactly this purpose. The more predictively accurate our models become, the more it constrains the set of possible mechanisms that can explain the exact behavioural regularities we find. Take, for example, Stocker & Simoncelli's model. While their account does not tell us what the mechanisms underlying the system are, it does provide *insights* into possible mechanisms. They say:

> The form of the contrast-dependent measurement noise in our model suggests that the locus of representation for measurements *m* is likely to be cortical. Neurons in area MT are a natural

> choice: They are highly motion selective, and their responses have been directly linked to perception. (2006, p. 583)

Even though Stocker & Simoncelli's model does not directly describe the neurological mechanisms responsible for the Thompson Effect, by describing and predicting the phenomenon in a detailed way, it provides insights into what the mechanisms must be like. The more detailed our account of the phenomenon, the more it tells us about what the unknown mechanisms must be capable of producing given known constraints. And this in turn helps us narrow the field of possible mechanisms. In this case, it means that whatever mechanisms are responsible for producing the Thompson Effect must meet the regularities observed by Stocker & Simoncelli's model. To demonstrate, consider their postulation that MT neurons may be part of the mechanism producing the effect:

> This implies that the MT population responses should reflect the influence of the prior, varying with contrast in a way that is consistent with the perceptual biases exhibited by the Bayesian observer model. (2006, p. 583)

Now let us turn to our intentional model. By applying an intentional model to Stanley, we can predict the sorts of discriminations he will be able to make. He can discriminate between "drivable" and "undrivable" terrain for instance. And perhaps between "rocks" and "bales of hay". The more predictive our intentional model becomes of a wide range of behaviours, the more fine-grained our account of Stanley's discriminations becomes possible. This, in turn, will put constraints on the possible mechanisms that can produce these discriminations given other known constraints. Not any implementation will do, and knowing how the system behaves in the appropriate circumstances tells us what the system must be like. As a clear example of this, consider the following passage from Eliasmith:

> …consider the example of deciding whether an object in the environment is a friend or foe. Suppose we have two different implementations of the function that needs to be computed to successfully achieve this recognition. One of these implementations, Athlon Alan, can compute this function in less than a second given its architecture, computational primitives, and so on. The other implementation, Intel Alan, takes nearly 10 minutes to perform the same computation because its architecture, computational primitives, and so on aren't optimized for this kind of computation. In other words, the computational complexity of the algorithm on the second implementation is significantly higher than on the first implementation. Of course, if the object is a foe, Intel Alan may not have the 10 minutes required to make this decision and thus may not ever exhibit this cognitive behavior. (2002, p. 5)

In the example above, there are different ways of implementing the same behaviour (e.g., identifying something as a predator), but situational constraints tell us which of these possible implementations of the system are more likely given the context in which the behaviour is observed. In the case of Stanley, for instance, we know that the mechanisms underlying his behaviour must be able to make the appropriate discriminations *g*iven the sort of environment he is in, and the time constraints we observe of his actual behaviour on the DARPA course. So, for example, we might rule out the idea that Stanley is implemented by a series of water-pipes since, given the observed speed of his actual behaviour, water might simply travel too slowly to effectively implement the sorts of behavioural discrepancies identified by our intentional model within the appropriate timeframe. And so the more detailed and predictive our intentional model becomes, the more constraints it places on the sorts of implementations that are possible for the system given other known constraints.

6.2.5 Generating a mechanistic account of a system does not make either model obsolete

While both statistical and intentional models are used to help generate mechanistic accounts of systems, this does not imply that such models become obsolete the moment we have a mechanistic account in

hand. We use different scientific models for different purposes. And while mechanistic models are ideal for characterizing the structure of systems, this does not mean they are ideal for other sorts of purposes.

For instance, a mechanistic model may not always be the best model to use for generating predictions. As Mark Wilson notes, "the reasoning requirements natural to design tasks are often quite different than those pertinent to prediction et al. and greatly influence the descriptive vocabulary we find suitable" (Wilson, 2006, p. 326).

As a clear example of this, consider two different ways of modeling the behaviour of water. One which models water based on its atomic constituents and how they causally interact to produce behaviour, and the other which says nothing about the mechanisms responsible for the behaviour and only describes behavioural regularities:

> If one is studying diffusion or Brownian motion, one adopts a molecular perspective in which water is regarded as a collection of particles. […] However, if one's concern is the behavior of water flowing through pipes, the best-fitting models are generated within a perspective that models water as a continuous fluid. Thus, one's theoretical perspective on the nature of water depends on the kind of problem one faces. Employing a plurality of perspectives has a solid *pragmatic* justification. There are different problems to be solved, and neither perspective by itself provides adequate resources for solving all the problems. (Giere, 2006, p. 34)

In the case of water's movement through pipes, the use of a phenomenological model allows us to predict and describe the behaviour of water better than the mechanistic model. The model that is best for mechanistically explaining the behaviour of water (in terms of molecular motion) is not necessarily the best model to use for predicting the behaviour of water as it flows through pipes.

Similarly, knowing the mechanisms that produce the Thompson Effect does not mean that we will not use a statistical model like Stocker & Simoncelli's in order to predict the effect. And the same lesson applies to intentional models. Consider the problems that faced the neuroscience of vision in the

mid-20th century. The neurological mechanisms responsible for vision were exceedingly difficult to predict based only on a mechanistic understanding of the system. This is because, as Bechtel points out,

> Understanding how the visual system is organized, coordinated with other physiological systems, and responsive to external stimuli, requires knowledge beyond the specification of the parts of the visual system and their operation. (Betchel 2007, p. 184)

It was not until neuroscientists began attributing *intentional content* to neurological systems that relevant predictions could be made. Certain neurons, for example, were found to contain visual information (in the *intentional* sense, not merely the technical sense) *about* the edges of objects in one's visual field. According to David Marr, this information needed to be "analyzed and understood in a way that [was] **independent of the particular mechanisms and structures** that implement them in our head" (Marr, 1982, p. 19, my emphasis). Only by intentionally modeling the system could we generate relevant predictions.

Put simply, intentional and statistical models are used to represent the behavioural regularities or patterns produced by systems, while mechanistic models are used to represent internal parts and operations. In this regard, the different types of models are used to represent different aspects of systems for different purposes, and so will have different pragmatic virtues.

6.2.6 Given these similarities, both models function as phenomenological models

The similarities listed above between statistical and intentional models provide evidence for the idea that intentional language functions within scientific practice as a type of phenomenological model. Consider that a benefit of both statistical and intentional models is that they allow us to generate predictions of systems whose mechanisms we cannot identify. In other words, both models are implementation independent. This is one of the defining features of phenomenological models: "[They] are a means for extracting *stable phenomenologies* from unknown, and perhaps unknowable detailed theories" (Batterman

2002a, p. 35). This is a point echoed by Craver, who tells us that phenomenological models "are complete black boxes: they reveal nothing about the underlying mechanisms and so merely 'save the phenomenon' to be explained" (Craver, 2006, p. 360).

Recall that phenomenological models come in all shapes and sizes, each useful in different contexts and for different purposes. Intentional models are simply another breed of phenomenological model, one useful for predicting behaviour that other models are not well suited for. So for instance, intentional models allow us to generate predictions when we do not have the luxury or ability to quantify over relevant features of the system needed to generate statistical or dynamical models. Similarly, intentional models may be essential in helping to construct other phenomenological models that require such quantification. Consider that, despite a lack of quantification, our intentional model of Stanley will help us determine the sorts of discriminations that Stanley can make of his environment. This, in turn, will help us determine the relevant probabilities to ascribe to various states of Stanley needed to create a statistical model.

If we understand intentional models not as descriptions of the structure of systems, but instead as a species of phenomenological model, then their role in science becomes much clearer. However, up until this point, I have focused only on the similarities between intentional and statistical models to make this case. In order to defend my position, I must also rule out any relevant dissimilarities between the two models that might disqualify intentional models from being a type of phenomenological model. In the following chapter, I examine such dissimilarities and demonstrate that they are not a threat to the account I have provided above.

Chapter 7
Differences Between Statistical Models and Intentional Models

Just as there are similarities between intentional and statistical models, so too are there key differences. In this chapter, I examine these differences in detail and see if any of them pose a threat to the idea that intentional models function as phenomenological models. Ultimately, I demonstrate that the differences between statistical and intentional models are *shallow*, and insufficient to threaten the account I provide.

7.1 We Use Intentional Models to Explain Behaviour, Not Just Predict It

The first important difference to note is that while we commonly explain behaviour by appealing to intentional models, we typically do not explain behaviour by appealing to statistical models. Stocker & Simoncelli's model is not considered to be an *explanation* of the Thompson Effect, but only a means of predicting it. Intentional models, on the other hand, are commonly used for explanations.

Consider that we do not only *predict* Stanley's behaviour on the DARPA course with an intentional model, but *explain* it as well. We can say that Stanley avoids the rocks in his path *because* he has representations of them, and instructions for how to behave when confronted with such representations. This explains why Stanley swerves to avoid obstacles. Or consider our day-to-day application of intentional concepts. I can explain Julian's getting a sandwich from the fridge in virtue of him *wanting* a sandwich and *believing* that one is in the fridge. In this regard, intentional models do not seem to function the way statistical models do given that they are explanatory as well as predictive.

I propose that this distinction between statistical and intentional models is ultimately not a reason to deny that intentional models are phenomenological models. While it is true that we often do not use statistical models to generate explanations in the behavioural sciences, this does not mean that we do not use phenomenological models *of any kind* to provide explanations in these domains. Phenomenological models can often provide us with scientific explanations depending on the particular question being

asked. Batterman, for example, argues that some phenomena can only be seen and characterized by employing phenomenological models, and thus these models are an essential part of our scientific explanations (Batterman 2000, 2002).

But even if we were to suppose that phenomenological models were *always* insufficient to function as rigorous scientific explanations in the sciences of the mind, this would still not be enough to challenge the idea that intentional models are phenomenological models simply because they are used to provide explanations. Consider what it means to explain a mechanistic system: "explanations afford the ability to say not merely how the system in fact behaves, but to say how it *will* behave under a variety of interventions" (Craver 2006, p. 358, emphasis in text). The use of phenomenological models would still allow us to determine a limited range of counter-factual behaviours in virtue of being predictively adequate:

> Because phenomenal models summarize the phenomenon to be explained, they typically allow one to answer some ["what-if-things-had-been-different"]-questions. (Craver, 2006, p. 358)

In this respect, such models would still have a degree of explanatory power. This appears to be why Craver considers phenomenological models to be on a continuum with mechanistic models when it comes to providing explanations of phenomena in the life sciences (2006, p. 360). In this respect, intentional models would still provide us with limited explanations, in exactly the same way that all predictive phenomenological models do. To further emphasize this point, consider Dennett's story of how we can use evolution to *explain* the spots on a butterfly's wings:

> No one ever has ever supposed that individual moths and butterflies with eye spots on their wings figured out the bright idea of camouflage paint and acted on it. Yet the deceptive rationale is there all the same, and to say it is *there* is to say that there is a domain within which it is *predictive* and, hence, explanatory. (Dennett 1987, p. 259)

Explaining the spots on the wings of butterflies in terms of their ability to fool predators does not tell us the mechanism by which evolution works. However, in virtue of being predictive, it provides us with a degree of explanatory power.

Of course, if one is committed to the idea that mechanistic models are the only models that truly explain in the sciences of the mind, then these sorts of phenomenological explanations would be too limited to function as rigorous scientific explanations. Craver, for instance, tells us that a mechanistic explanation "shows why the relations are as they are in the [phenomenological] model, and so reveals conditions under which those relations might change or fail to hold altogether." (Craver, 2006, p. 358). In other words, talking about Stanley on the DARPA course in terms of his representations does not tell us *why* his representations are as they are, nor *how* or *why* they might change or cease to exist under different circumstances. And in this respect, intentional models would not be explanatory *enough*. However, even if one were committed to this, it does not imply that such models would have no explanatory power *whatsoever*. And so their ability to provide explanations would not make intentional models non-phenomenological, it would simply mean their explanatory power is limited.

It is also important to note that even if mechanistic models are considered explanatory in the life sciences, they are not necessarily explanatory for everyday purposes, which is where we commonly find intentional explanations. To demonstrate, consider a variation on the following example offered by Van Fraassen:

> Suppose a father asks his teenage son, 'Why is the porch light on?' and the son replies 'The porch switch is closed and the electricity is reaching the bulb through that switch.' At this point you are most likely to feel that the son is being impudent. This is because you are most likely to think that the sort of answer the father needed was something like: 'Because we are expecting company.' But it is easy to imagine a less likely question context: the father and the son are re-wiring the house and the father, unexpectedly seeing the porch light on, fears that he has caused a short

> circuit that bypasses the porch light switch. In the second case, he is not interested in the human expectations or desires that led to the depressing of the switch. (1980, p. 131)

Now imagine a similar scenario where the son *is* being impudent in his response given that the father is looking for a more mundane reason for the light being on. Only instead of claiming that "The porch switch is closed and the electricity is reaching the bulb through that switch", the son instead gives a detailed mechanistic explanation of how the neurons in his brain interacted, and how they resulted in his lifting his finger to flip the switch. The father, annoyed, tells his son to stop being childish and give him a real answer. The son capitulates and tells the father that the light is on *because they are expecting company*. Now in this sort of everyday context, the mechanistic explanation is the wrong sort of answer given the interests of the father. He simply does not care about neurological and physiological mechanisms. What this shows is that what counts as an explanation in everyday contexts is often not going to be the same as what counts as explanations in various scientific contexts. And so the fact that intentional attributions are commonly used to provide explanations in everyday life is not a threat to the idea that intentional models are phenomenological models. It just means that some phenomenological models can be explanatory for everyday purposes.

7.2 Intentional Models Are Normative, Statistical Models Are Not

Unlike statistical models, it has been argued that intentional models are inherently *normative*, and as a result, they are not *descriptive*. In other words, when we describe a system in terms of intentional states, "we project ourselves into what, from his remarks and other indications, we imagine the speaker's state of mind to have been" (Quine, 1960, p. 210). Put simply, when we describe a system using an intentional model, we attribute to it a set of intentional states that we feel it *ought to have* if it were a rational agent in that scenario. Thus to describe a system in terms of intentional states requires adopting an interpretation of the system as being rational. Consider Stanley and his navigation of the DARPA course. If Stanley's

behaviour were completely irrational (i.e., he behaved erratically and randomly), we would be unable to understand his behaviour in terms of what he can, or cannot, represent. Part of what it means for Stanley to have representations is for him to behave in such a way that uses these representations to guide his behaviour in a meaningful way. And so we assume that: *if* Stanley has goal *x* (to successfully navigate his environment), and has representations *p, q, and r* ("There is a rock 10 meters ahead", "the terrain to the left of the rock is undrivable", "the terrain to the right of the rock is drivable"), he will do *y* (turn right). But this presupposes that Stanley is rational (e.g., that if Stanley wants to successfully navigate his environment, and believes that avoiding obstacles and driving on drivable terrain is the best way to achieve this goal, then he *will* try to avoid obstacles and drive on drivable terrain).

The problem is that most systems do not behave rationally (at least not completely). Rationality is an ideal that is often not attainable by the system being predicted. Instead, we merely choose to *interpret* the system as being rational, and as having rationally connected intentional states, in order to generate our predictions. However, given that our predictions of the system are based on what a *rational system would do* if it had the intentional states we attribute to it, and not on how the system genuinely works, it means that our predictions are not based on an *empirical* description of what the system is actually like, or how it is operating. This means that we can always attribute different sets of intentional states to the same system that will result in the same predictions (just so long as the rational connection between these intentional states leads to the same behaviours). As a result, the same physical system can always be interpreted as having vastly different sets of intentional states. This is why Dennett tells us that "deciding on the basis of available evidence that something is (to be treated as) an Intentional system permits predictions having a normative or logical basis rather than an empirical one" (Dennett, 1971, p. 97).

Given this normative and interpretative feature of intentional descriptions (i.e., given the unavoidable indeterminacy of translation between intentional and mechanistic descriptions), intentional models appear ill-suited to play a role in the empirical sciences which trade explicitly in structural and

causal descriptions. As Fodor puts it, every branch of science "is in the business of causal explanations" (1987, p. 33). Yet intentional models do not provide us with causal explanations. This is precisely why Quine argues that "the underlying methodology of the idioms of propositional attitudes contrasts strikingly with the spirit of objective science at its most representative" (1960, p. 218). As a result, there appears to be an important distinction in kind between intentional models (which are normative) and statistical models (which are not).

This sort of argument is problematic on multiple fronts. First, it mistakenly assumes that science is only in the business of describing the causal and/or structural properties of systems. But this is simply untrue. Characterizing the behavioural regularities produced by systems is an integral part of science as well. As Batterman points out, "a broad goal of scientific theorizing is to recognize and explain observed patterns in the behavior of systems of a given type" (Batterman 2000, p. 120). Both intentional models and statistical models are important tools in our recognition of such patterns. In this regard, they both play an important role in scientific inquiry.

Second, the fact that intentional models do not describe causal or structural features of systems just means that such models are implementation independent. And this is one of the defining features of phenomenological models more broadly. Statistical models similarly tell us nothing about the causal and structural properties of systems. Yet, we do not consider them to be at odds with scientific methodology as a result. In fact, as I mentioned in section 6.2.5, the fact that phenomenological models do not predict based on an account of the system's underlying structure is one of the very *benefits* of phenomenological models. This makes them useful in contexts where the mechanisms of the system are unknown (and thus such models are essential in *learning about* mechanisms). Similarly, they are often useful for identifying behavioural regularities that structural descriptions miss.

Third, the suggestion that we can attribute a radically different set of intentional states to a system and always account for the same behaviour is not so obvious. Even if multiple sets of intentional

attributions can account for the given behaviour of a system in a particular context, this does not mean that they will all work just as well at predicting the system when it is placed in a variety of contexts. It is an empirical question which sets of intentional attributions will be predictive of the system as a whole in a wide range of situations and environments. To take a simple example, consider the following passage from Dennett:

> When [a frog] looks around for flies, can it be said to be looking for flies *qua* flies, or merely *qua* dark, darting, edible things or *qua* something still less specific? (Dennett 1987, p. 108)

When predicting the behaviour of a frog that is presented with a fly, we might predict its behaviour by attributing to it the representation "fly", or we might make the same predictions by attributing to it the representation "dark, darting, edible thing". Or by attributing some other representation altogether. Any of these attributions will make the predictions come out right in this case. This, however, does not mean that any of these attributions are just as good as any other. We can, for instance, run experiments and see which of these intentional attributions best predicts the frog's behaviour in a wider array of contexts. If, for instance, we present the frog with darting black objects and find that it *always* reacts to such objects the same way it does to flies, then we have reason to view one of these intentional attributions ("fly") as less useful than others ("dark, darting, edible things"). Recall that the more detailed our account of the phenomena under investigation (in this case: the behaviour of the system), the more it constrains the set of possible phenomenological models that can account for it. This is a clear case of that. The more detailed our account of the discriminations and behaviour of frog, the more it constrains the set of possible intentional attributions that will account for the behaviour.

This does not necessarily mean that only one intentional model will always win out over all others in terms of predictive power for a given system. We might still be left with multiple intentional interpretations of the system that can account for all the same behaviours in all the same contexts. To

borrow a classic philosophical example, there may be no empirical test we could run that would determine whether the system has a representation of "rabbit", or "undetached rabbit parts" (Quine 1960). In this situation, what are we to say? There are two possible ways to deal with this. First, we could claim that if nothing in principle could ever distinguish between these two representations in terms of the behaviour or inferences of the system, then the meaning of the two are isomorphic. Second, if we insist that the meaning of the two really are distinct even if no test could distinguish between them, then I propose there is simply no fact of the matter which the system really has. For our scientific purposes, such fine-grained distinctions would have no import *by definition.* And so this sort of case is not a substantial argument against the scientific worth of intentional models as a type of phenomenological model.

Fourth, the normative argument suggests that intentional models are ill-equipped to play a role in scientific practice because they make predictions by *interpreting the system as being rational when we know it likely is not* instead of predicting the system based on its underlying structure. Thus, our model is not a descriptive claim about the system, and is instead an account of how the system *ought* to behave *if* it were rational. In this respect, the model is not empirical, and thus at odds with scientific practice. But I propose that this sort of move is no different in kind from our use of idealized models more generally in science. So for instance, we often use Newtonian mechanics to model systems in scientific contexts even though we know that the system is not actually Newtonian. Just as with our intentional model, our Newtonian model describes how the system *ought* to behave *if* we assume that Newtonian physics is true. And this interpretation is not based on an *empirical* description of the actual system, since we already know that Newtonian physics is *not a true description of the system*. In this regard, such models are normative in the same respect. But this hardly makes such models unscientific.

As an aside, it is also important to note that we cannot simply dismiss the presence of such idealized models in science as being temporary evils that we only tolerate until we can provide more correct empirical descriptions of the system (and thus confine intentional models to a class of scientific

models that are only temporarily useful, and which are to be discarded later). Such idealizations are unavoidable and integral to science:

> Any mathematical theory of physics must idealize nature. That much of nature is left unrepresented in any one theory, is obvious; less so, that theory may err in adding extra features not dictated by experience. For example, the infinity of space is itself a purely mathematical concept, and all theories within this space must share in the geometrical idealization already implied. (Truesdell 1960, p. 31)

Batterman makes a similar point, telling us that idealized models are often the *only* way to scientifically model certain phenomena, and solve certain problems, in physics. He tells us that it is "because of the extreme idealizations involved that [such models] are candidates for exact solutions." (2002a, p. 22) Thus the fact that intentional models attribute an idealized notion of rationality to systems as a means of generating predictions is not something that disqualifies it from being a phenomenological model. In fact, this is a move that is actually quite common in our use of phenomenological models.

Ultimately, the normative feature of intentional models is simply not a relevant concern. The scientific value of intentional models is based on their ability to predict systems. Thus, it is *qua* predictive model that intentional models are relevant to scientific inquiry, not *qua* non-normative model. The normative feature of intentional models is irrelevant to their scientific worth.

7.3 Statistical Models Are Based on Well-Defined Axioms, Intentional Models Are Not

Another potential difference between intentional and statistical models is that while statistical models are based on the clearly-defined axioms of probability theory, the axioms of rationality (on which intentional attributions seem to be based) are not so apparent. It is questionable whether there even are any clearly definable axioms of rationality. When we predict the behaviour of others, we do so without any explicit

understanding of what such axioms might be. In this respect, intentional models may be too different in kind from statistical models to warrant inclusion into the class of phenomenological models.

The problem with this sort of objection is two-fold. First, it is a mistake to think that all domains to which statistical models are applied are therefore axiomatized. That there are statistical axioms does not mean that any particular statistical model (e.g., of stock markets) is also axiomatized in the relevant respect. We do not, for instance, have the axioms of stock market behaviour just because we have a statistical model of the stock markets. You could, after all, have a statistical model of rationality.

Second, and more importantly, this sort of objection is a red-herring. Even if we assume that there are no explicit axioms of rationality (which is still up for debate), it is hardly a necessary characteristic of phenomenological models that they be based on explicitly identified axioms. That was never the claim. Intentional models are, however, relevantly like statistical models in key respects: First, they are predictively valuable in scientific practice. Second, they make predictions without telling us structural or mechanistic details of the system. Third, they are often used to identify patterns and regularities in behaviour produced by mechanistic systems. Lastly, they are used in conjunction with mechanistic models to provide more complete understandings of systems. These similarities provide us with compelling reasons to consider intentional characterizations as a species of phenomenological model, just as statistical models are.

7.4 Statistical Models Involve Quantification While Intentional Models Do Not

Statistical models are mathematical in nature and involve precise measurements and calculations in order to make predictions. Intentional models, on the other hand, often do not involve this kind of mathematization of the system, and thus are substantially less rigorous and precise. This means that such models may be far too insubstantial to have merit as a type of phenomenological model.

The problem with this suggestion is that it assumes that *only* those scientific representations that involve explicit quantification are truly scientifically relevant. But this is simply untrue. Images and

pictorial representations are commonly used in science even when they do not explicitly mathematize the system (Larkin & Simon 1987; Nersessian 1988; Giere 1996; Meynell 2008).[23] As Nancy Nersessian (1988) tells us, "the history of science abounds with instances of the use of imagery and of analogy to articulate vague notions into socially shared, scientifically viable conceptualizations of a domain. Periods of 'scientific revolutions' are most fertile with examples." (p. 42) As an example, consider that when trying to learn about a mechanistic system, it is often common practice to provide an abstract sketch of a mechanism involving boxes and arrows (Machamer et al. 2000, p. 8; Cummins 2000, p. 125; Zednik 2011, p. 249). Even in contexts where such images cannot yet be mathematized, they can still be scientifically informative by identifying or characterizing "spatial relations and structural features of the entities in the mechanism" (Machamer et al. 2000, p. 8).

To claim that such visual representations of systems are unscientific simply in virtue of not mathematically characterizing the system does not do justice to way such representations are used in science. Of course, it is true that images which lack explicit quantification are often imprecise. However, such accounts are often a necessary first step in generating more exact mathematized models. As Carlos Zednik points out, "mechanistic explanation frequently starts as a relatively abstract mechanism sketch that leaves ample room for elaboration" (2011, p. 249).

I propose that all this is similarly true of intentional models. Their value to science is not diminished by their lack of mathematization. In fact, they can often be a necessary step in developing such mathematical models, and in learning about mechanistic systems, by allowing us to predict when mathematization is not an option (see section 6.2.6). In this respect, the fact that such models do not mathematize the system is not sufficient to deny intentional models a place among our other scientific models.

[23] It is extremely important to note that the question of whether we *could* mathematize such images is irrelevant. The question is whether such images, *in their pictorial non-mathematized form*, are informative to science.

7.5 Intentional Models Are a Species of Phenomenological Model

These differences between statistical and intentional models are simply not enough to support the claim that intentional models behave in a fundamentally different manner from the way we use phenomenological models in science. In fact, some of the supposed differences between the two models are only illusory. In terms of the normative/descriptive difference, statistical models are just as normative as intentional models in the sense that neither model describes the structure and causal features of systems (e.g., they are both implementation independent).

Ironically, the differences between intentional and statistical models that are *not* illusory provide us with even more reasons to think that intentional models behave as a species of phenomenological model. While intentional models work with an idealized notion of rationality that statistical models do not, it is actually quite common for phenomenological models to employ similar kinds of idealizations. In fact, such idealizations are an integral and essential part of science. Meanwhile, the fact that statistical models make predictions by mathematizing the system in a way that intentional models do not only shows that intentional models and statistical models are ideal for predictions in different contexts (depending on what kind of information we have available about the system).

With this understanding of intentional models as phenomenological models, we can now develop a much clearer idea of what the methodological study of the mind looks like. In the next Chapter, I provide a general overview of how we use different types of models (including intentional models) to provide a more complete understanding of the mind.

Chapter 8
The Bigger Picture

With a clear sense of how intentional models are used in science, we can now take a step back and examine the implications that this account has for the scientific study of the mind more generally. In the first part of this chapter, I focus on the methodological benefits of employing a wide variety of different kinds of scientific models when studying the mind. A diversity of models is, at the very least, pragmatically necessary in developing a mechanistic explanation of the mind. Intentional models are amongst those that we employ for such pragmatic purposes, helping us to generate mechanistic explanations (even if they are often not explanations themselves).

In the second part of the chapter, I examine the possibility that intentional models may be far more than merely pragmatically useful for helping to generate mechanistic explanations. The idea that a single type of model (such as a mechanistic model) can account for everything that is scientific relevant about the mind is far from obvious. This claim should not be confused with the claim that everything that matters about the mind may be the result of physical mechanisms. One is a claim about physical mechanisms, while the other is a claim about our capacity to adequately represent all the relevant features of physical mechanisms using only one type of scientific model. Mechanistic models capture certain aspects of the physical mechanisms that make up the brain, but they may do so at the cost of identifying or representing other aspects. If this is the case, then different types of scientific models may be integral to our continued understanding of the mind even with a detailed mechanistic explanation of the brain in hand. Our scientific picture of the mind may not be one that can be captured by any single type of scientific model, and so attempts to develop one all-encompassing model of the mind may be based on a mischaracterization of the nature of scientific representation. At most, we may be left with a *set* of different types of scientific models that we must stitch together, each capturing different aspects of the

phenomenon. While I will not argue that this sort of fragmented pluralism of models definitely *is* the case, I will argue that we have compelling reasons to consider it a very real possibility for a final state of science. If it turns out to be true, then intentional models may be essential to science regardless of whether we have a complete mechanistic explanation of the mind or not.

8.1 The Pragmatic Benefits of Using Different Types of Models

If we are interested in a mechanistic explanation of the mind, then how can we go about identifying the mechanisms working within the brain? The assumption that we need only look at the brain to easily determine the parts and operations responsible for generating a given mental phenomenon is, of course, naïve. Understanding the mechanisms working inside the brain is fraught with complications and problems:

> The challenge in constructing mechanistic explanations is that normally operating mechanisms do not reveal either their parts or operations. Not just any way of carving up the mechanism reveals the appropriate parts. The relevant parts are those that actually perform the operations in the mechanism. To consider an example, although neuroanatomists over several centuries sought to delineate parts of the brains of humans and other species in terms of the gyri and sulci produced by the folding of the cortex, and these still serve as useful landmarks when identifying where operations occur in the brain, they do not represent the working parts. […] As challenging as it is to identify candidate working parts, it is even harder to identify the component operations. (Bechtel 2005, p. 316)

Part of the problem is that mechanisms are often hierarchical. In other words, a component part of a mechanism may itself be a mechanism, which can be decomposed into further parts and operations (see Section 2.2). This hierarchical feature of mechanisms gives the misleading impression that the parts of the embedded mechanism (often referred to as the mechanism at the "lower level") must be smaller in size

than the parts of the mechanism in which it is embedded (the mechanism at the "higher level"). But this is untrue:

> When we identify levels in terms of causal interactions within a mechanism, entities that are structurally alike may appear at different levels. Protons for example interact with membranes in the chemiosmotic mechanism responsible for converting energy liberated in oxidative reactions in cells into a proton gradient that drives ATP synthesis. Protons also occur in the molecules that comprise the membrane, but these are at a lower level than the protons that are transported across the membrane and thus interact with it. There is not a level of protons, but levels corresponding to the entities that causally interact in a given mechanism. The result is a hierarchy of levels, but one that is characterized relative to the phenomenon an investigator initially sets out to explain. (Bechtel 2005, p. 315)

What this means is that merely looking inside the brain and identifying objects does not tell us which mechanisms those objects are parts of, what level they belong to, or how they interact with other objects. And so attempting to generate a mechanistic model of the brain simply by cracking open the skull and looking inside is inherently problematic. So how then do we proceed?

Keep in mind that we often learn about neurological mechanisms by way of reverse engineering. Recall the example from Section 2.2 of Bartolomeo Panizza, who was able to learn about the mechanisms responsible for vision by studying people that went blind after damage to the occipital lobe. In this instance, Panizza learned about the relevant mechanisms for vision by first identifying and understanding the phenomenon (vision), and then determining how the phenomenon changed when alterations were made to the neurological mechanisms. Thus, understanding and characterizing the phenomenon produced by the mechanism was the first step in learning about the parts and operations that brought it about. Of course, Panizza could not use a mechanistic model of vision to identify and characterize the relevant phenomenon, since none was available to him (the mechanisms responsible for vision were what he was

trying to *discover* after all). And so characterizing the relevant phenomenon had to be done by employing different sorts of scientific tools. And this is where phenomenological models become essential.

Our use of intentional models, for instance, can help us refine our understanding of the phenomenon of vision by identifying the sorts of discriminations that the system can make of objects appearing in its visual field. Similarly, statistical models like Stocker & Simoncelli's model allow us to identify and predict the way in which we visually discriminate the speed of objects. Both of these models allow us to better understand the phenomenon under investigation. These phenomenological models can provide us with invaluable information about the conditions under which the phenomenon is produced, which in turn places constraints on what the mechanisms that produce the phenomenon must be like. Consider the phenomenon of fermentation:

> …knowledge of how to set up conditions for fermentation was acquired by brewers long before the development of biochemistry in the early 20th century, and the knowledge was not supplanted by the investigations of biochemists. On the contrary, biochemists employed such knowledge in setting up the experimental conditions in which they could study the operation of the enzymes. (Bechtel 2007, p. 183)

Phenomenological models allow us to identify constraints on possible mechanisms, which helps us to narrow the list of potential mechanisms responsible for the phenomenon. In this way, phenomenological models can act as pointers, showing us where to look for possible mechanisms (just as Stocker & Simoncelli's model pointed them towards cortical neurons).

Of course, these sorts of insights are often, by themselves, insufficient to tell us exactly what the mechanisms are, or how they function. However, they can often give us a general sense of what the mechanism might be like, which allows us to construct a first pass at a mechanism sketch —even if this sketch is extremely vague. Recall Zednik's claim mentioned in Section 7.4 that "mechanistic explanation frequently starts as a relatively abstract mechanism sketch that leaves ample room for elaboration" (2011,

p. 249). The next step is to take this abstract sketch, and to return once again to our various phenomenological models. Would the abstract mechanism we posit be able to produce the exact same regularities that both our statistical and intentional models were able to identify from the actual system's behaviour? If so, then we have some evidence that the mechanism sketch may be on the right track. If not, then what sorts of modifications to our sketch will better account for the regularities? It is not simply a linear methodological path from phenomenological models to mechanistic ones. Instead, there is a constant back and forth between the different models throughout the process of scientific inquiry that allows them to constantly benefit from the insights of each other.

The relatively abstract mechanism sketch that we develop (with help from our phenomenological models) is often fed *back* into our phenomenological models. Just as knowing more about the phenomenon allows us to refine our account of the mechanisms, so too does knowing more about the mechanisms allow us to refine our phenomenological models:

> Sometimes knowledge about the components of a mechanism can guide inquiry into how the mechanism engages its environment and when such knowledge is available, ignoring it is fooldhardy. The same, though, applies in the opposite direction —knowing how a mechanism behaves under different conditions can guide the attempt to understand its internal operation. (Bechtel 2007, p. 174)

Knowing more about how the mechanisms of a system allow it to make discriminations of its environment allows us to better refine the intentional content we attribute to the system (making for better intentional models). Similarly, knowing more about the underlying mechanisms allows us to better assign probabilities to the various states of the system required for making better statistical models. This, in turn, allows us to further refine our understanding of the intricacies of the phenomenon, and puts still further constraints on what the mechanisms must be. This back and forth process between our various models "enables the refinement of methodologies, the clarification of concepts, the design of experiments and

studies to control for causal factors demonstrated by others. All this makes for more knowledge, which, judged by means of the evaluative tools available [to] each [model], is also better knowledge" (Longino 2006, p. 127).

Similarly, by employing a variety of different phenomenological models (intentional, statistical, dynamical, etc), we can provide a more comprehensive understanding of the phenomenon under investigation by predicting it in a wider array of contexts. Different models use different information about the system in order to form predictions. As a result, different models will be ideal for predictions in different contexts. Thus even if we could, in principle, use a single phenomenological model to predict all the behaviours of a given system, in practice we often lack the information needed to generate such an all-encompassing model. We might lack the information required to generate a statistical model in one context (given a contextual inability to appropriately mathematize the system), but not an intentional model. Therefore, our use of an intentional model in such a context would provide additional information about the phenomenon that we would lack by only relying on statistical models to form our predictions. In this regard, the methodological value of employing multiple phenomenological models comes from their pragmatic benefits under varying circumstances. This allows us to gather a greater, and more varied, pool of information that we can use in our understanding of the mechanisms.

Given that different phenomenological models have different pragmatic constraints on (and ideal conditions for) their application, each model we use provides information that other models may need to take into account, but which we may be unable to gather using those models alone due to contextual limitations. The different models play off each other, providing information that helps to refine and augment the predictive and/or explanatory capabilities of one another. The different kinds of scientific models work in tandem, building off one another.

As this process develops, we occasionally find an overlap between the theoretical objects in one model, and theoretical objects in another. For example, our intentional model, when refined enough, may

posit an intentional state that happens to fit the behaviour of a particular neurological mechanism. In this way, we can correlate certain intentional states with certain mechanisms when our mechanistic and intentional models are sufficiently refined. This sort of overlap is extremely informative and helpful to our understanding of the system when it occurs, but it should be stressed that such overlaps are usually serendipitous, and not the *expected* outcome of our application of such models. We do not use intentional models *only* when the intentional states they refer to functionally correspond to particular mechanisms (see, for instance, the theory of planned behaviour discussed in Section 5.2 and 5.3). We use intentional models because we can generate predictions of systems that are informative, and this provides scientifically relevant data that other models can use to improve upon their own accounts of the phenomenon. The use of an intentional model may be essential in learning about the mechanisms of a system even if no overlap is found between the theoretical objects in the two models. [24] Put simply, while it is immensely beneficial when there *is* an overlap between the theoretical objects of two different types of models, this is not what makes either model worthwhile to our scientific methodology.

This account of the methodological interdependence of multiple scientific models within a given scientific domain may similarly carry over to the interdependence of scientific domains more generally. Consider the relationship between psychology and neuroscience. While some have argued that the theories of psychology can (or at least *should*) be reducible to the theories of neuroscience (Bickle 2003), this idea is often based on the assumption that a completed neuroscience could in principle do everything that a completed psychology does (predictively and explanatorily). But even if we suppose that this is true (which is contentious; see Section 8.2 for details), we do not *have* a completed neuroscience to use in place of psychology. And, the study of psychology is an essential part of our *development* of a complete

[24] Just as, for instance, we do not insist that statistical models are only beneficial to science if and when we can find an overlap between the theoretical objects in statistical models (numbers, averages, variances) and the theoretical objects of other models (like the objects of mechanistic models). Averages are not made up of physical mechanisms any more than physical mechanisms are made up of averages. But that does not mean that statistical models are therefore not beneficial to science.

neuroscience (and vice versa). Just as we use a plurality of models to refine and improve upon our understanding of some phenomenon within a scientific domain, so too is it the case that we use a plurality of scientific disciplines to refine and improve upon our understanding of some aspect of the world.

Recall that our understanding of the mechanisms responsible for vision began with a detailed account of the phenomenon produced by those mechanisms (which pragmatically involved modeling the phenomenon of vision in different ways). By a similar token, in order to develop an understanding of the neurological underpinnings of the brain as a whole, we must similarly begin with a detailed understanding of the phenomena produced by the brain as a whole. And psychology, as a scientific domain, exists for exactly this purpose. Consequently, psychology plays a necessary role in helping to develop a neuroscientific account of the brain by telling us essential information about the behaviour and function of the system that a complete neuroscience will have to account for, and be constrained by. Just as different scientific models within a given domain inform one another (with the results of each model feeding into the others), so too do psychology and neuroscience inform one another (with the results of each domain feeding into the other).[25]

The fact that psychological theories make reference to theoretical objects that do not necessarily overlap with the theoretical objects of neuroscience (beliefs, desires, etc.) is irrelevant to the methodological necessity of psychology. What matters is that psychological theories represent and characterize the way in which people behave in a variety of contexts, the way they react to environmental

[25] Patricia Churchland (1989) offers a similar account of the interaction between psychology and neuroscience. However, she insists that this process of co-evolution always leads to a convergence between the domains, resulting inexorably in a single mechanistic neurobiological account of the system. In this respect, she argues that the co-evolution of psychology and neuroscience inevitably results in a reduction of psychology to neuroscience as the two domains grow together (1989, p. 374).

There are indeed cases where a reductive convergence between psychological and neurological models is ideal. However, this is not *universally* the case. There are cases where such reductions are neither inevitable, nor desirable, depending on what our pragmatic purposes are for modeling the system. The fine-grained neurobiological details of the system are not always what we care about, and psychological models may identify features of systems that neurobiological ones cannot. The inter-dependence and co-evolution of psychology and neuroscience is important even in cases where the two do not converge on a singular model (for more details, see Section 8.2 and 9.4).

conditions, and the way they interact with others in complex social settings. This information plays an essential role in our attempt to develop a neurological account of the mind given that these behaviours are a direct result of neurological mechanisms. So any good neurological account of the brain must conform to the data gathered by psychological studies regarding the way in which the system behaves (just as psychological accounts must conform to the data gathered by neuroscience regarding the way neurological mechanisms produce behaviour). However, the two domains need not share the same vocabulary or theoretical objects in order for this to be the case.

Of course, we will occasionally find overlaps between the theoretical objects of psychology and the theoretical objects of neuroscience as the two domains inform, and feed into, one another. But just as is the case with our use of intentional and mechanistic models, the *value* of psychology and neuroscience is not contingent on this overlap of concepts. What matters is that psychology characterizes important aspects of neurological systems that allows us to generate and acquire essential data for, and determine possible constraints on, our neurological theories. And similarly, these neurological theories in turn generate and acquire essential data for, and determine possible constraints on, our psychological theories. A complete overlap of concepts is not a requirement for this methodological process.

Therefore, even if one wishes to argue that psychological concepts ought to be replaced entirely by neuroscientific concepts at the end of the day, one first needs to have a complete set of neuroscientific concepts and theories before this can be accomplished. And the practice of psychology (along with its conceptual tools) is needed to acquire this. Therefore, one must still climb the psychological ladder before it can be kicked away, and we are still very much in the process of climbing that ladder.

It is also important to note that all of this assumes that a completed neuroscience could, at least in principle, account for all the phenomena described by a completed psychology. In the following section, I will demonstrate why this may in fact be false. And if this is the case, then the role of intentional models

in science may be far more integral to scientific practice than merely existing as a pragmatic tool for generating mechanistic explanations.

8.2 A Patchwork View of Scientific Practice

If the mind *is* the brain, and the brain *is* a set of physical mechanisms, then it seems reasonable to assume that a complete account of these mechanisms will tell us everything there is to know about the mind. But even if this is true, we must be extremely careful not to make the further inference that a complete *mechanistic model* of the brain will necessarily provide us with a complete account of these mechanisms. This is not because scientific models do not provide us with an understanding of the system (they do), but because different types of scientific models might always, out of necessity, distort the system being modelled in different ways so as to only ever provide us with a *partial* understanding of it. Thus, adopting different types of scientific models may be required in order to represent all the relevant features of it.

A growing number of philosophers and scientists suggest that the complexity of nature, and the essential idealizations and distortions that scientific representations must employ, may make it impossible for a single type of model to capture everything that is scientifically relevant about a complex phenomenon (see for instance: Truesdell 1980; Dupré 1993; Suppes 1993; Hacking 1996; Cartwright 1999; Longino 2002, 2006; Batterman 2000, 2002a, 2002b; Fehr 2006; Chemero & Silberstein 2008). As Patrick Suppes notes:

> The application of working scientific theories to particular areas of experience is almost always schematic and highly approximate in character. Whether we are predicting the behavior of elementary particles, the weather, or international trade —any phenomena, in fact, that has a reasonable degree of complexity— we can hope only to encompass a restricted part of the phenomenon. (1993, p. 53)

The assumption that there must be one type of description or model that can correctly and completely account for all aspects of some phenomenon is not something we can take for granted. No model can capture everything about a complex system, and so different models may be necessary in order to fully represent different aspects of a given system or phenomenon. As Truesdell tells us:

> One good theory extracts and exaggerates some facets of the truth. Another good theory may idealize other facets. A theory cannot duplicate nature, for if it did so in all respects, it would be isomorphic to nature itself and hence useless, a mere repetition of all complexity which nature presents to us, that very complexity we frame theories to penetrate and set aside. (1980, p. 72)

What this means is that a particular type of model may be essential for characterizing certain aspects of a system while necessarily being unable to characterize or identify others. In this sense, developing a complete account of the neurological mechanisms that make up the brain (and thus the mind) may *require* that we employ different sorts of scientific models (including intentional models) to represent different aspects of those mechanisms.[26] Consider Bechtel's claim that

> ...researchers have produced knowledge [about neurological mechanisms] that could not have been acquired by [strictly inquiring into the parts and operations of those mechanisms]. Psychophysicists and ecological psychologists complement the reductive inquiries of neurophysiolgists and have not been rendered unnecessary by the neurophysiologists' success. (2007, pp. 184-185)

[26] This point is stronger than the mere *pragmatic* position defended in Section 8.1. The claim is not that multiple models are instrumentally necessary for *generating* a single correct model, but instead that there may not *be* a single correct model. And thus a plurality of different kinds of models may themselves be *part of* our best account of the phenomena. We will always have to employ a variety of different models that characterize a target system in very different ways.

Mechanistic models, like all types of models, emphasize some features of a system at the cost of others, and so they may only be able to provide information regarding certain aspects of a phenomenon while necessarily ignoring others.

To take another example, consider that the more detailed our mechanistic models become, the less informative they tend to be about features that span a range of radically different mechanistic systems.[27] In other words, by increasing the structural details of the system in our model, we correspondingly decrease the model's ability to see certain types of similarities that exist across various types of mechanistic systems:

> Structural theories include more information about singular materials and, as a consequence, less information about a class of materials. For example, the dependence of a macroscopic variable such as viscosity on temperature could be predicted by a kinetic theory. But in this case, for each specific law of intermolecular force, the explanation offered by the structural theory would differ. Due to complexity, there are many cases where the solution to the force equations would be mathematically intractable. Under a phenomenological theory, on the other hand, such a dependence would be ignored, i.e., the theory would be less definite regarding the relation between a macroscopic variable and its molecular support. (Fillion 2008)

This would make the implementation-independence of phenomenological models an invaluable asset to scientific practice. The fact that there is a many-to-one relationship between mechanistic models and phenomenological ones means that the same phenomenological model can be used to identify and characterize similarities that exist among systems that are implemented in very different sorts of ways. It is for this reason that Batterman argues that phenomenological models may be necessary to explain some phenomena that span a variety of mechanistic systems that mechanistic models may be unable to identify (2002a). In such cases, the fine details of the mechanism "may, in fact, actually detract from an

understanding of the phenomenon" (Batterman 2002a, pp. 21-22). Truesdell makes a similar point regarding the value of phenomenological models:

> [We do not consider] a traffic engineer stupid for neglecting to make use of physical and chemical principles determining the motion of the automobile when he sets up his stochastic theories for traffic control. […] For most of the physical phenomena of ordinary experience, considerations of the structure of matter do not yield a finer or more accurate theory: They do not yield any theory at all, any more than nuclear physics, however true, and however useful for studying nuclei in small numbers, gives any information at all about the behavior of a rat, or than classical mechanics, however true, and however useful for explaining motion of a single automobile, gives any information at all about traffic flow. (Truesdell 1966, pp. 215-16)

As a result, the different types of models (mechanistic and phenomenological) are often used to represent different aspects of systems. Each represents features that the other may miss.

To further emphasize this point, consider the common analogy of scientific models as maps of the world. Different types of models act as different sorts of maps that we can use to represent and navigate the world. This analogy between models and maps is apt, given that any two-dimensional map of the earth will necessarily distort the terrain it represents. This is a point that Mark Wilson emphasizes quite strongly:

> As is well known, it is impossible to map terrestrial topography onto a sheet of paper without introducing considerable distortion in the result. At best, we can select a few features that we would like to register in our maps accurately and conveniently, while abandoning other critical qualities to their representational fates. (Wilson 2006, p. 289)

[27] This is not universally true of *all* cross-system similarities, but it need not be to make the point. There need only be a scientifically relevant set of such similarities for the point to be made.

Take Mercator projections for instance. This way of mapping the earth is extremely beneficial to sailing vessels for use in navigation given that "the compass and sextant routes that a sailing vessel might reasonably follow appear on such maps as straight lines" (ibid, p. 289). However, this comes at the cost "of great distortions in areal representation, especially within the higher latitudes (as manifest in Greenland's extremely deceptive size upon a Mercator map)" (ibid, p. 289). Meanwhile, Hammer projections compensate for the distortions in areal representations found in Mercator projections, but "at the price of considerable distortions in shape (worse than on the Mercator, although its depictions of shape are not exactly terrific either)" (ibid, p. 290). When it comes to the different mapping techniques, "each embodies its own distinctive *personality*, which is never in complete harmony with the physical system it attempts to describe: the spherical earth. As we attempt to maximize selected representational virtues (accurate areal representation), we mislead in others (shape)" (ibid, p. 290).

How then do we create a truly accurate map of the earth? The mistake is to assume that there must be one type of map that "gets the world right", with others being dismissed as merely pragmatically useful. Instead, Wilson proposes that the solution is simply to have an atlas of different maps that we constantly move between as the need arises:

> How do we correct for these representational pitfalls in our maps? The effective scheme is to supply a rich *atlas* of maps that cover the earth several times over, each of which is dedicated to answering questions best suited to its own personality. [...] In other words, a competent employer of an atlas will address the questions she seeks by thumbing to the right pages of the atlas, often in a rather complex fashion: a seaman plots sailing routes combining the information supplied in several maps, often without knowing the underlying theory that explains why this bustle of procedures supplies suitable sailing instructions. (Wilson 2006, p. 291)

The analogy between maps and models is importantly informative in this respect. No single type of scientific model may be able to completely account for a given phenomenon, leaving us with a plurality

of models that we must switch between as the need arises. If this sort of account turns out to be true, then intentional models may not simply be essential in our methodological development of mechanistic explanations, but may characterize aspects of systems (aspects that exist across mechanistic systems) that mechanistic models may miss altogether. In which case, intentional models may themselves be *part of* our best scientific account of the mind, as opposed to merely a means of helping us *generate* such an account.

Similarly, if this sort of pluralism turns out to be true, then different scientific domains are not likely to be reducible to others (at least not completely). Neuroscience, for instance, may represent certain aspects of the phenomena of human behaviour at the cost of others, and so may be unable to represent those features characterized by psychology. And indeed, there is some evidence that this is the case. According to Bechtel, our application of social psychology "characterizes regularities in the way cognitive agents respond to situations arising in their environment. This is not information that neuroscientists themselves are interested in or have the tools to procure" (2007, p. 194, endnote 3). Or consider Longino's claim that different "features of competing behavioral research programs […] are better accommodated in a framework that is open to pluralism than one constrained by a commitment to monism" (2006, pp. 126-127). Similar claims can also be found in Scarr 1995, and Chemero & Silberstein 2008.

In this respect, different scientific domains complement each other, and bring different insights to our understanding of the world that other domains lack. Or as Suppes puts it, "the rallying cry of unity followed by three cheers for reductionism should now be replaced by a patient examination of the many ways in which different sciences differ in language, subject matter, and method as well as by synoptic views of the ways in which they are alike" (Suppes 1993, p. 48).

Of course, we cannot rule out the possibility that a unifying model may one day be forthcoming. It would be poor science to merely dismiss the possibility of finding such models in the future (none of us can predict the scientific revolutions that await us in the future). However, by the same token, given what

we know about current scientific practice, it would be just as careless to simply dismiss the sort scientific pluralism discussed here. Put bluntly, those who insist that a unifying model of the mind *must* be forthcoming do not necessarily have the current state of scientific methodology, or the history of science, on their side.[28] The necessity of a plurality of models is something we must take seriously as a live option.

8.3 Summing Up the Big Picture

Ultimately, my intention here is not to argue for the stronger version of scientific pluralism over the methodological one. Instead, it is simply to highlight the essential role that intentional models currently have, and may continue to have, within our scientific practices. Intentional models are methodological tools that co-evolve with, and help to refine, different scientific models that together contribute to our study of the mind. This methodological co-dependence of models is part of the process by which we learn about mechanistic systems like the brain.

It is also a possibility worth considering that a plurality of different types of models is an ineliminable feature of scientific explanation and theory more generally. In which case, the representational benefits of intentional models may be unattainable using other types of models. This being the case, intentionality will remain an integral part of our understanding of the mind in future scientific practice.

[28] That being said, it may be worth striving for such a unified model, since we have no principled reasons to think that such a model is *a priori* impossible, and such pursuits may provide us with important scientific results. The point here is merely to emphasize that such a unified model may not be forthcoming, and thus the value of intentional models to science may be far more profound than previously implied.

Chapter 9
A Brief Philosophical Survey

With my account of the benefits of intentional language on the table, we can now see and appreciate the many insights that different philosophers of mind have brought to our understanding of intentional language in the past. While many of the accounts offered by these philosophers involved commitments that I will demonstrate were unwarranted, they nonetheless emphasized important features of intentional models that are worth noting. In this chapter, I highlight some of these accounts, and demonstrate what their benefits and drawbacks are.

9.1 Eliminative Materialism

To begin, consider Paul Churchland's account of Eliminative Materialism (1981). According to Churchland, intentionality should not be understood as a metaphysical phenomenon unique to the mind, but instead as a structural property of particular kinds of sentences; specifically, the propositional attitudes commonly used in traditional psychology (which he considers to be a form of folk psychology). As he puts it:

> Another conundrum is the intentionality of mental states. The 'propositional attitudes,' as Russell called them, form the systematic core of folk psychology; and their uniqueness and anomalous logical properties have inspired some to see here a fundamental contrast with anything that mere physical phenomena might conceivably display. The key to this matter lies again in the theoretical nature of folk psychology. The intentionality of mental states here emerges not as a mystery of nature, but as a structural feature of the concepts of folk psychology. (Churchland 1981, p. 70)

Like Churchland, I similarly propose that intentionality is best thought of as a means of describing certain kinds of systems, and not necessarily as a unique metaphysical phenomenon in dire need of scientific explanation (at least in its scientific usage).

Where Churchland and I disagree is on the value that these intentional sentences have to scientific discourse. Churchland proposes that intentional descriptions are ultimately rooted in a folk understanding of the mind, and thus will ultimately be displaced by neuroscientific descriptions which eschew intentionality. In other words, he believes that folk psychology has no place in our growing corpus of neuroscientific research, and as such, "intentional categories stand magnificently alone, without visible prospect of reduction to that larger corpus" (1981, p. 71). As a result, he predicts that intentional concepts will ultimately be eliminated from scientific practice. I propose that this position is problematic on numerous fronts.

First, it attempts to contrast neuroscientific descriptions with intentional descriptions, as though the two are mutually exclusive. This is clearly a false dichotomy. Neuroscientists commonly make use of intentional concepts like *representations* and *information*[i]. Intentional descriptions are as much a part of neuroscientific theories as mechanistic descriptions are (see Section 6.2.5).

Second, Churchland argues that the scientific worth of traditional psychology's intentional concepts is contingent on their place within a neuroscientific framework (which Churchland is sceptical we will find). However, as I have argued in Section 8.1, the scientific value of the theoretical objects posited by one domain of science is in no way contingent on their overlap with the theoretical objects of other scientific domains. While such overlaps are certainly nice when they exist, it is hardly a condition on scientific merit. What matters is that the two domains tell us important features about the phenomenon under investigation, not that their concepts or their vocabulary be the same.

Third, Churchland assumes that intentional descriptions describe and characterize the *same aspects of human behaviour* that neuroscientific theories do, but with much less success (and thus will be

replaced). But this may be to misunderstand the way in which the different domains represent phenomena. While there is indeed a degree of overlap in terms of the phenomena studied by both psychology and neuroscience, the two scientific domains might describe different aspects of this phenomena (see Section 8.2). As such, each domain may provide important data and constraints that the other domain cannot identify but must still account for. This would mean that the two domains complement each other. The assumption that neuroscientific theories will trump the theories of traditional psychology once everything shakes out in the end assumes that neuroscience can do everything that psychology does. But this is hardly obvious. Psychology employs different tools than neuroscience, and is often used to characterize different features of human behaviour.

And so the important insights brought to us by Churchland's theory of eliminative materialism are: 1) that intentionality may be rooted in our descriptions of systems as opposed to being a mysterious metaphysical property of systems; and 2) that we should not expect that all our intentional categories will be reducible to neurological mechanisms.

9.2 Functionalism and The Multiple-Realizability Thesis

Next, consider the Multiple Realizability (MR) Thesis, as argued for by Hilary Putnam (1967) and Jerry Fodor (1974). The general idea of MR is that a heterogeneous set of physical processes can instantiate, or bring about, the same higher-level states within systems (in this case: intentional states). Putnam and Fodor propose that intentional states are *functional* states of a system, and thus can be realized by any physical state that plays the appropriate functional role in the behaviour of the system. To explain how a variety of mechanistic systems can bring about the same functional states, they propose that we conceive of the different physical systems as all implementing the same kind of higher-level behavioural mechanism (e.g., one which outputs the same sets of functional states).

So, for instance, even though people, animals, computers, and (potentially) aliens may be composed of different sorts of physical mechanisms, it is possible for all of them to have the same

intentional states in virtue of having the same functional states. By thinking of the different physical processes as alternate ways of implementing the same Turing-style computational system, they can all generate the same intentional states used to guide behaviour.

I propose that one of the main insights of Fodor and Putnam's accounts is that they highlight the fact that we can, and do, successfully predict the behaviour of vastly different sorts of mechanistic systems by appealing to the same sets of intentional states. Previous attempts to find type-type identities between intentional states and physical states (such as brain states) meant that different sorts of systems would be incapable of having the same intentional states that humans could. Yet, it was (and still is) common practice to speak of very different sorts of systems in terms of the same types of intentional states. Thus the development of the MR thesis acknowledged the fact that intentional descriptions apply across a wide variety of mechanistic systems.

Another insight that Putnam and Fodor offer is that a higher-level science (like psychology) need not reduce to a lower-level one (like neuroscience) in order for that domain to scientifically represent genuine features of systems. The idea behind MR is that while the higher-level functional state cannot be reduced to the lower-level mechanisms, it is still a genuine aspect of the system worth understanding. As Fodor puts it:

> It seems to me (to put the point quite generally) that the classical construal of the unity of science has really misconstrued the *goal* of scientific reduction. The point of reduction is *not* primarily to find some natural kind predicate of physics co-extensive with each natural kind predicate of a reduced science. It is, rather, to explicate the physical mechanisms whereby events conform to the laws of the special sciences. (1974, p. 111)

The problem with Putnam and Fodor's account is their insistence that the only way to explain the cross-system application of intentional models is if all these systems have the same higher-level mechanism generating their behavior (one that is instantiated by different lower-level mechanisms). But

this assumption is neither necessary to explain the successful use of intentional language across systems, nor supported by our best empirical evidence.

Consider the empirical problems with this sort of position. The suggestion that different physical processes in a variety of different systems can instantiate the exact same functional state is far more problematic than it intuitively seems. As Bechtel points out, "if one uses a fine-grained account of both mental and neural processes, there is no evidence of the same mental state being realized in different ways" (Bechtel 2007, p. 173).[29] Moreover, we know that there is a tight connection between implementation and function, meaning that two systems that are implemented in very different ways will often be unable to have functional states that are identical in all the relevant respects. As Eliasmith notes:

> ...two implementations of a given functional description can not be usefully considered equivalent unless they are almost *identical*. This is because an algorithm running on one implementation can only be run by another implementation with the addition of an *emulator* program of some sort. Running this emulator adds computational complexity making the second implementation significantly different from the first. It is *only* in the limiting case of *infinite* symbol-strings that this overhead can be ignored (a limiting case often adopted by Turing machine proofs). For *finite* strings, this overhead will significantly affect the performance of the computer –especially if we place *time* and *resource* restrictions on its behavior. (2002, pp. 4-5)

And so this position is, at the very least, empirically problematic. More importantly however, this sort of account need not be true in order to explain the successes of intentional models applied across multiple systems.

As I have argued, the ability of intentional models to apply across a wide range of mechanistic systems is due to the fact that, as a type of phenomenological model, they are implementation

[29] Bechtel *does* note that if we describe neurological processes abstractly enough, we do find course-grained categories that apply across systems. But he proposes that this is mostly just a useful *heuristic* we can employ to help us characterize similarities between mechanistic systems (Bechtel and McCauley 1999), and is not a means by which radically different systems instantiate the exact same intentional states through a higher-level mechanism.

independent. Putnam and Fodor assume that the predictive successes of intentional models imply that they must refer to actual causal states of systems that generate behaviour. And so different systems described by intentional models must all be instantiating these identical causal states in different ways. But there is no reason to assume that this must be the case. There is an implicit intuition that the scientific respectability of intentional descriptions are contingent on their denoting causal spatiotemporal regions of the brain. But this is hardly true (as our use of statistical models and dynamic models demonstrates).

I suspect that there is an ontological motivation for this implicit intuition that is a reactionary response to substance dualism. The assumption is that if intentional states are not causal spatiotemporal chunks of physical matter, then they must be either *mental* objects adrift in some nebulous mental realm somewhere, or else they are mere fictions to be discarded once our science is complete. But, of course, neither of these options need be the case, as I argued in sections 8.1 and 8.2. Not all intentional attributions must coincide with sub-personal physical mechanisms in order for the descriptions to be integral to our scientific understanding of the mind.

Even if one chooses to adopt an ontology in which the only things that are "real" are causally interacting chunks of physical matter, this does not imply that only descriptions which denote such things have a respectable role in scientific practice. If it did, then large swathes of contemporary physics would be abandoned. This would be to make the mistake I highlight in Section 2.3: to insist that scientific methodology should conform to fit one's chosen ontology as opposed to using the fruitful and successful methodological tools it has at its disposal. Putnam and Fodor's assumption that intentional states must be brain states in order for intentional models to be successful is simply not well grounded.

And so the important insights brought to us by Putnam and Fodor are: 1) The fact that neuroscientific mechanisms do not always coincide with intentional attributions does not make such attributions irrelevant to our scientific accounts of the mind; and 2) intentional language applies *across* mechanistic systems.

9.3 Anomalous Monism

Donald Davidson, in his 1970 paper "Mental Events", offers a theory of intentional language that manages to overcome some of the problematic assumptions made by Putnam and Fodor, while unfortunately falling prey to others. To understand Davidson's position, however, we must first change the way we think of the relationship between intentional and mechanistic models. Instead of talking in terms of physical mechanisms and intentional states, Davidson talks in terms of physical and mental *events*. An event is "physical" when it falls under a physical description (one employing physical terminology), and "mental" when it falls under a mental description (one employing intentional terminology). Under this interpretation, mechanistic interactions are understood as physical events, while the having of intentional states are understood as mental events. According to Davidson, there is no dualism regarding the sorts of events that exist, only the sorts of descriptions we use to characterize them. Thus any event that we describe as a mental event can also be described as a physical event.

Davidson's theory, dubbed *Anomalous Monism* (AM), proposes that when we speak of behaviour being caused by intentional states (beliefs, desires, etc.), we are attempting to characterize the causal interactions of an event described under a mental language (the having of a belief for instance) with the event described under a physical language (the resulting behaviour).

Davidson proposes that all causal interactions between events fall under strict nomological laws or regularities. However, the causal interactions between mental events and physical events do not fall under such regularities. The idea being that the law-like relations that link events by way of cause and effect are only identifiable under a strictly physical description. Intentional descriptions, in contrast to physical descriptions, do not describe physical events in a manner that allows us to subsume them under strict laws. Put simply, intentional descriptions are a different means of characterizing causal events, and they do so in a way that obscures the nomological relations that causally connect them to physically described events.

In this respect, Davidson proposes that AM can be unproblematically committed to three *seemingly* contradictory principles. The first is that mental events can causally interact with some physical events (given that all mental events *are* physical events, and physical events can causally interact). The second is that all events are causally related by way of strict nomological laws. The third is that there are no strict nomological laws relating mental events to physical events (given that the strict law-like relation connecting events can only be seen under a physical description).

While Putnam and Fodor want to make intentional language autonomous from the mechanistic language of the lower-level sciences (like neuroscience), they are still committed to the idea that intentional language characterizes the behaviour of higher-level physical mechanisms. In this sense, intentional descriptions still act as abstract mechanistic descriptions of systems. This idea is one I've challenged on the grounds that intentional models predict the behaviour of systems in a way quite distinct from the way in which mechanistic models do. In this respect, mechanistic models and intentional models should not be confused for one another. AM better captures this distinction between intentional language and mechanistic language.

While Davidson talks exclusively in terms of the relation between mental and physical events, as opposed to the relation between mechanistic and intentional models, we can draw a parallel between the two cases. When mechanistic models are used to explain and predict the behaviour of systems, they do so by characterizing the system in terms of the causal relations between different physical objects that make up the system. Meanwhile, when we switch to an intentional description, we describe the system in a decidedly non-mechanistic way. The causal interactions that go on between the constitutive parts of a mechanism are what necessitate that mechanism's behaviour. However, the interactions of intentional states we attribute to a given system do not necessitate that it will act in any given way. Intentional models presuppose that a system is rational, which is almost always an idealization. Thus the strict cause and effect we find when we model a system mechanistically is not something we find when we model the

system intentionally. This captures Davidson's point that the necessity of a system's behaviour can be better understood given a particular kind of description of the system.

Davidson also argues that we should not expect to be able to reduce mental language to physical language. As he puts it, "no purely physical predicate, no matter how complex, has, as a matter of law, the same extension as a mental predicate" (2002/1970, p. 123). Since mental and physical languages characterize events in drastically different ways, there is no way to capture the meaning of a mental sentence by using a strictly physical sentence. I too propose that there exists an irreducibility between intentional and mechanistic models, given that intentional models are implementation independent, while mechanistic models are not.

Unfortunately, there are problematic assumptions with AM that ultimately lead Davidson astray. First, he suggests that there are two major sorts of descriptions to be contrasted: mental and physical (the latter being subsumed under strict nomological laws or regularities). But this idea does not do justice to the subtle and important differences that exist *within* our so-called physical descriptions. There are a great many different kinds of physical descriptions (those of physics, chemistry, biology, etc.), and not all of these different physical languages characterize events in a way that subsumes them under strict nomological regularities. Consider, for instance, our use of mechanistic descriptions in neuroscience. While we commonly use mechanistic descriptions to characterize causal neurological events, it does not follow that these neurological events are connected by strict nomological regularities. As Bogen points out,

> If the production of an effect by activities which constitute the operation of a mechanism is what makes the difference between a causal and a non-causal sequence of events, Mechanists need not include regularities and invariant generalizations in their account. (2005, p. 399)

The assumption that causality between two events, when characterized by a physical language, can be characterized as falling under nomological regularities is simply not something we can take granted. To demonstrate, consider the following example offered by Bogen:

> The mechanisms which initiate electrical activity in post-synaptic neurons by releasing neurotransmitters are a case in point. They are numerous enough, and each of them has enough chances to release neurotransmitters to support the functions of the nervous system. But each one fails more often than it succeeds, and so far, no one has found differences among background conditions which account for this (Kandel, Schwartz, & Jessel, 2000, p. 261). No one takes the irregularity of their operation as a reason to deny that on the relatively rare occasions when they do operate successfully these mechanisms release neurotransmitters and exert a causal influence on post synaptic neuronal activity (2005, p. 399).

Thus we are left claiming either that, contra Davidson, physical descriptions (like neuroscientific descriptions) do not characterize nomological regularities between events, or else that neuroscientific language *is not a physical language*. Indeed, given that many different domains of the life sciences primarily explain by way of mechanisms, and the fact that mechanistic explanations do not depend on (or necessarily identify) nomic regularities between events, Davidson's definitional criteria of physical languages is much too crude.

Instead of a single physical language, we have a multitude of physical languages that emphasize different aspects of systems. Some of these descriptions may characterize nomological relations that hold between events, while others may not. Thus insisting that the lack of nomic regularities connecting a set of intentional states to the behaviour of the system is due to the fact that we are attempting to characterize the same event under two different linguistic frameworks is not obvious (since even within a physicalist linguistic framework we may not find such nomic regularities). Similarly, we have little reason to think that intentional states must always coincide with physical events.

To emphasize this point, consider that throughout this dissertation, I have spoken of intentional language as a particular means of modeling systems; one to be contrasted with other sorts of models (such as mechanistic models). Davidson, meanwhile, does not speak in terms of different models, but instead in terms of different means of *classifying events.* The suggestion being that any event characterizeable under a mental description is always characterizeable under a physical description. Thus leaving us with a token-identity theory of the mind (every token mental event is co-extensive with some token physical event). But this sort of position can be understood in one of two ways:

1. Given that we use mental language to characterize physical systems (such as people, or animals), we can always carve the system into *some* set of to-ings and fro-ings (e.g., some set of physical events) that we stipulatively make co-extensive with our mental events.

2. There is an ideal non-arbitrary way of carving the system into physical events, and an ideal non-arbitrary way of carving the system into mental events. It is then an empirical fact that every mental event picked out by a mental description will always happen to be co-extensive with a physical event.

I propose that (1) is trivial. The fact that we use intentional language to describe physical systems means that this will always be the case, and so in an important sense it is uninformative. Meanwhile, I have argued at length in this dissertation that something like (2) is not something we should expect to be the case. This sort of move falls into the same trap that Putnam and Fodor fall into: assuming that the only way to make mental descriptions respectable (either scientifically or metaphysically) is if intentional states always denote some causal region of space-time. But this hardly needs to be the case. The successes of intentional models do not require that intentional states denote physical events. And even if Davidson's

account turns out to be true, it is hardly an *a priori* truth. He has no grounds at present to *stipulate* that this must be the case.

And so the important insights brought to us by Davidson are: 1) Since intentional and mechanistic descriptions characterize systems in drastically different ways, we should not expect to reduce one type of description to another; and 2) the predictions we make via intentional language do not tell us what the behaviour of a system will *necessarily* be, given that this linguistic framework does not characterize the sorts of causal interactions that necessitate its behaviour. In order to identify such relations, we need to change the language we use to describe the system. This nicely captures the contrast between the idealized predictions of intentional models with the causal-based predictions of mechanistic models.

9.4 The Co-Evolutionary Research Ideology

While the different philosophical accounts mentioned above capture some important aspects of intentional language, none have emphasized the important interactions that go on between intentional descriptions and other sorts of descriptions in a scientific context. Patricia Churchland's *co-evolutionary research ideology* highlights the interplay between intentional models and mechanistic models in our study of the mind (1989). Churchland proposes that there is a co-evolution between intentional models in psychology and mechanistic models in neuroscience. The two inform one another, and so grow and change together:

> In these instances discoveries at one level often provoke further experiments and further corrections at the other level, which in turn provoke questions, corrections, and ideas for new explorations. (Churchland 1989, pp. 363-364)

As intentional psychology and mechanistic neuroscience grow together, each will be revised, and their concepts will be reorganized and restructured, in order for them to converge on a singular account of the phenomenon. In this way, Churchland proposes that a reductive account of intentional psychology to

mechanistic neuroscience (after the appropriate modifications to our theories based on co-evolution are made) "is more or less inevitable" (1989, p. 374). As she puts it:

> As long as psychology is willing to test and revise its theory and hypotheses when they conflict with confirmed neurofunctional and neurostructural hypotheses, and as long as the revisions made with a view to achieve concord with the lower-level theory, then the capacities and processes described by psychological theory will finally find explanations in terms of neuroscientific theory. (The same is true, of course, for the revisions and reconstructions in neuroscience.) (1989, p. 374)

This convergence means that, at the end of inquiry, there will no longer be a one-to-many mapping between intentional models and mechanistic models. The only reason that the same intentional model can apply to different mechanistic systems, according to Churchland, is because we have not refined our intentional categories sufficiently enough yet for them to account for the mechanistic differences that exist between the different systems. As the intentional and mechanistic accounts co-evolve together, the intentional concepts that apply across systems ("beliefs", "desires", "mental representations", etc) will fragment into concepts that are always mapped one-to-one onto the particular mechanisms of a given system. As she puts it:

> Now from the reductionist viewpoint, this possibility [that we will be unable to find a one-to-one mapping of intentional psychological categories to neurobiological categories] does not look like an obstacle to reduction so much as it predicts a fragmentation and reconfiguration of the psychological categories. Indeed, there are already signs of a fragmentation of the folk psychological category of memory. (1989, p. 365)

According to Churchland, if we think that different mechanistic systems can have the same intentional state (e.g., a belief that *x*), then this is not an argument for the irreducibility of intentional

states to neurobiological ones, but instead suggests that we must simply revise our intentional categories to better fit with the more fine-grained mechanistic categories that we use to differentiate these systems.

The benefit of Churchland's co-evolution account is that it nicely captures the important methodological value of employing both intentional and mechanistic models in our study of the mind, and the important interdependence that exists between different types of models in scientific practice (see Section 8.1). The application of intentional models in psychology is an important step in our understanding of neurological mechanisms:

> Crudely, neuroscience needs psychology because it needs to know what the system does; that is, it needs high-level specifications of the input-output properties of the system. Psychology needs neuroscience for the same reason: it needs to know what the system does. That is, it needs to know whether lower-level specifications bear out the initial input-output theory, where and how to revise the input-output theory, and how to characterize processes at levels below the top. (Churchland 1989, p. 373)

If what we care about is representing or characterizing, in detail, the mechanisms that constitute a particular system, then an intentional model (as a phenomenological model) will provide us with a place to start by characterizing the system's general behaviour. This points us towards possible mechanisms. As our knowledge of the mechanisms improve, the intricacies and details of the system's behaviour become more recognizable, allowing our phenomenological models to become more fine-grained and accurate. As our models inform one another and evolve together, our broad intentional concepts (like "beliefs") may be fragmented into multiple different intentional concepts in order to better account for the mechanistic distinctions we find within the system, and the mechanistic differences that exist between systems. Eventually, as we progress, this fragmentation of intentional concepts can result in a highly-detailed and fine-grained phenomenological model that will map one-to-one onto the mechanistic model of a target system.

The problem with Churchland's account is her insistence that there will *always* be this reduction of intentional models to mechanistic ones after the appropriate tweaks to both models are made. Such a claim disregards some of the key scientific virtues of employing phenomenological models. Even if some intentional models converge with mechanistic ones when we are interesting in providing a detailed account of a particular system's mechanisms, science is not *only* in the business of describing such mechanisms. Recall Batterman's point that a broad goal of scientific theorizing is to recognize "observed patterns in the behavior of systems of a given type" (Batterman 2000, p. 120). In such cases, a phenomenological model which *ignores* many of the fine details of the system may be a better candidate for charactering these regularities than one which is detailed enough to map one-to-one onto a mechanistic account of the system (see Batterman 2002a, 2002b; Fillion 2008).

This means that Churchland is incorrect when she claims that a fragmentation of intentional concepts in order to better fit with the particular mechanisms of a given system is inevitable, or that a reduction of intentional concepts to mechanistic ones should be our desired outcome. Such a fragmentation of intentional concepts in order to facilitate reduction may be extremely useful in some contexts, but can actually be *undesirable* and *problematic* in others, depending on what our pragmatic purposes are for modeling the system. Very broad (unfragmented) intentional categories may be beneficial in some circumstances precisely *because* they apply across systems, allowing us to see commonalities that exist between them. Similarly, broad patterns in human behaviour may only be identifiable when we adopt a model which leaves out many of the details of the system.

This means that, despite Churchland's instance to the contrary, the one-to-many mapping of intentional models to mechanistic models is not always a vice that must be overcome through sufficient refinement and co-evolution with mechanistic neuroscience. It can often be a virtue given that it is used to model behavioural regularities that mechanistic models might miss altogether, and is used to identify regularities that exist across a range of mechanistic systems. One of the advantages of phenomenological

models is that they are not mechanistic models, and so are applicable to, and useful in, different scientific contexts. Adapting and altering all intentional models so that they *become* detailed mechanistic models, as Churchland suggests, ignores the very reason we employ phenomenological models in many scientific contexts.

Another way to characterize this problem is to realize that intentional models and mechanistic models are radically different means of characterizing the input/output relation of a system. Depending on our pragmatic reasons for characterizing this relation, some phenomenological models will be beneficial the more detailed and fine-grained they become (which may eventuate in a convergence with a mechanistic account of the system), while others will be more useful when their account of the relation is more abstract. As Betterman notes, "often times this 'details is better' approach is misguided" (2002a, p. 21).

If what we care about is identifying and understanding the complex mechanisms that underlie a certain system's behaviour, then an extremely detailed account of the input/output relation will be ideal, and a convergence of mechanistic and phenomenological models will be extremely beneficial. On the other hand, if we are uninterested in the specific details of a given system, and instead want to identify very broad patterns of behaviour, then a much more abstract account of the input/output relation may be preferable. In such a context, a convergence between our intentional and mechanistic models would be unhelpful, since the more detailed account of the input/output relation would be *less* useful for seeing the patterns we care about.

We adopt models for different pragmatic purposes, and there are distinct advantages to employing a more abstract characterization of the input/output relation over the more fine-grained accounts. Therefore, the broad intentional categories of traditional psychology may have important scientific value *in virtue of being* broad intentional categories that do not map onto the detailed mechanisms of a system.

And so the important insights brought to us by Churchland are: 1) That we need to adopt multiple scientific models (intentional models from psychology and mechanistic models from neurosicence) in order to fully explain and understand the mechanisms of the brain; and 2) that in many important cases, these models inform one another, change, and co-evolve together.

9.5 The Intentional Stance

Lastly, let us consider Daniel Dennett's *Intentional Systems Theory*. Of the accounts discussed in this chapter, Dennett's is closest in spirit to the account of intentional language that I provide. According to Dennett, intentional language works as a predictive framework (see: Dennett 1971, 1987, 1991a, 1991b). This framework, which he calls the "intentional stance", employs intentional descriptions to predict how different kinds of systems will behave. Dennett contrasts this predictive strategy with two others: the "physical stance", and the "design stance". The former predicts systems based entirely on their physical construction and the laws acting on them, while the latter makes predictions based on the system's designed function.

According to Dennett, we switch between these three predictive strategies –or stances– in order to predict systems of various complexities. Predictions made via the physical stance "are based on the actual physical state of the particular object, and are worked out by applying whatever knowledge we have of the laws of nature" (Dennett 1971, p. 88). Dennett suggests that the physical stance is always effective, but not always convenient. It may be useful for determining the behaviour of simple systems, but it does us little good when we need to make quick predictions of extremely complex systems. It is far too cumbersome and impractical for such purposes.

Of course, we often need to make predictions of systems when we simply have no practical means of employing the physical stance. In some of these situations, Dennett proposes that we can use a different predictive strategy: *the design stance*. With this stance, we can make predictions of complex systems based on how the system is designed. For instance, we do not need to know the internal parts of a

computer, a television, or a car, to predict how such objects will behave, since we know how they are *designed* to behave. As Dennett tells us,

> …almost anyone can predict when an alarm clock will sound on the basis of the most casual inspection of its exterior. One does not know or care to know whether it is spring wound, battery driven, sunlight powered, made of brass wheels and jewel bearings or silicon chips –one just assumes that it is designed so that the alarm will sound when it is set to sound. (1987, pp. 16-17)

Unlike the physical stance, not all systems are going to be predictable by taking the design stance; only systems which can be said to have a proper function. Similarly, there are constraints on when the design stance applies to a given system. If our alarm clock gets physically damaged for example, we will not be able to accurately predict it with the design stance anymore, and will instead have to revert to the physical stance (Dennett 1987, p. 17). These conditions aside, a great many systems are predictable using the design stance.

There are times, however, when we simply do not have the luxury of knowing either a system's proper function, or what parts it contains. In such situations, we can still make accurate predictions about the system without trying to understand its underlying parts or its design. To do this, we can switch to the *intentional stance*. When we take this stance,

> …we must treat the [system] as an agent, indeed a rational agent, who harbors beliefs and desires and other mental states that exhibit *intentionality* or 'aboutness,' and whose actions can be explained (or predicted) on the basis of the content of these states (Dennett 1991b, p. 76).

Certain systems (which he terms "intentional systems") display a genuine pattern of behaviour that can be predicted by attributing to them an appropriate set of intentional states. While human beings are the most obvious example of intentional systems, Dennett tells us that there are many non-human

examples to choose from as well. The intentional stance can effectively predict the behaviour of everything from fish, to birds, to reptiles, to insects and spiders, to clams, and to computers (Dennett 1987, p. 22).

Dennett's idea that we can adopt different "stances" for the purposes of prediction captures quite nicely the idea that we employ different types of models in scientific practice to predict and characterize systems. Similarly, his distinction between the physical stance (which predicts systems based on their structural properties and interactions) and the intentional stance (which makes predictions without appealing to structural features of systems, but by identifying patterns in behaviour that the physical stance cannot identify) nicely captures the distinction between mechanistic models and phenomenological models as I have discussed them here. Dennett similarly points out that intentional models can be used to predict a wide range of mechanistic systems while also noting, contra Putnam, Fodor, and Davidson, that the intentional states referred to in these models need not correspond with physical states (or events) in order to be predictive.

It is in many of the important details that Dennett and I part ways however. Before going into these details, it should be noted that given how prolific the corpus of Dennett's work is, he can often be found saying conflicting things at different points. As a result, I cannot say with certainty which parts of his previous work he is currently committed to. That being said, my intention is not to determine what Dennett currently believes, but simply to highlight some of the claims made by Dennett that do not cohere with the sort of account I have offered in this dissertation, and to demonstrate how those claims lead to problems.

The first thing to note about Dennett's account is his assumption that the physical stance is one monolithic predictive strategy. The assumption being that there is the intentional stance, the design stance, and then *the stance that covers everything else* (the physical stance). This makes the same mistake that Davidson makes in ignoring the huge amount of variation that exists within physical descriptions (or,

in Dennett's terminology, within the physical stance). While Dennett's distinction between stances highlights the fact that we do indeed adopt different perspectives in order to make different sorts of predictions of systems, his distinction between stances is simply much too simplistic.

To demonstrate, consider that there are some types of scientific models that fall under none of Dennett's stances. Take, for instance, our use of statistical models to generate predictions. What stance do they function as? According to Dennett, if you want to predict the behaviour of a system via the physical stance, you must "determine its physical constitution (perhaps all the way down to the microphysical level) and the physical nature of the impingements upon it, and use your knowledge of the laws of physics to predict the outcome of any input" (Dennett, 1987, p. 16). Yet statistical models say nothing about the physical constitution of the system, nor how the fundamental laws of physics act upon it. As a result, they seem not to predict based on the tenets of the physical stance. However, they similarly do not make predictions based on the target system's design (excluding it from the design stance), nor by appeal to intentional states (excluding it from the intentional stance). And indeed, this is true of *most* phenomenological models. In this respect, Dennett's distinction between *predictive stances* is greatly problematic.

The second problem with Dennett's account is his commitment to the idea that characterizing a system in terms of the objects and causal laws that make it up will *always* yield correct predictions. He grants that predictions via the physical stance may be *pragmatically* unfeasible, but insists that they always work in principle. As Dennett puts it, "[the physical stance] is not always practically available, but that it will always work *in principle* is a dogma of the physical sciences" (Dennett 1987, p. 16, emphasis in text). But this is untrue, and betrays Dennett's confusion between scientific representations, and the physical system being represented.

The dogma of the physical sciences is that the existence of physical objects and laws is sufficient to explain the existence of all the physical phenomena that supervenes upon them. It is not that

descriptions which *only employ terminology that makes reference to* physical objects and laws are sufficient for all our predictive scientific needs (with practical considerations being the only limitation). These are very different claims. In other words, Dennett's implicit assumption is that if we know everything there is to know about the physical objects that make up a system, and the laws acting on them, then we can predict what it will do next without fail. However, this is very different from claiming that *adopting the physical stance* will allow us to predict what a system will do next without fail. The obvious inference that Dennett makes is that the physical stance can, at least in principle, tell us everything there is to know about the objects that make up a system and the laws acting on them. But this may simply not be possible given the nature of scientific representations. Scientific representations, even in our most fundamental physics, might always involve distortions, idealizations, and abstractions (thus making it impossible for any single "stance" to always be predictively successful of all phenomena).

To demonstrate, consider our use of the fundamental theories of statistical mechanics to predict phase transitions, like water turning from liquid to ice:

> The problem is that phase transitions —as understood by statistical mechanics— can only occur in infinite systems, yet the phenomena that we are trying to explain clearly occur in finite systems. (Callender 2001, p. 549)

Batterman highlights this problem as well, claiming that:

> From the point of view of the underlying fundamental theory whose proper focus is on the interactions of a *finite* number of molecular components of the macrosystems, these qualitative changes are genuinely novel. The upshot is that the statistical mechanics of finite systems is explanatorily insufficient. While it gets the ontology of blobs of gases and fluids right (they are composed of a finite number of interacting molecules), there remain macroscopic phenomena — universal patterns of behavior— that cannot be explained by this fundamental theory. (2011, pp. 1033-1034)

Put simply, the only way for statistical mechanics to account for phase transitions is by interpreting the system as having an infinite number of molecules (which we know it does not have). According to Batterman, this demonstrates that even our fundamental theories in physics are importantly incomplete in the way they represent systems. If this is true, then it poses a serious problem for Dennett's insistence that the physical stance is universally successful. Recall that the physical stance predicts by characterizing the physical objects that make up a system, and the laws acting on them. However, we *know* that we cannot predict phase transitions by characterizing all the finite particles that make up a system, and the laws of statistical mechanics that act upon them. Such a prediction will always be unable to predict and account for phase transitions. Instead, we *must* idealize the system to get the predictions to come out right.[30]

To further emphasize this point, consider once again the analogy discussed in section 8.2 between scientific representations and maps. Dennett often implies that the physical stance is much like map of the world that leaves out no details at all. While the intentional stance predicts by characterizing patterns in behavior, the physical stance predicts by characterizing everything about the system.[31] While he grants

[30] Of course one might insist, as Callender (2001) does, that this may simply show that the laws of thermodynamics are false (or not something we should take *too* seriously). But where does this leave Dennett? Such laws clearly cannot be part of the physical stance if Dennett insists that the stance will *always* yield true predictions. And so does this imply that only some future description of the ultimate laws of physics can count as part of the physical stance? If so, then virtually nothing we have at present counts as an application of the physical stance. And, in fact, this means that all our current scientific models have more in common with the *intentional* stance than they do with the *physical* stance.

[31] We can see strong shades of this idea in Dennett's 1991 paper "Real Patterns". He tells us to imagine a virtual universe characterized by the "Game of Life" computer program. This universe consists of a two-dimensional grid in which any square on that grid can be either empty or filled. There are also rules for whether a square should change from empty to filled, or filled to empty, depending on the state of the surrounding squares. Dennett tells us to imagine that, given the rules of this virtual universe, it can be used to simulate a Turing-machine; one that is running a program of chess in which two AI-controlled players face off against each other. He tells us that adopting the physical and intentional stances allow us to make predictions of what will happen in this universe in different ways. While the intentional stance predicts by interpreting the rational moves of the AI-controlled chess players instantiated by the simulated chess program (i.e. by identifying a genuine pattern of behaviour in the workings of the system), the physical stance predicts by characterizing the entire universe bit-by-bit and making predictions based on the rules applied to the system. In other words,

> ...when we adopt the physical stance toward a configuration in the Life world, our powers of prediction are perfect: there is no noise, no uncertainty, no probability less than one. Moreover, it follows from the two-

that taking the physical stance makes it impossible to *see* the patterns in behaviour used by the intentional stance to form predictions (which is a position that has much in common with what I advocated here), he is still committed to the idea that physical stance can always predict in virtue of characterizing *everything* about the system, and so it carries with it the patterns in behaviour that the intentional stance identifies (even if such patterns cannot be seen *as* patterns). He proposes that it may not always be feasible to use such a detailed "map" of the world, but it will always get you to where you need to go.

The problem with this idea is that a map which contains all the details of the world is not simply impractical, it literally *cannot be used as a map*. In other words, a map that contains all the details of the world is no map at all. In order to be of any use at all for navigating it must, by necessity, focus on the particular things we are interested in finding, and in presenting the information in a form that we can use. Otherwise, the map is no different from the world we are seeking to navigate through. This is precisely why Truesdell points out that "a theory cannot duplicate nature, for if it did so in all respects, it would be isomorphic to nature itself and hence useless, a mere repetition of all complexity which nature presents to us, that very complexity we frame theories to penetrate and set aside." (1980, p. 72) A similar point is made by Kellert et al. (2002), who point out that "all representations are partial in that any representation must select a limited number of aspects of a phenomenon (else it would not represent, but duplicate)" (p. xv).

The moment we adopt the physical stance *for the purpose of prediction*, we automatically have reasons to focus on describing some aspects of the world at the cost of others. This may necessitate distortion and abstraction in order to identify and analyze the features we care about. Dennett proposes that the intentional stance predicts by characterizing patterns in behaviour, while the physical stance predicts by characterizing everything about the system. But this may be impossible in principle. The

---

> dimensionality of the Life world that nothing is hidden from view. There is no back-stage; there are no hidden variables; the unfolding of the physics of objects in the Life world is directly and completely visible. (1991a, p. 38)

physical stance, like the intentional stance, might equally predict by characterizing patterns in the world. They just characterize *different* patterns. And even *within* the physical stance, we find different physical descriptions characterizing different patterns (those of neuroscience, those of chemistry, etc.). There might be no stance, or model, that captures everything about a system in the way Dennett supposes that the physical stance does. In which case, the differences between the intentional stance and the physical stance become much less pronounced and significant than Dennett insists they are. Moreover, the intentional stance becomes more like a *part of* the physical stance (a descriptive framework that characterizes real patterns in the world), as opposed to something else altogether.

And this brings us to the final problem with Dennett's account that I wish to discuss: his paradoxical view regarding the scientific merits of the intentional stance. On the one hand, Dennett seems to grant that intentional descriptions are present in scientific practice when he claims that "the decision to adopt 'the intentional stance' is not an unusual sort of decision in science" (1987, p. 239). On the other hand, he also warns us that we should not take the intentional stance "*too* seriously" given that it is merely a "heuristic overlay"(Dennett, 1987, p. 350, emphasis in text). He is also quick to point out that "Intentional theory is vacuous as psychology because it presupposes and does not explain rationality or intelligence" (Dennett 1971, p. 99). Because intentional descriptions cannot provide physical explanations for the theoretical objects they posit, Dennett proposes that their role in science is best thought of as a heuristic rule of thumb.

The problem with this sort of view is that it hangs a lot on the assumption that respectable scientific accounts always provide explanations, and that such explanations are always compatible with the implicit reductionism of the physical stance. Consider our use of mechanistic models to explain phenomena in the life science. Attempts to provide explanations in terms of covering-laws have failed in the life sciences, since no such laws seem to exist. But if Dennett is right, and neuroscience is part of the physical stance, then neuroscience is really just a shorthand for physics. In which case, appeals to

mechanisms are not truly explanatory, since explanations in neuroscience must reduce to explanations in physics, which are primarily *covering-law* explanations (and mechanistic models in neuroscience do not identify or characterize the universal physical laws that govern physics). And so just as intentional theory is vacuous as psychology because it presupposes and does not explain rationality or intelligence, so too would mechanistic neuroscience be vacuous as physics because it presupposes, but does not explain, the fundamental laws of subatomic physics that constitute macro-level neurological phenomena. Does this mean that, like the intentional stance, neuroscience is thereby just a useful heuristic account of physics that we shouldn't take *too* seriously? If so, then the intentional stance is on the same scientific footing as our best explanatory accounts in all the life sciences.

And what if neuroscience does not cleanly reduce to physics? Should we similarly conclude that neuroscience must just be a heuristic overlay not to be taken seriously given its irreducibility to physics? Is physics the only science that we are ever allowed to take seriously, and thus all descriptions must be unified under the banner of physics at the cost of being dismissed as merely a heuristic device? This sort of unificationism under physics may appeal to some, but there is no scientific reason to think it must be the case. As Ian Hacking points out, "the unity of science is rooted in an overarching metaphysical thought that expresses not a thesis but a sentiment" (Hacking 1996, p. 44). I propose that even if there is no clean reduction from neuroscience to physics, this does not mean that we should no longer take neuroscience seriously. The value of neuroscience is in helping us learn about the world in a way that the formalism and models of physics are not ideal for. As Alan Richardson points out:

> Suppose there is a clear sense in which the quantum formalism of physics does not quite match the quantum formalism of chemistry –does this hinder cooperation and sharing of theoretical knowledge? When? How? (2002, p. 19)

I similarly propose that the essential methodological value in employing different models in science is substantial, irrespective of whether they reduce to one type of model or not. And this is a good reason to take such models very seriously. In this sense, we have no reason to dismiss the intentional stance as something we ought not take too seriously without potentially dismissing a great many of the descriptions and models that are part of the physical stance as well. Dennett's desire to cleave intentional descriptions away from every other sort of scientific description requires that he show us what makes those accounts less well grounded than other models in science. And as I've shown, the value of intentional models to science is substantial.

And so the important insights brought to us by Dennett are: 1) That we can, and do, use multiple interpretations of systems in order to make predictions in different ways; 2) Our use of intentional descriptions acts as one such predictive interpretation; and 3) we have no reason to expect that the objects in intentional descriptions will necessarily correspond to causal states within a system.

9.6 Historical Insights

We can see that the history of philosophy is rich with insights as to the role of intentional language in science. While different philosophers espouse very different sorts of stories regarding the nature of intentional language, each captures something important. The eliminativist highlights the fact that intentionality may be a product of the language we use to describe systems, while the functionalist points out that such language applies to a wide range of mechanistic systems. The anomalous monist emphasizes the important differences between intentional language and mechanistic language, while the co-evolutionary research ideology describes the important ways they interact. Lastly, the intentional stance theorist calls attention to the predictive value of intentional language. Together, each gets a piece of the puzzle right, and points us towards a story that best fits with scientific practice. While these accounts each have weaknesses, the story I've presented in this dissertation overcomes these weaknesses by keeping in mind important historical lessons while keeping a firm eye on actual scientific methodology.

Chapter 10
Conclusion

It is now finally time to take stock of things, and to make some concluding remarks. I will begin by highlighting some of the key lessons learned in our examination of the scientific merits of intentional language. Then, once this is done, I will at long last turn to questions of ontology and metaphysics. What does my account say about the realism of intentional states? Are beliefs and mental representations *real*, or just a pragmatic and instrumentally useful way of talking? I have resisted discussing metaphysics up to this point so as to avoid letting the metaphysical tail wag the methodological dog. But with my general story in place, I will finally be able to explore the ontological implications of my account.

10.1 Lessons Learned

So what are intentional states? Intentional states are theoretical objects that are part of a linguistic framework that we use to model the behaviour of systems. These intentional models are holistic (they always require multiple intentional states in order to account for a system's behaviour), and are used to generate predictions. Given that not all predictive devices in science are useful in the same contexts, or generate predictions based on the same set of data, intentional models have their own set of pragmatic benefits.

Intentional descriptions function as a type of phenomenological model. Such models are ideal for making predictions when we do not know the underlying mechanisms of the system that produce its behaviour. Similarly, intentional models allow us to predict even if we are unable to quantify over the system, making them ideal for predictions in contexts in which other sorts of phenomenological models (such as statistical and some dynamical models) cannot be easily generated. We also use intentional models to see similarities in behaviour that exist across a wide range of different mechanistic systems. This information is invaluable in helping us learn about the unknown mechanisms that underlie a

system's behaviour. In this way, intentional models play an essential methodological role in our scientific understanding of the neurological mechanisms that constitute the mind. Moreover, given that different types of models can often be required for representing different aspects of phenomena, intentional models may turn out to be the only sorts that can identify very particular kinds of regularities and patterns in the behaviour of systems.

There is often a great deal of discussion in philosophy of mind about the nature of reduction. Can we reduce intentional descriptions to descriptions of neurological mechanisms? Can the principles of psychology more generally be reduced to the principles of neuroscience? The account I have presented here suggests that such reductionist projects often miss the methodological value of having different means by which we can represent systems. The value of intentional models does not come from their reducibility to other sorts of models, but from the fact that they represent systems in different ways, allowing us to overcome the gaps in knowledge we might have with other sorts of models. Similarly, the value of psychology does not come from its reducibility, or lack thereof, to neuroscience, but from the fact that psychology allows us to identify properties of neurological systems that helps *inform* our neuroscientific study of the mind. The fact that the concepts in one domain may not reduce to the concepts of another is irrelevant to the pragmatic value we gain from employing different fruitful representations of systems.

10.2 The Ontological Implications

But what does this say about *ontology*? The sort of methodological descriptivism I have on offer here still leaves open the question: are intentional states real, or just pragmatically useful? Ultimately the account I provide tells us very little about the ontology of intentional states. It is, in fact, compatible with varieties of both realism and anti-realism. I will highlight here just some of the ontological stories one can offer that would mesh well with the methodological account I provide. Instead of giving a treatise on the nature of realism (which would be impossible in the limited space remaining), I will instead just briefly highlight

a few ontological possibilities. Which of these ontological accounts we *ought* to adopt I leave to the reader.

Let us begin with a common ontological view regarding the realism of intentional states that can be seen in the works of philosophers such as Jerry Fodor (1974) and Hilary Putnam (1967). This is the idea that intentional states are real when they correspond to physical sub-personal mechanisms operating within a system. In other words, an intentional state is real if our intentional concept denotes a causally efficacious physical part of a system. This sort of realism often has ties to *nominalism* (the metaphysical view that only concrete particulars exist, and not abstract objects or universals). If we accept this interpretation of realism, then some intentional states will turn out to be real, while others will not (depending on the particular system and the particular intentional states we attribute to it). It is extremely important to note, however, that this would not change the methodological necessity of employing the so-called "unreal" intentional attributions in our discovery of the real ones. Nor would it change the fact that our use of unreal intentional attributions may be an unavoidable means of characterizing very real behavioural aspects of systems that other models may be unable to characterize.

While nominalism, in some form or another, strongly appeals to many philosophers, others have suggested that it does not necessarily fit comfortably with our scientific practices. Many versions of nominalism, for instance, commit one to an anti-realism regarding mathematical objects (since such objects are often considered to be abstract objects), and this does not do justice to the essential role that mathematics plays in scientific practice. As John Burgess notes:

> …almost everything that has come forth […] from the nominalist camp has represented the light-fingered larcenous variety, which helps itself to the utility of mathematics, while refusing to pay the price either of acknowledging that what mathematics appears to say is true, or of providing any reconstrual or reconstruction that would make it true. The usual label for this variety of nominalism is '[mathematical] fictionalism'. (Burgess 2004, pp. 18-19)

A similar worry is raised by Penelope Maddy, who notes that philosophers that are quick to label mathematical objects as "fictions" do so based on decidedly *metaphysical* considerations, not necessarily *scientific* ones. If one thinks that science ought to be our guide for determining what exists[32], then we must not make the mistake of stepping *outside* of science in order to metaphysically pass judgement on the objects posited by successful scientific endeavours. This sort of move involves comparing the objects *in* fruitful scientific practices (mathematical objects) to the *things that are really out there in the world* (the supposed "concrete particulars"). According to Maddy, this is exactly the wrong sort of move for us to make. If science is to be our guide for what is real, then we must work entirely *within* the confines of scientific practice, and so these extra-scientific comparisons are ill-conceived. As she puts it:

> In this humdrum way, by entirely natural steps, our inquirer has come to ask questions typically classified as philosophical. She doesn't do so from some special vantage point outside of science, but as an active participant, entirely from within. (Maddy 2011, p. 39)

The fact is that scientists do talk of abstract objects, such as mathematical objects, in their successful scientific practices. Because of this, Maddy proposes that these abstract objects are real. Of course, mathematical objects are not the same sorts of things as tables and chairs. Keeping in mind their *abstract* nature, Maddy proposes that we consider them to be "thinly" real (Maddy, 2011, pp. 60-83).

If intentional attributions are as methodologically important to science as I suggest, then they too might be thinly real. I have demonstrated that scientists can and do talk about intentional states in successful scientific practice. And this is the case even when the postulated intentional states do not appear to denote any particular physical structures. In this sense, they might be thinly real, and thus still deserve some place within our ontology.

[32] A version of Naturalism that Maddy subscribes to.

Others have proposed ontological accounts that share much in common with Maddy's. Dennett, for instance, similarly grants that intentional states are abstract objects, and also notes that abstract objects often play extremely useful roles in science. As an example, he highlights our use of *centres of gravity* in physics. He proposes that their relevance to scientific practice warrants being realists (of a certain sort) about them:

> …we should be […] more interested in the scientific path to realism: centers of gravity are real because they are (somehow) good abstract objects. They deserve to be taken seriously, learned about, used. If we go so far as to distinguish them as real (contrasting them, perhaps, with those abstract objects which are bogus), that is because we think they serve in perspicuous representations of real forces, "natural" properties, and the like. (Dennett 1991a, pp. 28-29)

Dennett then goes on to draw an explicit analogy between centres of gravity and intentional states, claiming that intentional states "are best considered to be abstract objects rather like centres of gravity" (1991a, p. 29). The analogy holds given that, like centres of gravity, intentional states are part of scientific representations (intentional models) that are used to characterize very real patterns that exist in the world (in the behaviour of certain systems). And so Dennett concludes that intentional states are real *in the same way that* centres of gravity are real.

Of course, Dennett's qualifications (he is a realist *of a certain sort* about centres of gravity; he thinks intentional states are *as real as* centres of gravity) have made him notorious amongst philosophers who feel that he is purposefully vague regarding his ontological commitment to the existence of such abstract objects. Dennett responds to these critics by pointing out that the realism/anti-realism distinction is itself greatly problematic, and that he does not wish to legitimize the unhelpful distinction by taking sides one way or another:

> I wouldn't want to trot out *my* ontology […] and then find I had to spend the rest of my life defending and revising *it*, instead of getting on with what are to me the genuinely puzzling issues –like the nature of consciousness, or selves, or free will. […] When and if professional ontologists agree on the ontological status of all my puzzle examples, my bluff will be well and truly called; I will feel a genuine obligation to make things clear to them on their terms, for they will have figured out something fundamental. (1993, p. 212)

That being said, Dennett has often made claims that seem to put him quite firmly on one side of the debate or the other. So, for instance, at some points he seems to advocate a position very similar to Maddy's, suggesting that scientifically fruitful abstract objects may not be the same *sorts* of objects as tables and chairs, but that they are still real in every way that matters to science. He tells us, for example, that intentional states are "*instrumentalistic* in a way the most ardent realist would permit: people really do have beliefs and desires, on my version of [intentional language], just the way they really have centres of gravity and the earth has an equator" (1987, pp. 52-53). He similarly states that belief "is a perfectively objective phenomenon" (1987, p. 15).

Meanwhile, these claims can be contrasted with others that would appear to put him unapologetically on the anti-realist side of the debate. In his paper *The Self as a Center of Narrative Gravity* (1992), for example, Dennett says that a centre of gravity is "a theorist's fiction. It is not one of the real things in the universe in addition to the atoms. But it is a fiction that has a nicely defined, well delineated and well behaved role within physics" (p. 103). In which case intentional states (in virtue of being *as real as* centres of gravity) are similarly fictional. We also find shades of this anti-realism in Dennett's suggestion that intentional descriptions are simply useful heuristics that we ought not take too seriously (see Section 9.5).

However, the fact that Dennett might not be completely sold on the brand of realism that he advocates in some of his writings does not mean that the account itself (or some modified version of it) is not worth holding. In fact, numerous philosophers have provided more consistent and convincing

arguments for the brand of realism that Dennett proposes (see, for instance, Kenyon 2000, and Ross 2000).

It is, however, also important to note that using science as our guide to ontology does not necessarily lead us to a realism about intentional states. Quine (1960) has suggested that a scientific posit having instrumental value is simply much too weak a criterion for ontological inclusion. Instead, he proposes that ontological inclusion be based on the indispensability of the theoretical posit to scientific practice. Only those objects that science, in principle, cannot do without are real. Under this interpretation, intentional states are far less likely to make the cut. But this is not surprising, since the vast majority of objects posited by science won't either.

The history of science may also lead one to a bleak outlook regarding the realism of intentional states. Those who subscribe to the *Pessimistic Meta-Induction* (see Putnam 1978; Lauden 1981) argue that throughout history, scientific theories may have become more successful, but they have almost all been proven false and displaced by better theories. Given this, we have no reason to assume that our *current* theories are in any better shape. We should therefore expect them to be overturned by future scientific practice as well. In which case, we should expect that the scientific virtues of intentional models will be similarly over-turned by future scientific practice (thus giving us good reasons to be anti-realists about the objects they describe). It should be pointed out that I am not arguing *for* the pessimistic meta-induction here, but merely pointing out that if one adopts it, then one would have good reasons to be sceptical of the existence of intentional states (for the same reason we should be sceptical of *all* current scientific objects).

Even if we do not use science as our explicit guide for ontology, there are many different ontological positions one can hold and still embrace the claims I have made in this dissertation. One might feel compelled by Carnap's view (1950) regarding linguistic frameworks, for instance. Very roughly, according to Carnap, the question of whether or not objects are real depends on the linguistic

framework we employ. And we employ different linguistic frameworks based on different pragmatic needs. Numbers, for example, are real so long as we are working within a mathematical language. On the other hand, if we are working within a language that makes no reference to numbers at all, then we have no reason to consider them real. Meanwhile, the question of whether numbers are real *outside of* any linguistic framework is ultimately an incoherent question for Carnap. This is because determining whether numbers are real or not requires understanding how we employ the concept "number". And it is the linguistic framework we are working in that determines how the concept is used (in virtue of dictating the assertability conditions for the concept). Thus, the question of whether numbers are real independent of any linguistic framework, and thus independent of any assertability conditions for the concept "number", is meaningless. Under this view, the question of whether intentional states are real depends on what linguistic framework we are employing. Intentional states are real just so long as we are working within an intentional language.

Or one might embrace something like pragmatism instead. In which case, the pragmatic benefits of intentional language will dictate the ontological inclusion of the objects it posits. Whether or not intentional models meet the required pragmatic threshold for being real is, of course, an issue for debate. But the pragmatist route is available for those who wish to take it.

I have even left the door open for the sort of intentional realism espoused by Searle. If there is some distinct ontological property of *aboutness* or *meaning* that we as biological entities possess but computers do not, then my account merely demonstrates that the scientific use of intentional language would not cut along the lines of this metaphysical distinction. The value of intentional language to science would therefore come apart from the ontology of intentionality in such a case, but it does not necessarily rule out such an ontology. Searle's metaphysics would not invalidate anything I've said in this dissertation, nor threaten my account of the value of intentional language to science. Ultimately what this

means is that my account of intentional language does not commit me to any particular ontological story of intentional states.

10.3 Final Thoughts

So where do we go from here? What does all of this mean for the future study of the mind (both scientifically and philosophically)? It means that we should not be afraid to embrace the different methodological tools we have at our disposal. Different scientific tools have different virtues, and we should not let our desire for a unified ontology undermine this important methodological point. We should not let our ontology dictate our methodology. The value of intentional descriptions to science is determined by the benefits they provide, and not by their reducibility to other sorts of descriptions.

Another important message to carry forward in our study of the mind is that we should not confuse mechanistic *models* with the physical mechanisms we use them to represent. The fact that we are attempting to understand physical mechanisms does not mean that a mechanistic model will be sufficient for all our scientific needs. There are important methodological tools we must employ in our study of mechanisms that do not necessarily characterize the parts, operations, or organization, of the mechanisms under investigation.

And finally, we should remember that intentional language is used in fruitful scientific practice, and this is not something we should ignore or trivialize. To forget this fact is to disregard genuine and productive scientific practices, and to invite confusion.

References

Andersen, R & Cui, H. (2009). Intention, Action Planning, and Decision Making in Parietal-Frontal Circuits. *Neuron* 63 (5): 568-583.

Ajzen, I. (1985). From intention to actions: A theory of planned behavior. In J. Kuhi and J. Beckmann (Eds.), *Action-control: From Cognition to Behaviour*. Heidelberg: Springer. 11-39.

Ajzen, I. (1988). *Attitudes, Personality and Behavior*. Milton Keynes: Open University Press.

Ajzen, I. (1991). The Theory of Planned Behavior. *Organizational Behavior and Human Decision Processes* 50: 179-211.

Ajzen, I. & Driver, B.L. (1992). Application of the theory of planned behavior to leisure choice. *Journal of Leisure Research* 24: 207-224.

Ajzen, I. & Madden, T. J. (1986). Prediction of goal-directed behavior: Attitudes, intentions, and perceived behavioral control. *Journal of Experimental Social Psychology* 22: 453-74.

Armitage, C. & Connor, M. (2001). Efficacy of the Theory of Planned Behaviour: A meta-analytic review. *British Journal of Social Psychology* 40: 471-499.

Batterman, R. (2000). Multiple Realizability and Universality. *British Journal for the Philosophy of Science* 51: 115-145.

Batterman, R. (2002a). Asymptotics and the Role of Minimal Models. *British Journal for the Philosophy of Science* 53: 21-38.

Batterman, R. (2002b). *The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence*. Oxford: Oxford University Press.

Batterman, W. (2011). Emergence, Singularities, and Symmetry Breaking. *Foundations of Physics* 41: 1031-1050.

Bechtel, W. (2005). The Challenge of Characterizing Operations in the Mechanisms Underlying Behavior. *Journal of the Experimental Analysis of Behavior* 84:313-325.

Bechtel, W. (2007). Reducing Psychology While Maintaining its Autonomy Via Mechanistic Explanations. In Schouten, M. and De Joong, H.L. (eds.), *The Matter of the Mind: Philosophical Essays on Psychology, Neuroscience and Reduction*. Blackwell Publishing.

Bechtel, W. (2008). *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. New York: Lawrence Erlbaum Associates.

Bechtel, W. & Abrahamsen, A. (2007). Mental Mechanisms, Autonomous Systems, and Moral Agency. *Proceedings of the Cognitive Science Society* 95-100.

Bechtel, W. & McCauley, R.N. (1999). Heuristic Identity Theory (or Back to the Future): The Mind-Body Problem Against the Background of Research Strategies in Cognitive Neuroscience. In M. Hahn and S.C. Stoness (Eds.), *Proceedings of the 21st Annual Meeting of the Cognitive Science Society*. Mahwah, N.J.: Lawrence Erlbaum Associates. 67-72.

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

Blue, C.L. (1995). The predictive capacity of the theory of reasoned action and the theory of planned behavior in exercise research: An integrated literature review. *Research in Nursing and Health* 18: 105-121.

Bogen, J. (2005). Regularities and Causality: Generalizations and Causal Explanations. *Studies in History and Philosophy of Science Part C: Studies in the History and Philosophy of Biological and Biomedical Science* 36: 397-420.

Bozionelos, G. & Bennett, P. (1999). The Theory of Planned Behaviour as Predictor of Exercise: The Moderating Influence of Beliefs and Personal Variables. *Journal of Health Psychology* 4: 517-529.

Braddon-Mitchell, D. & Jackson, F. (2007). *Philosophy of Mind and Cognition*, 3rd edition. Malden, MA: Blackwell Publishing.

Brentano, F. (1970/1874). Psychology from the Empirical Standpoint. In H. Modrick (Ed.), *Introduction to the Philosophy of Mind: Readings from Descartes to Strawson*. Glenview, Ill.: Scott, Foresman.

Burgess, J. (2004). Mathematics and Bleak House. *Philosophia Mathematica* 3 (12): 18-36.

Cajal, S. R. (1937) *Recollections of My Life*. Philadelphia: American Philosophical Society.

Callender, C. (2001). Taking thermodynamics too seriously. *Studies in History and Philosophy of Modern Physics* 32(4): 539–533.

Carnap, R. (1950). Empiricism, Semantics, and Ontology. *Revue Internationale de Philosophie* 4: 20-40.

Cartwright, N. (1999). *The Dappled World: A Study of the Boundaries of Science*. United Kingdom: Cambridge University Press.

Chemero, A. & Silberstein, M. (2008). After the Philosophy of Mind: Replacing Scholasticism with Science. *Philosophy of Science* 75 (1): 1-27.

Cherry, E. C. (1951). A history of the theory of information. *Proceedings of the Institute of Electrical Engineers* 98 (III), 389-393. Repr. with minor changes as "The Communication of Information" Scientific American 40 (1952), 640-664.

Churchland, P. S. (1980). Language, Thought, and Information Processing. *Noûs* 14: 147-70.

Churchland, P. S. (1989). *Neurophilosophy: Toward a Unified Science of the Mind-Brain*. Cambridge, Massachusetts: MIT Press

Churchland, P. M. (1981). Eliminative Materialism and the Propositional Attitudes. *The Journal of Philosophy* 78: 67-90.

Clark, A. & Chalmers, D. (1998). The Extended Mind. *Analysis* 58 (1): 7-19.

Connor, M. & Sparks, P. (1996). The theory of planned behaviour and health behaviours. In M. Conner and P. Norman (Eds.), *Predicting health behaviour*. Buckingham; Open University Press. 121-162.

Craver, C. (2006). When Mechanistic Models Explain. *Synthese* 153 (3): 355-376.

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

Craver, C., & Bechtel, W. (2006). Mechanism. In S. Sarkar and J. Pfeifer (Eds.), *Philosophy of Science: An Encyclopedia*. New York: Routledge.

Cummins, R. (2000). "How does it work?" vs. "What are the laws?" Two conceptions of psychological explanation. In F. Keil, and R. Wilson (Eds.), *Explanation and cognition*. Cambridge, MA: MIT Press. 117-145.

Davidson, D. (2002). Mental Events. In D. Chalmers (Ed.), *Philosophy of mind: Classical and contemporary readings*. New York: Oxford University Press. Originally published: 1970. In L. Foster and J. Swanson (Eds), *Experience and Theory*. Humanities Press. 79-101.

Dennett, D. (1971). Intentional Systems. *Journal of Philosophy* 68: 87-106.

Dennett, D. (1987). *The Intentional Stance*. Cambridge, Massachusetts: The MIT Press.

Dennett, D. (1991a). Real Patterns, *The Journal of Philosophy* 88: 27-51.

Dennett, D. (1991b). *Consciousness Explained*. Canada: Little, Brown and Company Limited.

Dennett, D. (1992). The Self as a Center of Narrative Gravity. *Self and Consciousness: Multiple Perspectives*. Hillsdale, NJ: Erlbaum.

Dennett, D. (1993). Back From the Drawing Board. In Dahlbom, B. (Ed.), *Dennett and His Critics*. Cambridge, Mass: Blackwell.

Dennett, D. (1994). Cognitive Science as reverse engineering: Several meanings of 'Top Down' and 'Bottom Up'. In D. Prawitz, B. Skyrms, and D. Westerstahl (Eds.), *Logic, methodology and philosophy of science IX*. Amsterdam, North-Holland: Elsevier Science. 679-689.

Dretske, F. (1981). *Knowledge and the Flow of Information*. Cambridge, Mass: MIT Press.

Dretske, F. (1985). Machines and the Mental. *Western Division APA Presidential Address*, April 26, 1985 (printed in *Proceedings and Addresses of the APA* 59: 23-33).

Dupré, J. (1993). *The Disorder of Things*. Cambridge, Mass: Harvard University Press.

Eliasmith, C. (1996). The Third Contender: A Critical Examination of the Dynamicist Theory of Cognition. *Philosophical Psychology* 9: 441-463.

Eliasmith, C. (2002). The Myth of the Turing Machine: The Failing of Functionalism and Related Theses. *Journal of Experimental & Theoretical Artificial Intelligence* 14 (1): 1-8.

Eliasmith, C. (2010). How we ought to describe computation in the brain. *Studies in History and Philosophy of Science Part A*, 41: 313-320.

Eliasmith, C. & Anderson, C. (2003). *Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems*. Cambridge, Massachusetts: MIT Press.

Fehr, C. (2006). Explanations of the Evolution of Sex: A Plurality of Local Mechanisms. In S. Kellert, H. Longino, and C.K. Waters (Eds.), *Scientific Pluralism*. Minneapolis: University of Minnesota Press: 26-41.

Fillion, N. (2008). The Role of Truth in Scientific Explanations: The Case of Phenomenological Theories in Continuum Mechanics. Accessed Online at: http://www.nfillion.com/docs/idealization.pdf

Fishbein, M. & Ajzen, I. (1975). *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research*. Reading, MA: Addison-Wesley.

Fishbein, M. & Ajzen, I. (1981). Attitudes and voting behavior: An application of the theory of reasoned action. In G.M. Stephenson and J. M. Davis (Eds.), *Progress in Applied Social Psychology*, Volume I. London: Wiley. 253-313.

Fodor, J. (1974). Special sciences and the disunity of science as a working hypothesis. *Synthese* 28: 77-115.

Fodor, J. (1975). *The Language of Thought*. New York: Crowell Press.

Fodor, J. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, Massachusetts: MIT Press.

Fodor, J. (1998). *Concepts: Where Cognitive Science Went Wrong*. Oxford: Clarendon Press.

Fodor, J. (2008). *LOT 2:The Language of Thought Revisited*. Oxford: Clarendon Press.

Fodor, J. & Lepore, E. (1992). *Holism: A Shopper's Guide*. Oxford: Blackwell.

Garcia-Bellido, A. (1984). Towards a Genetic Grammer. An English version of 'Hacia una Gramática Genética'. *Read Academia de Ciencias Exactas*, Fisicas u Naturales.

Gauker, C. (2009). Is Folk Psychology Predictive? A Challenge. *philpapers: online research in philosophy* Accessed Online at: http://philpapers.org/bbs/thread.pl?tId=329

Giere, R. (1996). *Science Without Laws*. Chicago: The University of Chicago Press.

Giere, R. (2006). Perspectival Pluralism. In S. Kellert, H. Longino, and C.K. Waters (Eds.), *Scientific Pluralism*. Minneapolis: University of Minnesota Press: 167-190.

Glennan, S. (2005). Modeling Mechanisms. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 36 (2): 375-388.

Godfrey-Smith, P. (2000). Information, Arbitrariness, and Selection: Comments on Maynard Smith. *Philosophy of Science* 67 (2): 202-207.

Godfrey-Smith, P. (2004). Genes do not Encode Information for Phenotypic Traits. In Christopher Hitchcock (Ed.), *Contemporary Debates in Philosophy of Science.* Malden, MA: Blackwell Publishing. 275-289.

Godin, G. & Kok, G. (1996). The theory of planned behavior: A review of its applications to health-related behaviors. *American Journal of Health Promotion* 11: 87-98.

Griffiths, P. (2001). Genetic Information: a metaphor in search of a theory. *Philosophy of Science* 67: 26-44.

Hacking, I. (1996). The Disunities of the Sciences. In Galison, P. and Stump, D. (eds), *The Disunity of Science: Boundaries, Contexts, and Power*. Stanford, California: Stanford University Press. 37-74.

Harris, S., Sheth, S., & Cohen, M. (2008). Functional Neuroimaging of Belief, Disbelief, and Uncertainty. *Annals of Neurology* 63 (2): 141-147.

Hausenblas, H. A., Carron, A. V., & Mack, D. E. (1997). Application of the theories of reasoned action and planned behavior to exercise behavior: a meta-analysis. *Journal of Sport and Exercise Psychology* 19: 36-51.

Hempel, C. & Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of Science* 15: 135-175. Reprinted in Hempel, 245-290, 1965a.

Hutto, D. (1999). *The presence of mind*. Philadelphia: J Benjamins Publishers.

Ji, S. (1997). Isomorphism between cell and human languages; Molecular biology, bioinformativ and linguistic implications. *BioSystems* 44: 17-39.

Ji, S. (1998). Cell Language (Cellese): Implications for Biology, Linguistics and Philosophy. *International Workshop on the Linguistics of Biology and the Biology of Language*, CIFN, Universidad Nacional Autónoma de Mexico, Cuernavaca México, March 23-27.

Ji, S. (1999). The linguistics of DNA: words, sentences, grammar, phonetics, and semantics. In *Molecular Strategies in Biological Evolution*, Volume 870 of the Annals of the New York Academy of Science: 411-417.

Kandel, E., Schwartz, J. & Jessel, T. (2000). *Principles of neuroscience* (4th ed.). New York: McGraw Hill.

Kellert, H., Longino, H. & Waters C. K. (2006). Introduction: The Pluralist Stance. In S. Kellert, H. Longino, and C.K. Waters (Eds.), *Scientific Pluralism*. Minneapolis: University of Minnesota Press. vii-xxix.

Kelso, J. A. S. (1995). *Dynamic Patterns: The Self-Organization of Brain and Behavior*. Cambridge, MA: MIT Press.

Kenyon, T. (2000). Indeterminacy and Realism. In D. Ross, A. Brook, and D. Thompson (Eds.), *Dennett's Philosophy*. MIT Press. 77-94.

Kirsh, D. (1991). Today the earwig, tomorrow man? *Artificial Intelligence*, 47: 161-184.

Larkin, J. & Simon, H. (1987). Why a Diagram is (Sometimes) Worth Ten Thousand Words. *Cognitive Science* 11: 65-99.

Lauden, L. (1981). A Confutation of Convergent Realism. *Philosophy of Science,* 48 (1): 19-49.

Longino, H. (2002). *The Fate of Knowledge*. Princeton: Princeton University Press.

Longino, H. (2006). Theoretical Pluralism and the Scientific Study of Behavior. In S. Kellert, H. Longino, and C.K. Waters (Eds.), *Scientific Pluralism*. Minneapolis: University of Minnesota Press: 102-132.

Ludvig, E. (2003). Why Pinker Needs Behaviorism: A Critique of The Blank Slate. *Behavior and Philosophy* 31: 139-143.

Machamer, P., Darden, L. & Craver, C. (2000). Thinking About Mechanisms. *Philosophy of Science* 67 (1): 1-25.

Maddy, P. (2011). *Defending the Axioms*. Oxford: Oxford University Press.

Marr, D. (1982). *Vision: A computation investigation into the human representational system and processing of visual Information.* San Francisco, CA: Freeman.

Meynell, L. (2008). Pictures, Pluralism, and Feminist Epistemology: Lessons from "Coming to Understand". *Hypatia* 23 (4): 1-29.

Morton, A. (1996). Folk Psychology is Not a Predictive Device. *Mind* 105 (417): 119-137.

Nersessian, N. (1988). Reasoning from Imagery and Analogy in Scientific Concept Formation. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, Volume One: Contribute Papers. 41-47.

Parisien, C. & Thagard, P. (2008). Robosemantics: How Stanley the Volkswagon Represents the World. *Minds & Machines* 18: 169-178.

Port, R. (2003). Meter and Speech. *Journal of Phonetics* 31: 599–611.

Putnam, H. (1967). Psychological Predicates. In W.H. Capitan and D. D. Merrill (Eds.), *Art, Mind and Religion*. Pittsburgh: University of Pittsburgh Press. 37-48.

Putnam, H. (1978). *Meaning and the Moral Sciences*. London: Routledge.

Quine, W.V.O. (1960). *Word and Object.* Cambridge, MA: The MIT Press.

Richardson, A. (2006). The Many Unities of Science: Politics, Semantics, and Ontology. In S. Kellert, H. Longino, and C.K. Waters (Eds.), *Scientific Pluralism*. Minneapolis: University of Minnesota Press. 1-26.

Ross, D. (2000). Rainforest Reaslism: A Dennettian Theory of Existence. In Ross, D., Brook, A., and Thompson, D. (Eds), *Dennett's Philosophy*. MIT Press. 147-168.

Ross, E. & Nisbett, R. (1991) *The Person and the Situation: Perspectives of Social Psychology*. Philadelphia: Temple University Press.

Sarkar, S. (1996). Biological Information: a skeptical look at some central dogmas of molecular biology. In S. Sarkar (Ed.), *The Philosophy and History of Molecular Biology: New Perspectives*. Dordrecht: Kluwer. 187-231.

Sarkar, S. (2000). Information in Genetics and Developmental Biology: Comments on Maynard Smith. *Philosophy of Science* 67 (2): 208-213.

Sarkar, S. (2004). Genes Encode Information for Phenotypic Traits. In Christopher Hitchcock (Ed.), *Contemporary Debates in Philosophy of Science*. Malden, MA: Blackwell Publishing. 259-274.

Salmon, W. (1989). *Four Decades of Scientific Explanation*. Minneapolis: University of Minnesota Press.

Scarr, S. (1995). Commentary. *Human Development* 38: 154-57.

Searle, J. (1980). Minds, Brains and Programs. *Behavioral and Brain Science*s 3: 417-457.

Searle, J. (1992). *The Rediscovery of the Mind.* Cambridge, Mass: The MIT Press.

Sellars, W. (1956) Empiricism and the Philosophy of Mind. In H. Feigl and M. Scriven (Eds.), *Minnesota Studies in the Philosophy of Science, Volume I: The Foundations of Science and the Concepts of Psychology and Psychoanalysis*. University of Minnesota Press. 253-329.

Sereno, M. I. (1991) Four Analogies Between Biological and Cultural/Linguistic Evolution. *Journal of Theoretical Biology* 151: 467-507.

Shannon, C. E. (1948). The Mathematical Theory of Communication. *Bell Systems Technical Journal* 27; 379-423, 623-656.

Smith, J. M. (2000). The Concept of Information in Biology. *Philosophy of Science* 67 (2): 177-94.

Sparks, P. (1994). Attitudes towards food: Applying, assessing and extending the 'theory of planned behaviour'. In D. R. Rutter and L. Quine (Eds.), *Social psychology and health: European perspectives*. Aldershot: Avebury Press. 25-46.

Sparks, P., Shepherd, R., Wieringa, N. & Zimmermans, N. (1995). Perceived behavioural control, unrealistic optimism and dietary change: An exploratory study. *Appetite* 24: 243-255.

Sterelny, K. (2000). The 'Genetic Program' Program: A Commentary on Maynard Smith on Information in Biology. *Philosophy of Science* 67 (2): 195-201.

Sterelny, K. & Griffiths, P. (1999). *Sex and Death: An Introduction to Philosophy of Biology*. Chicago: University of Chicago Press.

Sterelny, K., Smith, K. & Dickison, M. (1996). The Extended Replicator. *Biology and Philosophy* 11: 366-403.

Stich, S. (1983). *From Folk Psychology to Cognitive Science: The Case Against Belief*. Cambridge, Massachusetts: The MIT Press.

Stocker, A. & Simoncelli, E. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience* 9: 578-585.

Suppes, P. (1993). The Plurality of Science. In *Models and Methods in the Philosophy of Science: Selected Essays*. Dordrecht, The Netherlands: Kluwer Academic Publishing. 41-54

Thagard, P. (2009). Why Cognitive Science Needs Philosophy and Vice Versa. *Topics in Cognitive Science* 1 (2): 237-254.

Thrun, S. et al. (2006). Stanley: The Robot that won the DARPA Grand Challenge. *Journal of Field Robotics* 23: 661-692.

Timpson, C. (Forthcoming). *Quantum Information Theory and the Foundations of Quantum Mechanics*. Oxford: Clarendon Press.

Trafimow, D. & Finlay, K. A. (1996). The importance of subjective norms for a minority of people: Between subjects and within-subjects analyses. *Personality and Social Psychology Bulletin* 22: 820–828.

Truesdell, C. (1966). *The Mechanical Foundations of Elasticity and Fluid Dynamics*. New York: Gordon and Breach Science Publishers Inc.

Truesdell, C. (1980). Statistical Mechanics and Continuum Mechanics. In *An Idiot's Fugitive Essays on Science.* New York: Springer-Verlag. 72-79.

Van Fraassen, B. (1980). *The Scientific Image.* Oxford: Clarendon Press.

Watson, J.D., Tooze, J. & Kurtz D.T. (1983). *Recombinant DNA: A Short Course.* New York: W.H. Freeman.

Watters, A. E. (1989) Reasoned/intuitive action: An individual difference moderator of the attitude-behavior relationship in the 1988 U.S. presidential election. *Unpublished master's thesis*, Department of Psychology, University of Massachusetts at Amherst.

Wheeler, J.A. (1990). *Information, physics, quantum: The search for links*. In W. Zurek (Ed.), *Complexity, Entropy and the Physics of Information*. Redwood City, CA: Addison Wesley. 3-28.

Wilson, M. (2006). *Wandering Significance: An Essay on Conceptual Behaviour.* New York: Clarendon Press.

Woodward, J. (2000). Explanation and Invariance in the Special Sciences. *British Journal for the Philosophy of Science* 51: 197-254.

Zednik, C. (2011). The Nature of Dynamical Explanation. *Philosophy of Science* 78 (2): 238-263.

Zeilinger, A. (2005). The Message of the Quantum. *Nature* 438: 743.