"Expotition [sic] to the North Pole"
The 20th Century Search for Mind

by

Margo Harvie

A thesis submitted in conformity with the requirements
for the Degree of Master of Arts
Department of Theory and Policy Studies in Education
University of Toronto

"Expotition [sic] to the North Pole"[1]

As soon as he had finished his lunch Christopher Robin whispered to Rabbit, and Rabbit said, "Yes, yes, of course, and they walked a little way up the stream together.

"I didn't want the others to hear," said Christopher Robin.

"Quite so," said Rabbit, looking important.

"It's -- I wondered -- It's only -- Rabbit, I suppose *you* don't know, What does the North Pole *look* like."

"Well," said Rabbit, stroking his whiskers. "Now you're asking me."

"I did know once, only I've sort of forgotten," said Christopher Robin carelessly.

"It's a funny thing," said Rabbit, "but I've sort of forgotten too, although I did know *once*."

(A. A. Milne, 1926, p. 122)

[1]Taken from the title of Chapter VIII in Winnie the Pooh (1926) p. 111.

"Expotition [sic] to the North Pole":
The 20th Century Search for Mind

Master of Arts
Graduate Department of Theory and Policy Studies in Education
The Ontario Institute for Studies in Education of
The University of Toronto

2000

Margo Harvie

ABSTRACT

Functionalism, in one form or another, is widely accepted in philosophy and cognitive science as an account of the relationship between mental states and brain states. This thesis claims, however, that the functionalist model of mind is unstable and attempts to demonstrate how it consistently "falls back on" earlier theories (e.g., eliminativist behaviourism, identity theory) whose problems it was designed to overcome. The work of Robert Van Gulick, Daniel Dennett, and Fred Dretske is examined in order to show how each of these functionalist philosophers takes their representationalist explanation of intentional mental states (e.g., beliefs and desires) and uses it to develop an account of subjective consciousness. The thesis concludes that representational models which are based on the functionalist's tri-level account of

ACKNOWLEDGMENTS

I would like to express my sincere thanks and appreciation to my thesis supervisor Chris Olsen who (with some effort) kept me on track during the unmentionable number of months it took to complete this project and who, in the final weeks and days, allotted so much time and effort to its completion that his visits to the Fitness Institute dwindled from "occasional" to "never."

Thanks also to Bill Seager, who as the other member of my thesis committee, demonstrated a great deal of patience in waiting for the thesis to come together, and who provided several valuable suggestions for its improvement when it finally did.

To my family, nothing need be said except "it's done!"

CONTENTS

Chapter 1

The 20th Century Search for Mind

In a (not so recent) issue of Maclean's Magazine[2] it was reported that a psychologist at McMaster University was in possession of a substantial portion of Einstein's brain. This psychologist was entrusted with a segment of the brain in July 1996 by 84-year-old New Jersey pathologist Dr. Thomas Harvey. Dr. Harvey removed Einstein's brain (without permission) in 1955 and was so intent on keeping it, that he later lost his job at Princeton rather than turn over the specimen.

What information is this fragment of the physical universe likely to provide us with when it comes to understanding how mind and brain interact? Perhaps there will be some interesting neurophysiological findings that will come to light based on the relation between the particular brain matter in question and the indisputably impressive mental feats of the individual who once was responsible for getting it from one location to the next. But perhaps not. Exactly what can Einstein's brain (now entirely inert) tell us about the marvels of his mind?[3]

Thomas Nagel maintains that the history of philosophy of mind during the past fifty years can be characterized as an ongoing dispute between those who hold that mental phenomena and brain states are one and the same thing and those who maintain that no

---

[2] See the article entitled "Dissecting his genius," p. 12 in the October 20th 1997 issue of *Macleans Magazine*.

[3] An article entitled "Decoded in Canada: Einstein's brain," stated that there is "convincing evidence that the anatomy of Einstein's brain may have been as unusual as his intellect" (*The Globe and Mail*, June 18, 1999 p. A1).

convincing proof that they are (one and the same thing) has yet been given. Nagel writes: "If there really are mental phenomena, they must be either identical or nonidentical with physical phenomena. If dualism is to be avoided, and if behaviourism is not a viable way of avoiding it, one seems to be thrown into the arms of neural identity theory" (Nagel, 1995, p. 82). But this is going too fast. It is important to consider some of the "isms" mentioned above a little more closely in order to understand how each of various successive models of mind introduced during the last century was a reaction to the failings of the one which preceded it.

John Searle claims that what is striking about mainstream philosophy of mind during the past fifty years is how much of it seems "obviously false" (Searle, 1992, p. 3). Indeed, there doesn't *really* seem to be any conclusive explanation as to how our thoughts and feelings and conscious selves appear to issue effortlessly from the pinkish-grey mass of nervous tissue that rests comfortably inside each of our bonily protective skulls. Have we made any progress when it comes to understanding how human brains generate human thoughts or have we, as Hilary Putnam puts it, simply fallen in to the philosopher's typical "pattern of 'recoil'" (Putnam, 1994, pp. 445-6) in which we ignore the degree to which problems inherent in previous models of mind remain largely unsolved in the "solutions" that were designed to overcome them?

1.1 The Persistence of (and Problems Inherent in) the Cartesian Model of Mind

Although it seemed reasonable to Descartes that a person could best be understood as a physical body and an immaterial mind/soul, the need to reject any kind of dualist version of how mind and body interact has (with a few exceptions) acted as a central motivating

force behind the various materialist models of mind which have been dominant since the beginning of the 20th century. In fact, much of contemporary philosophy of mind appears to be motivated by what Searle refers to as a "terror of falling into Cartesian dualism" (Searle, 1992, p. 13). One of the most effective weapons which can be used against a given model of mind is the accusation that the theory in question is based on some sort of Cartesian, or dualist, division. Note also – and this is very interesting and relevant to the discussion of how various explanations of mind-brain evolved – that more than one of a group of quite diverse materialist accounts can be taken apart to reveal Cartesian tendencies. Indeed, that "spook stuff," as Daniel Dennett refers to it, seems to have a certain staying power.[4]

What is it about the dualist's description of a person as consisting of two substances – a physical body and a non-physical mind – that has so repelled the majority of philosophers of mind during this century? There are several interrelated problems inherent in the view. To begin with, it is difficult to understand what exactly an immaterial substance might be and how it might behave in an otherwise material universe. In addition, there is the question of causation. If the common intuition that our thoughts have causal efficacy in relation to our actions and behaviour is true, how exactly do these physical and non-physical substances interact? No one has yet come up with a clear and/or believable answer to this question.

---

[4] For example, Bechtel (1988) notes the similarity between substance dualism and eliminative materialism, p. 102; and equates property dualism and token identity theory, p. 110. Daniel Dennett refers to most, if not all, versions of materialism other than his own as Cartesian materialism. Likewise, the various versions of functionalism which have evolved from token identity theory often run up against the criticism that the very abstract descriptions of mind they promote take them in the direction of a dualistic model of mind-brain interaction.

To avoid the puzzling problem of substance interaction, it has sometimes been suggested that mind and body might run in parallel (i.e., with no causation between the two) or that they interact only partially (i.e., the material substance can cause an immaterial thought but not vice versa).[5] But neither form of parallelism is particularly effective in providing any kind of intuitively acceptable answer as to how (if we are comprised of two separate substances) mind and body work together so well as one. To claim that our beliefs and desires don't exist, or to imply that they are not causally connected to our actions, does "violence to our conviction that mental conditions are effective in human behaviour" (Campbell, 1984, p. 57).

Dualism, therefore, remains unacceptable to the vast majority of those who undertake to explore the connection between mind and brain. Defining persons in terms of two separate substances invariably leads to insoluble problems for the philosopher of mind. Whether it is objected to in terms of metaphysical extravagance[6] or as presenting the problem of overdetermination[7] in relation to causation of behaviour, dualism appears to postulate one too many items.

---

[5] This form of parallelism – in which brain events are seen to cause thoughts, but thoughts themselves are causally inefficacious – is referred to as epiphenomenalism. Epiphenomenalism, which provides a major challenge to nonreductive materialist accounts of mind, will be discussed in more detail in the chapters which follow.

[6] Bechtel (1988, p. 88) describes how dualism (whether the substance or property variety) violates the principle of *Occam's razor* – i.e., it postulates problematic and unnecessary entities.

[7] The term *overdetermination* refers to the fact that in an explanation of cause and effect, an effect cannot have two separate and independent causes (which, with dualism, seems to be the case). See Guttenplan, 1995, p. 85.

1.2 The Behaviourist Solution

Behaviourism deals with the problem of overdetermination quite easily. It simply eliminates any reference to one of the items – i.e., the mind. During the early decades of the twentieth century, psychological behaviourism emerged as a reaction to the approach of psychologists such as James and Titchner who proposed that *consciousness* was the rightful subject matter of psychology and who maintained that introspection was the correct methodology for its study. In contrast, J. B. Watson claimed that since psychology was a science, it should concern itself only with what was observable and objective (e.g., behaviour) (Kim, 1996, p. 25). To focus on consciousness, according to Watson, was to return "to the ancient days of superstition and magic" (Byrne, 1995, p. 134). The development of psychological behaviourism as an empirical research program was based, therefore, on the rejection of any reference to the mind and/or mental states.

Philosophical behaviourism, which developed alongside psychological behaviourism, is generally described as less eliminativist with respect to mental phenomena. For example, analytical, or logical, behaviourism was not as much concerned with denying the existence of mental states as it was with using the correct language to translate psychological concepts into physical concepts so that a given statement could be verified in the correct scientific fashion. The influence of logical positivism, from which analytical behaviourism evolved, can be seen in the behaviourist's insistence that statements that refer to mental phenomena (e.g., beliefs) must be translated into statements which contain terms that refer to observable physical states or occurrences. For example, a behaviourist would need to translate a given agent's belief *that it will rain* into the same agent's behaviour (or behavioural disposition) with respect to umbrellas or open

windows, and so on. Although most descriptions of the development of behaviourist thought make a point of distinguishing between the psychological and philosophical varieties, when it comes to denying that our thoughts have causal efficacy, their approach is one and the same. The behaviourist, whether a full-fledged eliminativist or not, must describe the causation of behaviour in purely physical terms, making no reference (at all) to mental states.

Behaviourism banished the mind in order to eliminate the need for an explanation of unscientific entities such as beliefs and desires which appeared to be inaccessible and unverifiable. According to the behaviourist, mental states are merely abstractions. They cannot be postulated as the causes of behaviour because, as explained above, this view leads in the direction of a potentially dualistic situation. Behaviourism, then, was intended to dispense with the insoluble problems that Cartesian models of mind invariably generate. The behaviourist account, however, is not immune to several tenacious problems of its own.

To begin with, the behaviourist's attempt to translate mental phenomena such as beliefs and desires using terms which refer (only) to physical behaviour(s) runs into the immediate problem that any explanation of behaviour is entirely open-ended. For example, my belief that it is 10:00 p.m. could result in any number of behaviours (or in no behaviour at all). Secondly, in describing mental states in terms of behaviour and/or dispositions to behave, it is extremely difficult to eliminate all reference to mental goings-on. For example, it is almost impossible to give an analysis of a belief without making reference to a desire, and vice versa (Searle, 1992, p. 34 and Kim, 1996, p. 34).

Another serious problem for the behaviourist model has to do with the issue of mental phenomena such as particular sensations (e.g. pain) which cannot be adequately described as behaviour or dispositions to behave. For example, to describe pain as the disposition to behave in a pain-behaving way (e.g., wincing, crying out, taking an aspirin, etc.) leaves out the essence of what, in fact, you are trying to get at – i.e., your first-person experience of discomfort.

The behaviourist approach, then, was successful in banishing Cartesian worries, but it came with its own set of problems. In particular, there is something about the behaviourist approach that is highly unintuitive. It just seems incomprehensible that our complex reactions to, and understandings of, the world might arise from nothing and no place at all.

1.3 The Physicalist's Reaction: Mind-Brain Identity Theory

During the 1950s, the problems inherent in logical behaviourism became more widely acknowledged, and a new theory – the identity theory – began to gain ground as a result of work published by U. T. Place and J. J. C. Smart. Proponents of the identity theory were committed to the proposal that mental states exist and that any given mental state is reducible to a specific brain state (e.g., pain is entirely reducible to the firing of C-fibres in the brain).[8]

---

[8] Place was happy enough with the philosophical behaviourist's explanation of most mental events as behaviour and/or dispositions to behave, but claimed that there remained an "intractable residue" of mental concepts having to do with sensation and conscious experience whose explanation would require reference to inner processes in the brain. (Place, 1956/62, p. 101).

Identity theory does seem to make much better intuitive sense than does behaviourism's denial of mental events. It is easier to understand my desire to answer the door when I hear the doorbell ring as being the result of some set of brain events than it is to understand the behaviourist's somewhat awkward version in which I, for no mentally-related reason at all, walk, or feel disposed to walk, towards the door. In addition, the reductionist approach to mind is compatible with the general direction that scientific discoveries are *supposed* take. Science, after all, is in the business of reducing the mysterious to the physical (e.g., heat to molecular activity, light to electromagnetic radiation, etc.). Like dualism and behaviourism, however, identity theory arrived with its own set of contentious issues.

Critics typically base their objections to strict reduction on Leibniz's Law which states that:

> if X is identical with Y, X and Y share all their properties in common so that for any property P, either both X and Y have P or both lack it.[9]

For example, if it is true that your belief has intentional content (i.e., it is *about* something), then it must be true that the brain event to which it reduces, also has the property of intentionality. The property of "aboutness," however, does not seem an appropriate one for brain states.

Smart defended the identity theory against the Leibniz's Law criticism by claiming that it is simply an issue of linguistic use that makes it seem that we attribute different

---

[9] For more details, see Kim's (1996) discussion of "indiscernibility of identicals," pp. 57-58.

properties to mental states and brain states.[10] Other defenders of the theory claimed that Leibniz's Law is not applicable to the type of identity (i.e., cross-category) in question. In general, however, most philosophers have interpreted the identity theory as requiring strict identity between mind and brain and have found this identity relation to be problematic.[11]

Another problem spawned by the requirement of strict identity between mental states and brain states is the fact that entities with neurophysiologies different than our own (of which there are plenty) must be denied mentality. They can't have pains and fears unless the brains states that accompany these mental states are identical to the brain states to which our own pains and fears reduce.[12]

1.4 Options for Those Who Reject the Identity Theory

Assuming that the human-chauvinistic stipulation described above is unacceptable, we are left with, at best, a *contingent* identity relation between our thoughts and behaviour and the neurophysiological events which take place in the brain. What's worse is that, in rejecting strict identity, we seem to be right back in the clutches of Cartesian worries. If mental states aren't exactly brain states, then what are they?

---

[10] According to Smart, "topic neutral" terminology can be used to solve the problem. For example, the experience of seeing a yellow-orange afterimage is to be described as "an event going on in me that is like the event that goes on in me when I see an orange" (Searle 1992, p. 37). Topic-neutral vocabulary, however, is just as awkward as the behaviourist's description of mental events as behaviour and/or dispositions to behave.

[11] Saul Kripke maintains that the terms to be identified must be rigid designators (i.e., all identity claims are necessary and not contingent) and that, therefore, mental states cannot be identical to physical brain states since it is not possible to determine the type of necessary conditions required in order to make the claim (Bechtel 1988, p. 99).

[12] Note that this same problem of human chauvinism shows up again in relation to some forms of functionalism, a model of mind designed to eliminate the errors of identity theory.

Generally speaking, there are two choices when it comes to responding to the problems of strict reduction. The first option is to eliminate half of the problem by classifying mental states as unreal, or illusory. The second option is to weaken the reduction requirement to terms which are less strict.

Philosophers such as Richard Rorty, and (more recently) Paul and Patricia Churchland, maintain that mental states are, in fact, radically different from brain states and that they, therefore, cannot be reduced to physical brain states. They maintain that, since the everyday (or folk-psychological) terminology used in the description of mind is irrelevant when it comes to a discussion of the workings of physical processes, reference to this terminology should be *eliminated* from any serious theory of mind. A similarity of approach between eliminative materialism and eliminative behaviourism can be seen here. In both cases, the folk-psychological terminology used in the discussion of beliefs and desires is said to have no place in materialist models which entail descriptions of physical behaviour and/or firings of neurons. As with behaviourism, however, the suggestion that our mental states are non-existent is highly nonintuitive.

There is an interesting point to make in reference to eliminative materialism. With the claim that mental events are radically different from brain events, the eliminativist appears to be, in some sense, acknowledging the Cartesian point of view in which the mind is said to be entirely distinct from brain. In other words, the insistence on the "elimination" of mind can be seen to imply that mental states – whatever they might, in fact, be – are entirely separate from the physical world.

For those who are skeptical that research will reveal a strict identity between mental and physical phenomena, but who refuse to accept the eliminativist stance with respect

to mind, there is a second way of responding to the problem. This is to promote a weaker version of identity theory which is referred to as token identity theory. Rather than claiming that there is one, and only one, type of brain state for each type of mental state, the token identity theorist claims that every instance, or token, of a mental state is identical with an instance, or token, of a bodily state, of some type or another (Rosenthal, 1995, p. 350). This acknowledgment of the multiple realizability of mental states allows organisms which are physiologically (and so on) dissimilar to ourselves to have the same (though numerically different) mental states. It also allows for the possibility that dissimilar (combinations of) brain states are responsible for similar mental states in different (or even the same) human individuals.

Although token identity seems somehow more reasonable than the strict reduction of mind to brain, it is faced with the problem of explaining exactly how it is that a given set of mental and physical tokens might relate. Token identity theories often seem to lead to a form of property dualism in which the same event is described in terms of both its mental and physical properties.[13] In order to solve this problem, token identity theory needs some kind of believable explanation of how mental events relate to the corresponding physical event(s) taking place in the brain. The most widely accepted explanation of this relationship is the notion of *supervenience* which refers to a form of nonreductive materialism in which mental states supervene on – are realized by – physical brain states while not being *reducible* to these states (Guttenplan, 1995, p. 94).

---

[13] Davidson's *anomalous monism* – which claims that the same event has both physical and mental properties – is sometimes given as an example of token identity theory which promotes a form of property dualism (see Bechtel 1988, p. 107 and Guttenplan, 1995, p. 92).

Token identity theory makes the claim, then, that there can be two different, but legitimate, accounts of the cognitive processes that take place in the brain – one which uses strictly neurophysiological vocabulary, and one which uses mentalistic terminology. But, as mentioned above, defining the relationship between mental events and brain events remains problematic. To avoid the unsettling proposition that mental states are epiphenomenal (i.e., that your beliefs, for example, play no causal role in the production of your behaviour), some way of understanding and/or categorizing mental phenomena and their relationship to physical brain states is needed. It is widely accepted by the majority of those involved in the philosophy of mind today that one or another form of *functionalism* best answers this need.

1.5 The Functionalist Description of Mind

Functionalism is often described as an intermediate position between philosophical behaviourism and identity theory in the sense that it provides a middle ground on which to resolve the problems spawned by behaviourism on one side and reductionism on the other. The distinguishing feature of functionalism is the claim that mental events are to be defined according to the function they perform within a given causal system. The behaviourist maintains that mental events are to be described in terms of behaviour (or dispositions to behave); the identity theorist claims that a mental event can be reduced to a neural event; the functionalist claims that a mental event is to be defined according to the causal/functional role it plays in the process of transforming input (i.e., sensory input and/or another mental state) to output (i.e., another mental state and/or behaviour).

While the behaviourist account allows for only input and output, the functionalist account provides for another (internal) step, or interface, between the two. It is, in fact, the task performed during this intermediate step that defines a particular mental state. For example, a mental event such as my belief that I hear the doorbell ringing is described in terms that encompass both the reception of sensory input (a neural event) and the movement of my body towards the door (a behavioural event). The belief itself, therefore, is described according to the role it plays in turning input into output.

Functionalism is based on the notion of multiple realizability – a notion, as was discussed above, which was introduced via the token identity theory. Rather than postulating a strict identity between a given type of mental event and a particular type of brain event, functionalism maintains that a mental event can be realized by any one of multiple events taking place in a brain or, for that matter, in some device with processing powers equivalent to a brain. The only restriction is that for each mental state token, there must be a corresponding brain state token of some type.

A functionalist model of one sort or another underlies the majority of present-day explanations of how our minds connect to the physical world. There are, however, several different versions and/or applications of functionalism which are sometimes confusingly conflated. Conceptual functionalism, or analytical functionalism as it is sometimes called, is modelled on the premise that mental states are to be defined according to the functional/causal role they take within the complex causal network which facilitates transactions between the mind and the outside world (Kim, 1996, p. 104). This form of functionalism maintains that common sense, or "folk," psychology can be used to understand, or define, the causal structure that underlies our mentality but does not

commit to any particular physical realization of the causal network it describes. Any system, no matter what its physical makeup, can be said to have mentality as long as the right causal relations exist between its input and output. This approach eliminates the problem of human chauvinism generated by the mind/brain identity theory but it provides no explanation of *how* different physical entities might be able to realize the same causal structure.

1.5.1 Putnam's Machine Functionalism

It was a theory referred to as machine functionalism which provided (according to some) a convincing answer as to how it might be possible for two dissimilar physical implementations to manifest the same mentality. In the early 1960s, Hilary Putnam used two standard functionalist concepts – the causal role of mental states and multiple realizability – to come up with a functionalist model of mind that claimed to be well grounded in the physical sciences.[14] Machine functionalism was based on the workings of a Turing computer for which a given operation is defined according to a set of explicit rules which stipulate what output is to occur based on the current input and the current state of the machine.

The machine functionalist compares the efficacy of a mental state to that of the processing unit which produces output as specified by the machine table of a Turing computer (i.e., both minds and machines are designed to turn input into output). What it

---

[14] Block describes the functionalism promoted by Putnam and Fodor as being based upon substantive scientific hypotheses as opposed to Lewis's conceptual functionalism which is based on a priori psychology (Block, 1978, p. 271).

means for something to have mentality is simply for it to physically realize a Turing machine of appropriate complexity (Kim, 1996, p. 88). A system has the capacity to feel pain, for example, if it has some mechanism that detects tissue (or some substance) damage and can produce an appropriate reaction (output). Of course, the mechanism whose function it is to detect pain can, and will, vary from one creature (implementation) to another.

In describing the mind as a probabilistic automaton, however, the problem arises that a given internal state of a Turing machine represents the total state of the machine at a given time, whereas a given mental state can never be described as comprising the total psychological state of the subject (Kim, 1996, p. 88). For this reason, machine functionalism evolved in such a way as to accommodate the view that mental states should be defined, not as a particular machine state, but rather as an operation that could be performed by variety of different machines (Bechtel, 1988, p. 117).

1.5.2 Computational or AI Functionalism

The proposal to view the mind as being comprised of numerous autonomous sub-systems, or components, was enthusiastically taken up by cognitive psychologists as well as by those involved in the area of artificial intelligence (AI). Computational functionalism facilitated research projects in which attempts were made to isolate particular cognitive processes found in humans and to simulate these in computers. Underlying the project to create "artificial intelligence" is the belief (carried over from machine functionalism) that our mental processes can be characterized according to the formal operations which are

carried out on symbols. According to this view, the processing in a given module of the mind consists of the right kind of symbol manipulation (Bechtel, 1988, p. 118).

The project of *strong AI* is committed to the belief that machines can be built which not only mimic human behaviour but which, in fact, use functionally the same operations as humans in order to produce equivalent behaviour. In other words, computational functionalism explains human cognition by describing it as particular "software" modules which run on the "hardware" of the brain.

In humans, unlike computers, the functional architecture is already set to a large degree by the biological constitution of the nervous system. The goal of strong AI, therefore, is to create modular virtual machines that simulate a particular procedure performed by the human brain and to allow this virtual machine to guide us in the development of the correct functional architecture (Bechtel, 1988, p. 121).

In fact, the main dispute in computational AI revolves around this issue of functional architecture. Originally, strong AI was committed to von Neuman architecture in which computation is defined strictly in terms of serial operations of symbol manipulation. Since the mid-1980s, however, cognitive theory has grown increasingly fond of the connectionist architecture of parallel distributed processing and the development of non-symbolic AI models. Smolensky's discussion of computationalism in Guttenplan's *A Companion to the Philosophy of Mind* maintains that the jury is still out on whether the computational approach taken will turn out to be a symbolic, connectionist, or some combination of the two (Smolensky, 1995, p. 176).

1.5.3 Homuncular Functionalism

Computational functionalism proposes three levels of description for each mental event including: 1) a neurophysiological description of a neural event in the physical brain, 2) a functional description in which the neural event is described in computational (but still physical) terms, and 3) a common-sense description of the functional event using everyday psychological terms. However, as William Lycan points out, neither living things nor computers are split between a purely physical and a purely abstract level (Lycan, 1995, p. 320). In order for functionalism to avoid accusations that it succumbs to some kind of Cartesian division, a more detailed description is required which would explain how these computational modules interact with other modules and within the brain itself.

According to homuncular functionalism, each modular component is controlled by a homunculus.[15] In order to avoid an infinite regress – in which the workings of the homuncular mind must be described as having its own internal homunculus, and so on – each functional component is further broken down into sub-components and then into sub-sub-components, with each descending level having a less complex/intelligent executive, or homunculus, function. If a flow chart were used to show the hierarchical structure of the functions involved in a particular task, it could be seen that the very lowest sub-sub-components would have task descriptions which are obviously mechanistic and therefore, require no homuncular "guidance" since they would be taking place at the neuroanatomical level (Lycan, 1995, p. 320). Or so the theory says.

---

[15] A homunculus is like a "little person" in the head who is postulated to have the limited intelligence required to take a specific role in our conscious cognitive functioning.

Note that the proposed structural hierarchy of the components and sub-components in homuncular functionalism is such that a given level can be described as functional (i.e., abstract) compared to the level below and structural (i.e., physical) compared to the level above it. According to Lycan, "homunctional [sic] characterizations and physiological characterizations of states of persons reflect merely different levels of abstraction within a surrounding functional hierarchy" (quoted in Bechtel, 1988, p. 123) so that, in a sense, it is unclear where functionalism ends off and the identity theory begins.

1.5.4 The Teleological Restraint

Although functionalism claims to have rescued mental states from the oblivion enforced by behaviourism, the fact that it describes such states in highly abstract terms can lead to the criticism that it allows for mentality to be assigned in far too a liberal manner. As Block points out in "Troubles with Functionalism," if all that counts when it comes to comparing functional processes is the causal interactions between components, entirely dissimilar entities will be found to be functionally equivalent. Block likes to use the example in which it would be possible for the population of China to be functionally equivalent to your current "state of mind" (Block, 1978, p. 276).

To avoid the problem of excessive liberalism in functionalist accounts, a teleological requirement is often imposed. This means that a system can be described in functional terms only if its operations are contributing to the overall needs of the organism (i.e., only if it is doing the job it was designed to do). In the case of humans, for example, it is proposed that a particular state of an organism can be described using functionalist terminology only if that state is performing the function that it was designed (i.e., through

evolution) to do.[16] But this proposal swings us back in the direction of human chauvinism[17] since it restricts functionalist description to systems which are very similar to ourselves.

1.6 Problems with the Functionalist Account

Based on this uneasy tradeoff between excessive liberalism and human chauvinism, Block makes the claim that functionalism – in attempting to define a middle ground between behaviourism and the identity theory – is, in fact, unstable. He maintains that functionalist descriptions of mind must either succumb to a form of eliminative behaviourism (in which just about anything could have mentality) or to identity theory (in which mentality must be denied to entities with neurological constitutions dissimilar to our own).

Computational functionalism runs into problems which lie outside of the liberalism versus chauvinism debate, however. For example, more than a few philosophers have rejected the claim that intentionality, or the *aboutness* of our thoughts, can be defined in terms of the formal processing of symbols. In *Representation and Reality*, Hilary Putnam rejects the computational functionalist's claim that propositional attitudes are computational states of the brain, stating that propositional attitudes are not definable as "parameters that would enter into a software description of the organism" (Putnam, 1988,

---

[16] According to both Lycan and Block, however, teleological claims are controversial. Block refers to the Swampman problem in which you discover that your grandparents were, in fact, formed from swamp particles and were not products of normal evolution. In this case, if you accept the teleological constraint, you have to admit that your mental states have no content since the normal evolutionary process which defines the functionality of mental states never took place in the case of your grandparents (and their descendants) (1995, p. 331).

[17] Chauvinism, here, refers to the limitations created by the assumption that only systems with neurological constitutions such as ourselves can be described in functionalist terms. Block's argument is that in avoiding liberalism, there is a tendency to fall back into chauvinism, and vice versa.

p. 73).[18] John Searle is well known for his claim that the formal syntax comprising a computer program could never produce true mentality which, according to Searle, requires semantic content. Likewise, Dreyfus has long argued that human cognition cannot be accounted for merely by means of formal representations and the rules for processing them (Bechtel, 1988, p. 126).

Block, however, does grant that computational simulation of our psychological processes is entirely possible and that these psychofunctional simulations would certainly be capable of the cognitive processing that lies behind our intentional states (e.g., beliefs and desires). If the problem of human chauvinism isn't an issue, says Block, psychofunctionalism can do the job since it is capable of reproducing our psychological states of mind using, for example, a particular hardware/software configuration. But there is, according to Block (and others), one problem. The psychofunctional simulation is without any qualitative experience – i.e., it is lacking conscious and continuous first-person experience of itself and the world around it (Block, 1978, p. 287).

1.7 The Particular Problem of Consciousness

For quite some time, those working in the field of cognitive science paid little attention to the issue of consciousness. The assumption was, perhaps, that functionalism had enough on its plate just trying to explain how something entirely abstract (such as our beliefs and desires) could possibly take a role in the causation of physical behaviour. In the past decade, however, the issue of whether consciousness can be explained in terms that will

---

[18] Note that it was Putnam himself who, in the early 1960s, introduced the theory of computational functionalism – a theory which he later rejected in *Representation and Reality (1988)*.

prove acceptable to our materialist view of reality has kept philosophers of mind very busy. One after another, books attempting to deal with the "consciousness problem" have hit the book stands and much of the discussion has been quite heated.[19]

There is a reason for the recent interest in the topic of consciousness. Functionalism is seen by the majority of those involved in cognitive science as the last best hope when it comes to providing a method of understanding how mental states connect to physical brain processes.[20] Evolving from the rejection of dualism, behaviourism, and type-type identity theory, functionalism has come to be seen as showing the greatest potential for providing a viable way in which to understand and describe the connection between our thoughts and the physical world. In addition, the fact that computers have been able to simulate certain of our psychological processes seems to provide convincing proof that (at least some) of our mental states can be described in functional and/or computational terms.

There is a widespread concern, however, about whether functionalism can handle *all* of the phenomena that comprise our mentality. Assuming that it might eventually be possible to create an accurate simulation of a complete human psychology, there is no evidence that the entity in which the simulation occurs would be conscious in the sense that it could feel the pains it "complained" about, or have first-person "understanding/ experience" of the content of its mental states.

---

[19] For example, see the debate between Searle and Dennett in Searle's book, *The Mystery of Consciousness* (1997).

[20] See Jerry Fodor's *The Language of Thought* in which he quotes Lyndon B. Johnson – "I'm the only President you've got" – in order to point to one of the reasons behind the continuing popularity of the functionalist model of mind. (Fodor, 1975, p. 27).

Davies and Humphreys maintain that philosophers of mind can be classified according to which of two stances they take with respect to the consciousness problem: elusiveness or demystification (Davies and Humphreys, 1993, p. 13). For example, philosophers such as Thomas Nagel, Colin McGinn, and Frank Jackson argue that the subjective first-person experience of being a conscious self is characterized by an elusiveness that no materialist account of mind (e.g., functionalism) can overcome. They attack functionalism using arguments[21] which claim to show that this materialist model of mind is unable to capture the most essential components of our mentality.

Demystifiers such as David Rosenthal, Daniel Dennett, William Lycan, Robert Van Gulick, and Fred Dretske, on the other hand, claim that there is nothing to mentality and/or consciousness that can't be explained using the right kind of functionalist account. What is interesting is that those committed to the demystification of consciousness have frequently used the same method of "solving" the problem of how to explain what lies behind our first-person experiences of the world.

1.8 The Representational Solution to the Problem of Consciousness

In 1978, Block claimed that he knew of only one serious attempt to fit 'consciousness' into information-flow psychology and this was the program described by Daniel Dennett in

---

[21] The *inverted spectrum* problem arises in the case of two systems which are functionally identical except that one has (for example) the qualitative experience of seeing red when the other has the qualitative experience of seeing green. Since their behaviour with respect to colour is identical, there is no way of determining which colour qualia they are experiencing. Likewise, with respect to the *absent qualia* argument, a functional simulation might demonstrate appropriate behaviour with respect to colour and yet have no qualitative experience whatsoever. Frank Jackson's *knowledge argument* claims that Mary the neuroscientist (who is entirely cognizant of the physics and neurophysiology that lies behind colour vision) still does not *know* what it is to have the qualitative experience of colour since she has lived her entire life in black and white.

his 1969 book entitled *Content and Consciousness* (Block, 1978, p. 290). Since then, however, several philosophers have taken on the challenge of explaining our first-person conscious experiences of the world in terms of a particular representationalist account of mind – an account which they had previously used to explain the intentional aspect of our thoughts. In the chapters which follow, I will take a look at three of these representationalist accounts of mind, and follow the development of each – from its early stages right up until the philosopher in question presents his "solution" to the consciousness problem. The question which I will attempt to answer is this: does the representationalist model of mind provide an adequate explanation of consciousness (or, for that matter, of intentionality) or must functionalism – like the approaches of behaviourism and identity theory which proceeded it – admit defeat when it comes to explaining just how our conscious and contentful states of mind relate to the physical world?

Chapter 2

Gaps and Transparencies:

Van Gulick's Representational Model of Consciousness

Some philosophers of mind – Colin McGinn, for example – have suggested that it's time to consider the possibility that we are never going to understand the relationship between our conscious experiences and the brain processes which are said to underlie them. Those committed to materialism, however, don't take kindly to this suggestion. After all, if the brain is a physical object (and it certainly seems to be) nothing but time should stand between us and a scientific explanation of how it works.

Robert Van Gulick, for example, rejects McGinn's claim that an understanding of the link between mind and brain must forever remain cognitively closed to us (McGinn, 1991, p. 3). Monkeys (it seems) may be unable to understand the concept of the electron, and armadillos will forever be befuddled by the most elementary mathematics, but surely, we, with our great capacity for learning and our comprehensive understanding of the physical sciences cannot be compared to creatures such as these! Van Gulick (for one) claims we cannot. We are not (just) armadillos, he says, and therefore we can (or at least it's more than likely that we can) come up with a good model of the way in which our conscious mental states supervene on our physical brain states.

Van Gulick is a functionalist and a realist about phenomenal consciousness. In other words, he remains optimistic that the functionalist model of mind will be able to provide

a convincing explanation of our first-person experiences – that sense of "something it is like to be you" experiencing the sights and sounds, etc., of the external world.

In order to understand Van Gulick's explanation of consciousness, it is important to have an understanding of the way his representationalist model of mind has developed over time. In this chapter, I will look at three of Van Gulick's articles,[1] and argue that his claims with respect to what his representationalist model of mind can explain have grown increasingly bold with each publication. His approach, in fact, reflects a progression which is typical of many functionalist philosophers during the past decade or so (i.e., having originally steered clear of the consciousness debate, he becomes increasingly committed to providing us with a functionalist account of our first-person experiences). The point to keep in mind, however, is that Van Gulick's explanation of consciousness can only be as viable as the representationalist base from which it was built.

2.1 Representation: the Basis of a Functional Mind

Van Gulick begins his 1982 article "Mental Representation – a Functionalist View," by making the distinction between representational "states of mind" and "mental representations." The former refer to mental states (e.g., beliefs) which are said to represent the world "as being in some particular way" (Van Gulick, 1982, p. 3). Mental representations, on the other hand, "are to be understood as formal or syntactic structures which function as internal symbols" (Van Gulick, 1982, p. 3). A representational system, then, is one which has the ability to recognize, and respond appropriately to, the formal

[1] The 1982 article "Mental Representation – a Functionalist View," the 1988 "Consciousness, intrinsic intentionality, and self-understanding machines," and the 1993 "Understanding the Phenomenal Mind: Are We All just Armadillos?"

syntactic structure of internal representations that are said to underlie the production of representational states of mind such as beliefs and desires. As a functionalist, Van Gulick holds the view that mental states are to be defined by the functional role they play within the mental economy. He maintains that a given mental state has the "content" or "meaning" that it does in virtue of the causal role it plays in regulating an organism's (or system's) interaction with its environment and that this causal role is always defined by the formal syntax of the corresponding internal representations.

The connection between these representational states of mind and the internal representations that are said to underlie them is, however, somewhat more tenuous than Van Gulick lets on. His claim is that a given mental state (e.g., the belief that it is raining) "causes" the system that possesses it to react to its environment in a particular way (e.g., the system goes looking for a raincoat). The actual physical cause of this action, however, is entirely reliant on the system's ability to interpret and process the formal syntax of internal representations. What Van Gulick's account fails to provide, however, are the details of the way in which mental states actually hook up to these internal representations.

2.1.1 Who and/or what has intentional states (of mind)?

In order to avoid assigning intentional states of mind too liberally (i.e., to organisms and/or systems that lack them), certain restraints must be applied to the functionalist account. Van Gulick applies a teleological constraint on the output/behaviour produced by a given system. A system can be said to have intentional mental states if it is able to "modify its behaviour in ways which are adaptive given the situation which it takes to

obtain" (Van Gulick, 1982, p. 5). In other words, the system's behaviour must indicate that it (in some sense) "understands" the information it is processing. It must indicate goal-directedness in relation to the specific environment it is operating in. "The basic idea," writes Van Gulick "is that of a system which tends toward certain end states and which exhibits plasticity and persistence in doing so in the face of disturbing influences" (Van Gulick, 1982, p. 6).

There is an additional constraint, however, which must be satisfied before a system can be described as operating on the basis of having an actual "understanding" of its mental states. This constraint has to do with the determination of rationality which, according to Van Gulick, is defined according to the number and complexity of inferential relations between mental states. For example, in order to ascribe sophisticated mental states such as beliefs and desires to a system, the assumption must be made that the system has the appropriate conceptual and inferential structure consisting of a logically-connected set of interdependent contentful states. As Van Gulick puts it, "considerations of holism apply pervasively in relating content to functional role" (Van Gulick, 1982, p. 6). In other words, with highly sophisticated intentional systems, the content associated with one mental state is inextricably tied to the content associated with one or more other mental states.[2]

Van Gulick rightly acknowledges the issue of holism here but he fails to note that it is exactly this issue that creates problems for his representationalist account of intentional content. I noted above that Van Gulick fails to provide any clear explanation of the

[2] For example, it is not possible to believe there is a coffee pot on the stove while having no (other) beliefs about stoves, kitchens, or cooking (Van Gulick, 1982, p. 6).

connection between mental states and the internal representations. With the acknowledgment that the content of a given mental state is inextricably tied to the content associated with an indeterminate number of other mental states, the possibility of a clear explanation of the connection between these states of mind and the formal syntax of internal representations appears even more remote. And, without an explanation, the claim that mental states are causally connected to behaviour is empty. Van Gulick maintains that it is the processing of formal syntax that causes behaviour in a representational system. If no clear connection between this syntax and what Van Gulick refers to as our representational states of mind can be established, our beliefs and desires run the risk of being classified as entirely superfluous.

Van Gulick, then, maintains that a true representational system must have a certain teleological design and meet certain standards of rationality, but he provides no clear explanation of how it might be possible to ascertain that either of these requirements is met. What is even more confusing in his account of representational systems is that, having discussed these two qualifications (i.e., teleology and rationality), he appears to suddenly "loosen up" on these requirements and argue that intentionality can exist, after all, outside of the context of representational states of mind such as beliefs and desires.

Not all informational states in a representational system, says Van Gulick, will be inferentially related to other states to the same degree as specified above. For example, he points to the impoverished behavioural consequences of certain perceptual states of a frog in which the frog's tongue lashes out indiscriminately at, for example, all black specs. The mental state which is said to cause the frog's behaviour cannot be described as any sort of belief or desire, since in the case of the frog's tongue response, the information being

processed is "opaquely embedded in fixed adaptive behaviour regulating mechanisms" (Van Gulick, 1982, p. 8). According to Van Gulick, even representational systems which can only respond to their environment in a very restricted manner (e.g., frogs) can still be said to have some degree of intentionality since their behaviour can be described in teleological terms – i.e., it meets the goal of adapting to the environment in which they must operate.

2.1.2 How a representational system works

As discussed above, Van Gulick's explanation of representation distinguishes between representational states of mind and the formal syntax of the internal representations which he claims underlie these mental states. A representational system S possesses two different sorts of information about a given mental representation f – 1) its intrinsic, or formal, character and, 2) the (semantic) information which f represents. The idea of a representation, he writes, is the idea of "one item going proxy for another" (Van Gulick, 1982 p. 11). By possessing syntactic information about one item (the representation), an intentional system "gains access to" semantic information about some other item (e.g., an object in the environment). S's "understanding" of the content associated with f is explained in terms of the set of operations it performs as a result of identifying f's particular formal structure. S knows what to do – it knows which set of operations is valid – given the syntactic structure of f.

Van Gulick's account of mental states is not always clear, however, when it comes to exactly which mental states are *fully* representational and which are not. He claims that the more some item f moves towards being used as a representation, "the more its content

is detached from any direct behaviour-regulating role and instead depends on the form sensitive processes which interpret it" (Van Gulick, 1982, p. 17). In other words, the wider the range of possible responses that S can have to f, the more likely it is that S is operating as a representational system with contentful psychological states such as beliefs and desires. Van Gulick repeatedly makes the point, however, that the distinction between the two categories of mental states – those which are fully representational and those which are not – is simply a question of degree. From this we can only conclude that all mental states are to be considered representational – some are just *more so* than others.

Why, then, does Van Gulick go out of his way to distinguish between those mental states whose output is based on a complex structure of inferential relations and those which operate as fixed behaviour-regulating mechanisms? The answer becomes clear during his discussion of the advantages of representational systems.

2.1.3 Hierarchies of intelligence

According to Van Gulick, the only way a complex intentional system can operate effectively is through the use of a comprehensive network of logically interrelated representations. He argues that the issue is one of design. Sophisticated information-processing systems have a need for economy when it comes to design, both in relation to storage and to the integration and processing of information. Representations provide for these design features (Van Gulick, 1982, p. 17).

The major advantage of using a representationalist model to understand how the mind works, however, is that it provides a *decompositional* strategy – a method of explaining how the (what seems to be) intrinsic intentionality of a highly sophisticated

representational system, can, in fact, be broken down into very simple component states through the use of hierarchically-nested systems of interactive homunculi (Van Gulick, 1982, p. 18).

Van Gulick's description of homuncular functionalism is as follows. The homunculi at the top of the hierarchy are able to interact with internal representations in a way that indicates that they have some sort of "understanding" of how these representations fit into the overall representational structure. At this level in the hierarchy there are numerous valid responses to the formal syntax of a particular representation and these options are based on the "context" of the processing which is currently taking place. It is possible to explain the apparent "intelligence" of the higher-level homunculi by claiming that each homunculus is supported by a set of less intelligent homunculi on the level below it. The further down in the hierarchy a given homunculus is located, the less "understanding" it has of the fact that a given formal structure represents something and the more it simply "reacts" to the structure in question with a simple and automatic operation. Low level homunculi, therefore, might be seen to correspond to what Van Gulick has referred to previously as fixed behaviour-regulating mechanisms (Van Gulick, 1982, p. 8).

The homuncular model of functionalism is popular with certain functionalists[3] because it appears to provide a way of explaining how the highly abstract information processing operations of S are traceable all the way down to its very lowest-level operations – operations which can be ultimately explained in physical and/or hardware terms. The apparent "intelligence" of the representational system is dealt with by positing a series of hierarchically-nested (and progressively less intelligent) homunculi whose

[3] In particular, see Dennett's account in his 1978 *Brainstorms*.

operations and interaction enable S to function intelligently in its environment. The lowest level homunculi perform duties which are so simple they can be explained in "nothing other than causal or physical hardware terms" (Van Gulick, 1982, p. 18).

There is, however, something unsettling about homuncular accounts. The decompositional strategy which Van Gulick promotes implies that the highly abstract intentional states he is describing can, in some sense, be traced back to one or more neural events. Van Gulick is not, however, proposing a reductive account of mind. There is no doubt about his commitment to the claim that content does not reduce to a purely physical level. For example, he quotes from Donald Davidson's "Mental Events" as follows: "It is not supposed that we will arrive at a complete and exact theory which generates true lawlike biconditionals (or conditionals) tying a physical description of a system and its environment to a description of the system's contentful states" (Van Gulick, 1982, p. 8). Just the same, the rationale behind hierarchical accounts of functional systems often seems to lead back in the direction of a reductionist-flavoured explanation since the intent is to blend the distinction between function and structure in such a way as to avoid any kind of strict division between the abstract and physical.

Lycan, one of the main proponents of the homuncular approach, writes: "homunctional characterizations and physiological characterizations of states of persons reflect merely different levels of abstraction within the surrounding functional hierarchy or continuum . . . [so that] we can no longer distinguish the functionalist from the identity theorist in any absolute way" (Bechtel, 1988, p. 123). Lycan's statement is interesting considering that the tri-level functionalist account of mind was developed to avoid the problems inherent in the reductive claims of identity theory. For the moment, however,

I want to ignore these problems and go on to look into how Van Gulick makes use of his theory of representation to provide a functionalist model of phenomenal consciousness.

2.2 Self-understanding Machines

In his 1988 article, "Consciousness, intrinsic intentionality, and self-understanding machines," Van Gulick makes the bold[4] move of attempting to use his functionalist model of intentional mental states to provide an explanation of our subjective conscious experiences. His discussion is based on a set of progressively refined questions but the eventual (and essential) question he asks is this: Can the representationalist model of mind he has developed to explain the intentional nature of mental states be extended/used to explain the first-person, subjective aspect of these states?

Van Gulick acknowledges that some philosophers of mind (e.g., Searle and Nagel) deny that functionalism *can* be used to explain intentionality.[5] Originally, says Van Gulick, he concluded that Searle's and Nagel's view (that a capacity for subjective experience is a *prerequisite* for intrinsic intentionality) and his own (that an intentional state is simply one that plays an appropriate causal role in mediating the system's interactions with its environment) were simply incompatible – and that, of course, his view was right. He claims, however, that eventually he began to entertain the notion that a close consideration of the subjective, first-person aspect of some intentional states might be helpful when it comes to developing a better understanding of how intentionality actually works.

---

[4] In "A Functionalist Plea for Self-Consciousness," (1988b) Van Gulick points out that the prevailing view of the time was that functionalism could handle an explanation of intentional states but not the subjective aspect of these states.

[5] These philosophers argue that 1) functionalism cannot explain consciousness 2) intrinsic intentionality cannot be separated out from consciousness.

2.2.1 The semantic transparency of experience

In order to explain what might lie behind the subjective "feel" of some mental states, Van Gulick builds on his original explanation of how different representational systems have different degrees of "understanding" of the symbols they process. To reiterate, his claim is that the more sophisticated a particular representational system is, the more it can be said to have an "understanding" of the content it processes. In other words, the further a system moves away from a design in which there is a single fixed response to a given representation and towards a design in which the organism responds to representations based on a complex conceptual/inferential structure, the more it can be said to be a fully-intentional system. Likewise, he says, the more a system can be said to "understand" the content it processes, the more likely it is to be capable of having conscious, subjective experiences. The way Van Gulick likes to put it is that first-person phenomenal experiences involve representations (i.e., symbolic structures) which have a very high degree of what he calls "semantic transparency" (1988a, p. 94).

What exactly does Van Gulick mean by this term? Take, for example, the experience of viewing a colourful perennial garden in full bloom. According to Van Gulick, the experience would seem to involve "a complex representation which has the form of a three-dimensional manifold, which is locally differentiated in a variety of ways" (Van Gulick, 1988a p. 94). When you see the garden, you understand "how that representation represents the world as being" (Van Gulick, 1988a, p. 94). In other words, you simply "see the garden." The complex network of representations which underlies phenomenal experience is so transparent that you "normally 'look right through them'"(Van Gulick,

1988a, p. 94). There is no awareness of any representational process at all. The experience is of the external world (as represented).

What is responsible for this transparency? Van Gulick claims that it results not from any difference in the representations themselves, but rather from the fact that our phenomenal experiences require the processing of a vast number of complex and interrelated representations – a process that takes place within a logically structured domain of interdependent concepts. These highly sophisticated processing capabilities – which require almost instantaneous movement from one representation to another semantically-related representation – result in the experience that you (as a subject) are perceiving multiple, interrelated, and complex objects in the external world. The wording that Van Gulick uses here, however, points to another potential problem. His nonreductive materialist account of mind appears to raise the question of whether you experience the external world or, rather, a representation of it.[6]

Nevertheless, Van Gulick's claim is that the most promising approach when it comes to accounting for the subjective experience of understanding, is one which focuses on the way in which a system's internal processing relates one representation to another, rather than one which focuses on a single (e.g., qualitative) aspect of a given experience. In order to understand phenomenal consciousness, we should concentrate not on specific qualia (e.g., the taste of Chase and Sanborn) but rather on the dynamic process that underlies and facilitates the qualitative experience we are having. Central to Van Gulick's explanation of the semantic transparency of phenomenal experience is his description of how the

[6] Hilary Putnam's claim is that this confusion is the fault of the artificial (and unnecessary) interface that functionalism creates between mind and world (Putnam, 1994, p. 454).

concept of self is spawned by the complexity of the system's representational processing. The sense that you are having experiences, and understanding these experiences, is simply the result of the system's ability to make the appropriate connections between representations. You, as a self, do not control the process; rather the self (that you think you are) results from "the organized system of subpersonal components which produce [your] orderly flow of thoughts" and experience (Van Gulick, 1988a p. 96).[7]

Van Gulick anticipates that certain questions might arise in relation to his model of semantic transparency. For example, the extremely sophisticated cognitive structure that is said to lie behind phenomenal experience would seem to be lacking in certain beings (e.g., infants, non-human animals) who certainly seem to be phenomenally conscious. In addition, Van Gulick acknowledges that it remains an open question as to whether any system (e.g., future AI creations) capable of representational processing at a level of complexity sufficient to ensure for "transparency" can be said to have conscious subjective experiences of the world it represents since – to the best of our knowledge – no amount of highly complex computational processing has ever resulted in a conscious machine.

As before, however, I want to pass over these potential problems in order to follow Van Gulick as he continues to build on his particular theory of representation in order to come up with an even more detailed description of what lies behind subjective consciousness.

---

[7] Note that this treatment of the "self" is very similar to the one Daniel Dennett provides in his 1991 account of consciousness – an account which is frequently referred to as having strong eliminativist tendencies.

2.3 Understanding the Phenomenal Mind

In "Understanding the Phenomenal Mind: Are We All just Armadillos?" (1993), Van Gulick begins by arguing that phenomenal consciousness poses no serious threat to functionalism – making rather short shrift of the standard arguments which are typically used against functionalist models of consciousness. The second half of the chapter is devoted to the further development of the connection between phenomenal mental states and Van Gulick's notion of semantic transparency. I want to proceed in the reverse order, however. First, I want to look at how Van Gulick's functionalist explanation of the role of phenomenal mental states has developed since 1988. Secondly, I'm going to argue that, in spite of the appeal of the theory of semantic transparency, Van Gulick has been too hasty in claiming that his model is robust enough to defeat the standard arguments used against functionalist accounts of consciousness.

2.3.1 Kantian support for transparent processes

In "Consciousness, intrinsic intentionality, and self-understanding machines," Van Gulick claims that the fact that we experience the world transparently is due to our brain's ability to instantly and effortlessly connect a multitude of conceptually interrelated representations. In his 1993 article, Van Gulick changes the wording used to describe semantic transparency somewhat. He makes the claim that it is the *density* of the interdependent relations between the many associated representations that gives phenomenal objects their 'thickness' and objectivity (Van Gulick, 1993, p. 151). The speed and complexity of the processing is responsible for the sense we have of being a self, or

subject, who is experiencing a world of objects. In support of his explanation, Van Gulick refers to Kant's notion of the experience of a world:

> Conscious experience involves more than just being in states that represent or refer to objects and their properties. In some sense, which is hard to articulate, it involves there being a world of objects inherent in the representation. Or perhaps one should say it inherently involves an interdependent structure of conscious subject and world of objects set over against one another since, as Kant taught us, the notions of subject and object are interdependent correlatives within the structure of experience. One might say that conscious phenomenal experience involves the construction of a model of the world that in some sense itself is a world, but is so only from the subjective perspective of the self, which in turn exists only as a feature of the organization of experience (Van Gulick, 1993, p. 150).

Van Gulick goes on to incorporate his take on Kant's notion of a continuous sensuous manifold in his explanation of why objects are present to us as particular things. The 'thing-liness' of phenomenal experience, he writes, requires an intuition in the Kantian sense of the word. The world must appear to us as a "continuous sensuous manifold ... in which objects can be present as particular things" (Van Gulick, 1993, p. 151). Although he doesn't explicitly say so, I understand Van Gulick to be claiming here that the continuous sensuous manifold is the product of the sophisticated processing carried out by a full-fledged representational system. The concreteness of the object we perceive derives from the complex processing which takes place within the vast network of interrelated representations which are associated with it.

Van Gulick suggests that the much-discussed problem of qualia can be explained in terms of the function qualia perform within the continuous sensuous manifold of experience. His claim is that the continuous sensuous manifold derives from the density of the relations among the (representations of the) objects it specifies. A semantically transparent representation is one which, by definition, carries an extensive amount of

information about how any particular object is spatiotemporally related to other objects within the continuous sensuous manifold. Qualia, says Van Gulick, can be thought of as the properties by which regions within a manifold, and objects within a region, are differentiated and delimited (Van Gulick, 1993, p. 151).

In order to understand what Van Gulick is saying here, let's return to the example of a large and colourful perennial garden. Within this spatiotemporal portion of the continuous sensuous manifold, a myriad of sensory experiences – such as the contrasting colours and shapes of the flowers and background objects such as house and sky, the sound of a buzzing insects, the feel of the hot sun or a cool breeze or both, etc. and so on – are processed in such a way as to provide concreteness to the objects present in your experience. It is the redness (for example) of one clump of flowers in contrast with the yellow of the wall of the house that helps to delimit flowers and house as separate but related concrete objects within the given sensuous manifold.

According to Van Gulick, this representational account of phenomenal experience manages to avoid the awkwardness of explanations which treat qualia as "basic simples" – i.e., in which the visual experience of (for example) a certain colour of red is separated out from the overall experience of which it is only a part and treated as some sort of free-standing mental entity. When it comes to defending his theory of semantic transparency, however, Van Gulick is a little too quick to discount the standard arguments which are typically used against the functionalist account of consciousness. In the section which follows, I will take a look at whether Van Gulick's explanation of phenomenal experience is, in fact, successful in overcoming what Joseph Levine refers to as the explanatory gap problem (Levine, 1983, p. 354).

2.4 Looking for gaps in the transparency

In his 1983 article "Materialism and Qualia: The Explanatory Gap," Levine describes his stance as weaker than Kripke's claim[8] that materialism is certainly false. Levine maintains, however, that materialism is threatened by a serious epistemological problem when it comes to explaining how our first-person experiences of the world relate to physical brain processes. Levine's claim is that psychophysical identity statements which attempt to reduce subjective experiences to either neurological brain processes *or* functional states are weakened by a certain explanatory gap – a gap which results in our being unable to determine whether any given psychophysical statement is, in fact, necessarily true.

In Levine's discussion of this problem, he asks the reader to consider the functionalist identity statement: *to be in pain is to be in state f*. According to Levine, in comparing this statement with the identity statement: *heat is the motion of molecules*, a problem arises in the former case but not in the latter (Levine, 1983, pp. 354-5). The latter statement, says Levine, is fully explanatory in the sense that our knowledge of the physical sciences makes it intelligible that the motion of molecules could play the causal role that we associate with heat (Levine, 1983, p. 357). In the case of the psychophysical statement, however, Levine allows that the causal role of pain can be explicated by a functionalist account but claims

[8] According to Levine, Kripke argued that materialism was false based on two claims: 1) identity statements using rigid designators on both sides of the identity statement must be true in all possible worlds and, therefore 2) if psychophysical identity statements (e.g., pain is firing of c-fibres) are *conceivably* false then, according to 1) they are false (Levine 1983, p. 354).

that the particular feeling of the pain cannot. Therefore, what it feels like to be in state f is not made fully intelligible by understanding the functional properties of state f.[9]

The fact that we cannot explain what the connection might be between a given experience and a given functional state leads to the possibility that the connection is contingent. In other words, in some cases, f does not produce the same – or for that matter, any – experience. There is no clear way of confirming that if you experience pain when you are in functional state f – and if the population of China is (somehow) functionally organized to match this state – that China is experiencing the same pain as you.

It is Van Gulick's intention to disallow the type of liberalism demonstrated by this Chinese nation example by imposing a teleological restraint on his functionalist account of mind. He stipulates that only systems which behave in an adaptive manner with respect to their specific environment can be described using functionalist terminology and this quickly eliminates entities such as the entire population of China, since even if being in state f *does* somehow create the experience of pain for all of China, it serves no adaptive purpose by doing so. However, although Van Gulick reassures us that the teleological restraint eliminates the problem of overly-liberal functionalist accounts, he cannot claim that his teleological model is entirely effective when it comes to dealing with the explanatory gap argument that Levine uses against functionalist accounts of conscious experience.

Van Gulick maintains that Levine's argument – that we are faced with a problematic gap when we try to explain how our subjective experiences result from physical/

[9] My argument in this thesis, however, is that neither what it is *to be in* state f, nor what it *feels like* to be in state f can be made intelligible by understanding the functional properties of state f. In other words, functionalism provides neither a robust account of intentionality, nor a viable description of consciousness.

functional processes – fails because it is based on the notion that qualia are basic simples (i.e., they are without structure of any kind). Van Gulick, however, explains qualia as essential elements within a highly complex representational structure. He suggests, for example, that phenomenal colour space – far from being only arbitrarily connected to some sort of structure – has "a *complex organizational structure* that allows us to establish *explanatory* rather than simply brute fact connections between it and underlying neural processes" (Van Gulick, 1993, p. 145).

With this argument, however, Van Gulick is back into dangerous territory. By relying on the statement that a particular complex neural structure underlies the qualitative experience of (for example) colour in humans to win the point against Levine, he falls prey to a form of human chauvinism. In other words – and Van Gulick himself admits this – the functionalist explanation of first-person experience of colour which he gives, applies (if it applies at all) to humans alone. There is nothing in his argument that eliminates the possibility that entities unlike ourselves (e.g. aliens of some sort) could be in the same functional state (and one which meets a particular teleological requirement) and still have dissimilar (or no) phenomenal colour experiences!

The problem that Levine articulates in relation to psychophysical statements, then, remains valid. We really have no way of determining which (if any) of many possible explanations of our subjective conscious experiences (whether they be couched in functionalist or physicalist terms) are true. Van Gulick acknowledges this problem, but claims that the more we are able to define the phenomenal realm in terms of an overall functional structure, the greater our chances of eliminating this residue of unintelligibility which is left over from the explanation of how representational systems generate

consciousness. He makes the assumption that if the functionalist model of mind hasn't yet provided a *complete* account of how the abstract mind relates to the physical world, it eventually will. In the chapters which follow, I will attempt to show why such an assumption is overly-optimistic.

There is something else to consider here. Van Gulick's functionalist description of phenomenal experience is, as we have seen, modelled entirely on his explanation of internal representation. What started out as a model of representation based on the processing of formal internal symbols has gradually blossomed into an explanation of subjective phenomenal experience. Levine's point is that it is not just physicalism, but likewise functionalism, which is unable to provide a legitimate explanation of our conscious subjective experiences of the world. However, if functionalism can be said to fail in this respect, it seems that it might also fall short of the mark when it comes to Van Gulick's model of the intentional mind since, for him, intentionality and first-person experiences go hand in glove. I am making the suggestion, therefore, that Van Gulick's representationalist account of intentionality suffers from its own explanatory gap![10] We have no way of confirming that his highly abstract representationalist model of mind is in any way accurate as an explanation of what lies behind our intentional mental states.[11]

My conclusion, then, is this: In the claim that our beliefs and desires supervene on an underlying neural structure, Van Gulick's homuncular functionalist account is able to

---

[10] It suffers from a gap in the sense that it fails to provide anything other than a tentative and very abstract account of how the mental states in intentional systems such as ourselves are able to refer to the objects in the world around them.

[11] In other words, there is no way to verify that his abstract model of mind is any better than some other (but quite different) description of intentionality.

make only a highly theoretical connection between our conscious mental states and brain processes. It never explains how the two levels he describes – the intentional and the neurophysiological – might actually relate to one another. My claim is, therefore, that since Van Gulick's functionalist explanation of how the intentional mind operates is flawed and/or incomplete, it is not able to provide any sort of solid base from which to build a convincing explanation of phenomenal consciousness.

Ned Block, who (in contrast to Van Gulick) claims that the functionalist model of mind cannot accommodate phenomenal consciousness,[12] speculates on why functionalist doctrines have gained such widespread acceptance throughout the domain of cognitive science. The functionalist approach, Block claims, was offered initially as a set of hypotheses but with the passage of time these hypotheses – since they sounded so plausible and came with a set of useful features – came to be accepted as established facts (Block, 1978, p. 287). If you repeat the hypotheses of functionalism enough times, says Block, you begin to believe that they're more than just possible – you believe that they're true. Is this the case with Van Gulick? Not exactly. Van Gulick maintains that the best chance we have of being able to explain how mind and brain relate lies in the continual refinement of the functionalist explanation of the representational process that underlies our thoughts and experiences.

Van Gulick must, and does, acknowledge that "there is indeed a residue that continues to escape explanation" (Van Gulick, 1993, p. 145). The question is: is this residue sticky enough to gum up the works for his functionalist account of mind? In the chapters that follow, I will look at two other representationalist models in order to examine the

[12] Block, like Levine, however indicates that he is happy with the functionalist account of intentionality.

issue of whether functionalism can provide an account of intentionality robust enough to act as a viable base for the explanation of the conscious aspect of mind.

Chapter 3

From Content to Consciousness and Back Again: Why Dennett Vacillates Between Explaining and Eliminating Our Conscious Selves

In his recently published book, *Brainchildren*, Daniel Dennett makes (or, rather, reiterates) the claim that the two main topics in the philosophy of mind are content and consciousness. Long before many of his contemporaries began to seek explanations of our first-person experiences of the external world in representational accounts, Dennett was arguing for an explanation of consciousness based on a theory of content. Back in 1969, Dennett proposed that it was the brain's ability to store and manipulate information about the environment (at a sub-personal level) which somehow resulted in our (personal level) experience of being a subject able to observe, understand, and participate in an objective world. Although three decades have passed since Dennett began to investigate these issues in *Content and Consciousness*, a return to his early ideas on content offers insight with respect to his more recent (and, often, hotly contested) writings on the subject of conscious selves. I this chapter, I will argue that, to a very large degree, Dennett's many subsequent discussions on the relationship between content and consciousness are based on an elaboration of the ideas he presented in his early (1969) book.

Several critics have, however, interpreted Dennett's more recent work on consciousness as a betrayal of his earlier – and, in their opinion, valid – approach to intentional systems.[1] In particular, some critics claim that he has reneged on his earlier commitment to intentional-level interpretation and has moved in the direction of explaining conscious entities using a design stance, thus treating selves and the beliefs they might hold with less realism than was previously the case (Sedivy, 1995, p. 48). I will argue, however, that the influence of eliminativism on Dennett's account of mind is not new and can, in fact, be found in his very earliest writings. In this chapter, I will attempt to show that Dennett's agenda has, all along, included a commitment to "progress" from talking about mind in intentional terms toward a more scientifically-credible interpretation of intentional systems using design stance terminology.

I argued in the previous chapter that Van Gulick's representationalist account of mind – although it makes use of constraints such as teleology and homuncularism – is unable to move beyond an entirely speculative account of what it means to be a conscious and thoughtful entity. What is interesting about Dennett's explanation of mind is that it comprises both a functionalist account (one which is very similar to Van Gulick's), as well as a strong commitment to an eliminativist stance in relation to the mental states that it is intended to explain. In this chapter, I will argue that it is Dennett's allegiance to a verificationist approach to the discussion of mind that forces him (some of the time) to deny that our thoughts, selves, and conscious experiences really exist.

---

[1] Dennett is well-known for his proposal that intentional systems such as ourselves – systems capable of generating rational behaviour on the basis of their functional design – can best be interpreted by means of an intentional stance which is to be used as a heuristic device in order to understand and predict behaviour.

3.1 The Issue of Content

As discussed in the previous chapter, Van Gulick uses the term "contentful" to describe states of mind such as beliefs and desires. Dennett also makes extensive use of the term 'content' in his work (e.g., *Content and Consciousness*). What exactly is meant by this term when it is used in relation to metal states? According to Christopher Peacocke, what centrally distinguishes mental states with content is that "they involve reference to objects, properties or relations" in the world (Peacocke, 1995, p. 219). The contents of our thoughts are normally specified using "that . . ." clauses such as the proposition "they are home now" in the propositional attitude sentence "I believe that they are home now."

From a folk-psychological stance,[2] our beliefs and desires (as well as our hopes, fears, angers, etc.) are seen as the reasons that motivate us to behave in one way rather than another. For example, the fact that I believe that you are home now, might influence me to behave in one way rather than another (e.g., I might walk up to your front door and ring the doorbell). In referring to the beliefs and desires of folk psychology to interpret and predict the behaviour of human (and non-human) entities there is, therefore, always an assumption of rationality. In the words of Christopher Peacocke "for a subject to be in a certain set of content-involving states is for attribution of those states to make the subject as rationally intelligible as possible, in the circumstances" (Peacocke, 1995, p. 220). Although it is not uncommon to observe (or participate in) irrational behaviour, "the possibility of irrationality depends on a background of rationality" (Davidson, 1995, p.

[2] Folk psychology is defined as a "common sense" method of interpreting a person's actions and behaviour in terms of their beliefs and desires, etc.

232). The assumption is made that, generally speaking, those around us will act to confirm their beliefs and satisfy their desires.

A folk-psychological interpretation of behaviour does work and we use it every day in order to make our way in the world. But how does it work? What can the possible connection be between physical brain states and the abstract proposition *that Tuesday is the third day of the week*? The issue of how to understand the connection between brain states and the "meaning" of our thoughts is the main problem that Dennett wrestles with – starting in his 1969 discussion of the ascription of content and continuing right up to the present day.

3.2 The Ascription of Content – circa 1969

Given that the bibliography in his latest (1998) book indicates that he has published somewhere close to 60 books and articles since then, why go back to Dennett's original 1969 discussion of content? There are at least two reasons to do so. To begin with, Dennett himself states that his early discussion of content remains the foundation of everything he has published since (Dennett, 1998, p. 355). And, secondly, although the ideas laid out in the chapter entitled "The Ascription of Content" are written in reaction to the behaviourist approach to mind, it is clear that certain behaviourist, or verificationist, tendencies still hold sway throughout the discussion. Since Dennett's 1991 book, *Consciousness Explained*, was criticized by some as leaning too far in the direction of an eliminativist/behaviourist interpretation of mind, perhaps an examination of these early writings will help to determine if this criticism is, in fact, valid.

In *Content and Consciousness*, Dennett sets himself the task of describing the way in which we should think about what is referred to as the intentionality, or the *aboutness* of our thoughts.[3] When attempting to ascribe content, he says, it is necessary to look beyond the afferent effect (i.e., whatever caused the event in question) in order to concentrate on discerning an appropriate efferent effect.[4] To put it another way (and one that should sound better to the functionalist ear), it is what the organism *does* when in a particular mental state that tells us what the meaning of that state/event might be. Dennett emphasizes that we don't know (and won't know) what a given mental state/event means except in terms of an after-the-fact interpretation of any behaviour which results from it.

The issue of the appropriateness of response is central to Dennett's view of the way in which content is to be understood. In his view, the appropriateness of a response is determined by the functional design of the system in question. The ascription of content, therefore, requires a two-step procedure. First, an account needs to be given of the general functional design of the system in order to determine what can be considered rational/appropriate behaviour in its case. Secondly, a heuristic overlay in the form of an intentional characterization of these structures must to be given in order to "flesh out" the functional level account with talk of beliefs, desires, and motivations. For example, in the case of human entities, our "design" has evolved over hundreds of thousands of years to help us adapt to, and survive in, a particular environment. Given this evolved design, it

---

[3] Searle (1995) p. 385 defines intentionality as "that property of the mind by which it is directed at, or is about objects and states of affairs in the world."

[4] Note that Dennett often relies on behaviourist terminology during this discussion. However, his references to *afferents* and *efferents* should not confuse us into thinking that he is falling back into behaviourist ways here. As a centralist, Dennett is committed to giving the mental realm its due. After all, the discussion of how content might relate to mind is one that no strict behaviourist would allow.

is possible to define some behaviours as rational and others as not. Dennett's claim, therefore, is that the behaviour of an organism is to be interpreted in the teleological sense.[5] It must be seen to be functional/purposeful when it comes to helping the organism to cope with the environment it finds itself in.

What are the benefits of interpreting a system or entity in this way? With highly evolved and complex systems, says Dennett, the further "into" the workings of the mind we attempt to go, the more complex and, therefore, "messy" things get. The farther away from the periphery of the nervous system a particular mental event occurs, the more helpful a folk-psychological characterization is likely to be in acquiring any kind of understanding of the relevant functional organization of mental states involved. Without such a strategy, we are left trying to individuate particular mental states which are "compound, ambiguous and apparently continuously changing" (Dennett, 1969, p. 82). Our chances of success in such an endeavour are, according to Dennett, more or less nil.

3.2.1 Problems with the 2-step Solution

Dennett, then, maintains that any ascription of content requires the use of the 2-step procedure outlined above. To reiterate, it is necessary to 1) define the system using a functional (specifically teleological) design, and then, 2) interpret the output of the system based on the supposition that the system will behave rationally according to the design in question. There are, however, serious problems inherent in both steps. Dennett describes step one – the process of individuating neural structures by functional design

[5] Dennett uses the same restraint on his functionalist account that Van Gulick did, as noted in the previous chapter.

– as somewhere between extremely difficult and "all but impossible" (Dennett, 1969, p. 82). Given the complexity of the neurophysiological workings of the brain – in relation to the multitude of inputs, outputs, and possible interference that take place between the two – the provision of anything more than a highly abstract description of neural processing seems unlikely since the functions to be defined are always global, not local – i.e., they refer to the system as a whole and not to discrete neurological happenings. The second step in the process of ascribing content – the provision of an intentional interpretation for the functional design in question – is, according to Dennett, just as problematic as the first.

Dennett's claim, then, is that the complexity of the design of the intentional systems we are trying to understand precludes the possibility of being able to ascribe content to any of their specific functional states. It seems, however, that he wants to go one step further with his claim. He writes that even if it were possible to determine at "what level of the afferent stimulus analysis in the neural net" (Dennett, 1969, p. 83) a neurological signal becomes contentful (i.e., by referring to a specific object in the external world), there is no verifiable way of knowing what that content might be. This is to say that, for all (of Dennett's) intents and purposes, the content he is ascribing does not exist in any ontological sense whatsoever. There is no such thing as content in the physical world.

The ascription of content, therefore, entails nothing more nor less than describing a given behavioural event using a particular verbal expression (Dennett, 1969, p. 82). For example, says Dennett, suppose that you observe Fido refusing to walk out onto thin ice in order to retrieve a succulent beefsteak. Notice that in attempting to provide an intentional description that matches the functional interrelations of Fido's nervous system, you are able to use nothing more precise than opinions expressed in ordinary language

(Dennett, 1969, p. 85) – probably something like "Fido's fear of thin ice keeps him from retrieving the attractive beefsteak." Considering that Fido's neural activities are entirely unsullied by concepts such as "thin ice" and "beefsteak", it can be said that this intentional description of Fido's bad experience is nothing more than an extremely imprecise interpretation of his current state of mind. We are simply using a heuristic device, says Dennett, in order to help us understand what might lie behind Fido's failure to behave as expected. For Dennett, then, content is what we, as observers, provide when we interpret the behaviour of the entities we are observing. The content is (in a sense) ours, and not theirs.

3.2.2 *Content as the harbinger of troubles to come*

Dennett claims that his insistence on making a clear distinction between the personal and sub-personal comes from Ryle and Wittgenstein who warned that these two levels of explanation must not be confused. But, according to Dennett, just because beliefs and desires are said to be in one category, or domain, of inquiry while neural events are in another, does not mean that a hard wedge should be driven between the two. Although there is no rigorous, or empirically sound, way of ascribing content to an intentional system, Dennett insists that there is no need to succumb to the admission of any sort of unbridgeable gap between the mental and the physical.[6]

The acknowledgment that we are dealing with two different levels of explanation, however, immediately spawns the need for some way of understanding the relationship

[6] As Sedivy points out in her article "Consciousness Explained: Ignoring Ryle and Co.," Dennett seems determined that – in spite of the conclusions of Wittgenstein and Ryle – when it comes to understanding mind, we should not "isolate the philosophical from the mechanical questions" (Dennett, 1969, p. 95).

between them. Dennett's solution is to claim that it is the terms used to describe beliefs and neural events (rather than the beliefs and neural events themselves) which refer to different ontological categories. What is the difference between these two categories? Terms which are used to describe neural events refer to the physical world, while the terms used in talk about beliefs and desires and the experience of pain are non-referential (Dennett, 1969, p. 95). Literally, there is nothing (i.e., no *thing*) for them to refer to.

It is here (way back in 1969) that the beginnings of Dennett's (still unresolved) conflict in relation to the explanation of content and consciousness can be found. On the one hand, it is clear that he agrees with Wittgenstein and Ryle that personal and sub-personal explanations should not be equated/confused. On the other hand, he is determined that the sub-personal level should not be closed off to philosophical inquiry. This "wanting it both ways" on Dennett's part is, I suggest, the source of his ongoing dilemma. At the same time that he is claiming that content can be understood only in the context of an intentional interpretation of behaviour, he is pushing to provide an explanation of content based on the functional design of a system.

What is the result of this unresolved problem? When Dennett is in one mood, he writes about the usefulness of interpreting intentional systems using an intentional stance; when he is another, he is inclined (in fact, forced) to take a stronger eliminativist stance with respect to any non-physical entities that are likely to get in the way of an "empirical" explanation. When Dennett wants to talk science, beliefs and desires, and even selves, must run for cover.[7]

[7] Dennett's approach to content often swings dangerously close to that of the well-known views of Paul Churchland who claims that our vocabularies will eventually evolve to the point where reference to non-existent entities such as beliefs and desires will no longer be necessary.

How, then, are we to understand the discussion of the ascription of content which Dennett undertakes in 1969? Is intentional interpretation the only available means of explaining our own and others' behaviour, or is it simply a heuristic device which can be discarded once we have a better (scientific) account of our functional design? In my opinion, Dennett himself is not entirely certain of the answer to this question.

3.3 Mirror, Mirror on the Wall: Which is the Loveliest Stance of All?

The same doubts with respect to the legitimacy of intentional interpretation appear to surface again in the article entitled "Intentional Systems." In this article, Dennett expands on his theory of how the ascription of content can help us to interpret certain complex intentional systems. He maintains that there are three ways of approaching, or understanding, such a system. We can take 1) a physical stance, 2) a design stance, or 3) an intentional stance to the design in question. In other words, an attempt can be made to explain and predict the behaviour of an intentional system in terms of its physical components, in terms of its design components, or by performing a folk-psychological interpretation of its behaviour based on the assumption that it will behave according to certain rational principles which are defined by a general understanding of its functional design.

As materialists, it would seem that a physical stance would be the stance of choice if we require a totally accurate description of the events which take place in our world. As Dennett points out, however, such an approach is possible only in the case where the processes in question can be worked out on the basis of our knowledge of the laws stipulated by physics, biology, and chemistry. This is easy enough to do in a very simple

or straightforward case (e.g., "it was the force of the moving baseball that shattered the glass of the window") but there are many, many more cases in which the taking of a physical stance gets us absolutely nowhere. For example, says Dennett, we would be quickly overwhelmed when attempting to describe the workings of even a simple computer by referring to the actual physical events which take place during its operation. Taking a physical stance, then, is generally not feasible given the complexity of most of the systems we are attempting to understand.

Alternatively, says Dennett, we might attempt to understand the output of an intentional system by taking a design stance in order to describe/predict its behaviour. In this case, however, we would have to have a detailed knowledge of all of the functional components and sub-components for the system in question; and, secondly, we would have to make the assumption that the system would perform as designed and without breakdown. But, again, these requirements are clearly beyond our reach in most cases. Most of us would have trouble providing an explanation of the operations of even a very simple system from a design stance. In the case of any sort of complex system (e.g., a chess-playing computer) Dennett notes that even those knowledgeable about the design in question (e.g., programmers, engineers) sometimes find it very difficult to predict the output/behaviour of the system using a design stance.

So what is the solution? Just as outlined in *Content and Consciousness*, Dennett advises that we take an intentional stance when trying to comprehend a system with any sort of complex design. When it is not possible "to beat the machine by utilizing one's knowledge of physics or programming to anticipate its responses, one may still be able to avoid defeat by treating the machine rather like an intelligent human opponent" (Dennett, 1978, p. 5).

As with a human entity, the most productive way of interacting with a system is to assume 1) that it will function as designed, and 2) that its design is such that it will (almost) always select the most rational move (Dennett, 1978, p. 5).

Dennett makes the claim in "Intentional Systems" that the decision to use one strategy, or stance, rather than another should be entirely pragmatic. In other words, the stance you select depends on who you are and how you are trying to relate to the particular system in question. For example, in the case of the chess-playing computer, the repairman might take a physical stance, the programmer/designer might take a design stance, and the chess-playing opponent would likely take an intentional stance (Dennett, 1978, p. 7).

All of this sounds very reasonable, and Dennett's account of stances is considered by many to demonstrate a certain flexibility and "open-mindedness" because it appears to acknowledge that there are different, but equally valid, ways of looking at the output/behaviour of intentional systems. I want to suggest, however, that Dennett is not (and never was) as open-minded with respect to the stances as he is sometimes seen to be. In *Content and Consciousness*, he seemed to argue that the intentional interpretation of behaviour, by means of the ascription of content, is a valid psychological project. In "Intentional Systems," however, he writes: "Where, then, should we look for a satisfactory theory of behaviour? Intentional theory is vacuous as psychology because it presupposes and does not explain rationality or intelligence" (Dennett, 1978, p. 15).

The intentional stance is helpful, says Dennett, but it doesn't explain anything – it isn't empirical and for Dennett, an empirical scientific theory of behaviour is the ultimate goal. "In the end," he says, "we want to be able to explain the intelligence of man, or beast,

in terms of his *design* [my italics]"(Dennett, 1978, p. 12). What the intentional stance really does, is show us where our theory of behaviour for a given system is incomplete. Since it is "based . . . on no particular picture of the system's design, [it] cannot be construed to confirm or disconfirm any particular pictures" of this design (Dennett, 1978, p. 13).

What exactly does Dennett mean by design here? Is he referring to the somewhat general, but teleological, notion of what the system was designed to do (e.g., find food and survive predators or play chess)? As described in *Content and Consciousness*, a general notion of design is used to justify an interpretation of rationality and, in this case, the design stance can be seen as simply the prerequisite step before taking the intentional stance.

In "Intentional Systems," however, Dennett moves in the direction of defining the design stance in much more exacting terms (e.g., using specific functional modules and sub-modules, and so on). In other words, he appears to be operating with two different definitions of design – a general notion of design which is used to determine what is rational in a given case of intentional interpretation, and a full-fledged design stance in which the behaviour of a system can be predicted using specific and clearly-defined functional modules. This latter definition seems to be the one Dennett has in mind in "Intentional Systems." Here he states that theory builders should (whenever possible) dispense with intentional explanations and move on to "more reliable design-stance explanations and predictions" (Dennett, 1978, p. 12). In *Content and Consciousness*, a design stance was discussed as a sort of prerequisite step before coming up with an intentional interpretation of a system's behaviour. In "Intentional Systems," however, this

prerequisite step gets promoted to a stand-alone approach and one that, given time, should eliminate the need for intentional stances.

I am arguing, then, that in "Intentional Systems" (and elsewhere) Dennett appears to waver with respect to the legitimacy of the intentional stance. In spite of providing long and detailed explanations of why this stance is essential when trying to interpret complex systems, and in spite of his claim that no stance is "intrinsically right or wrong" (1978, p. 7), Dennett's bias in favour of the design stance does tend to shine through. For example, he talks of the "intelligence loans" (Dennett, 1978, p. 12) that must be taken out whenever we describe a system using the intentional stance since it is necessary to borrow (some unexplained) intelligence from somewhere in order to talk in intentional terms. Eventually, says Dennett, these loans will have to be paid back with an explanation based on the cold, hard facts of empirical science.

In order to understand why Dennett sees the design stance as having greater value than the intentional stance, we have to acknowledge his particular and personal priorities. In "Self Portrait" (1998) he writes, "I have always been fascinated with how things worked – clocks, engines, magic tricks. (In fact, had I not been raised in a dyed-in-the-wool 'arts and humanities' academic family, I probably would have become an engineer . . ." (Dennett, 1998, p. 356). Dennett prides himself on being a philosopher who begins his philosophical investigations from a "base camp in the sciences" (Dennett, 1995, p. 242). As he has reiterated in most of his writings, he is committed to operating strictly within the confines of an objective, materialistic, third-person point-of-view (Dennett, 1995, p. 237). The ultimate concern, therefore, is to get down to empirical facts and this, according to Dennett, is something we will never do using the intentional stance.

There is an important point in relation to the design stance, however, that Dennett sometimes ignores. When it comes to understanding the actions/behaviour of a complex intentional system, it can be said that a design stance is just as much a heuristic device as any intentional stance. Just as the ascription of beliefs and desires can be used to help us predict the behaviour of systems whose design is too complex for us to deal with, so the design stance can be seen as simply another attempt to overcome our lack of understanding of the way in which physical processes lead to our thoughts and behaviour.

A particular functional design, consisting of a hierarchy of modules and sub-modules, does nothing more than describe the way in which the production of thought *might* take place. The fact is, we are not even close to understanding the design of the human mind in functional terms. It seems believable that evolution designed and re-designed the responses of certain entities in such a way that they were able to interact successfully with the world around them. At this point, however, taking a design stance with respect to our behaviour requires just as much (and often more) supposing as any intentional interpretation.

Dennett himself (sometimes) admits this. He sometimes does acknowledge that the design stance, just like the intentional stance, relies heavily on the use of metaphor. In *Content and Consciousness*, Dennett came out with the strange claim that "A computer is no more *really* an information processor than a river *really* ha[s] desires" (1969, p. 90). In other words, he has from the beginning (1969) recognized that the reason we talk about computers as "processing information" is because describing what they actually do using the terminology of physics would be impossible for (and meaningless to) us. Talk of beliefs and desires – and even information processing – is simply a way of helping us to

understand events in our world which, when described in purely physical terms, are too complex for us to comprehend.

3.4 The way in which beliefs are (sort of) real

Dennett's discussion of the intentional versus the design stance might lead us to conclude that he rejects realism with respect to beliefs. If only it were that straightforward! Dennett addresses the issue of realism in more detail in an article entitled "Real Patterns" in which he claims that he takes an "intermediate" position between realists, such as Fodor – who maintain that beliefs correspond to specific physical structures in the brain – and eliminativists, like Paul Churchland, who deny the reality of beliefs entirely (Dennett, 1991b, p. 30).

Dennett explains his position as follows. Beliefs have no ontological reality. They don't exist any more than abstract objects such as "centers of gravity" or "Dennett's lost sock center."[8] What is real, however, are the patterns of behaviour that are discernible when taking an intentional stance. The intentional stance works because it provides us with a reasonably reliable tool for predicting the behaviour of intentional systems.

In "Real Patterns," Dennett compares the patterns of behaviour that are discernible from an intentional stance to the patterns that are produced by computer software such as that used for an implementation of The Game of Life. In this game, visible and consistent patterns of movement are produced on the screen while "at the physical level there is no motion, and the only individuals, cells, are defined by their fixed spatial

[8] Dennett (1991b) p. 28. The latter is defined as "the center of the smallest sphere that can be inscribed around all the socks I have ever lost in my life."

location" (Dennett, 1991b, p. 39) What we see – the "motion of persisting objects" (1991b, p. 39) – seems real and (if we know how the program works) we can predict what movement will happen next quite reliably. Dennett, then, is claiming that these patterns have a certain reality - the same kind of reality as do the patterns of our behaviour when interpreted using abstract objects such as beliefs and desires. Note that in observing the output produced by The Game of Life software, the content of our visual experience is entirely dissimilar to the underlying processing (of rules) that generates the moving objects we see. Likewise:

> The process that produces the data of folk psychology . . . is one in which the multidimensional complexities of the underlying processes are projected through linguistic behaviour, which creates an appearance of definiteness, and precision, thanks to the discreteness of words. (Dennett, 1991b, p. 45).

Dennett, then, takes an intermediate position on realism with respect to belief. In acknowledging both sides of the argument, however, he fails to resolve the issue. What Dennett admits is useful (i.e., the intentional stance) has no real scientific credibility and Dennett, as we know, is committed to the provision of a model of mind which is empirically sound and scientifically impeccable.

In spite of his claim that the intentional stance gets the job done, Dennett is strongly motivated, therefore, to come up with an empirically sound theory of behaviour that will ultimately do away with any description of mind based on the fictions of the intentional stance. It is this distinct bias in favour of the design stance that Dennett carries over into his attempt to explain what lies behind our conscious experience of the world.

3.5 From Content to Consciousness

Why have I devoted so many of the preceding pages to the discussion of Dennett's theory of intentionality? I maintain that in order to make sense of why he proceeds as he does in *Consciousness Explained*, it is necessary to have a clear understanding of his view with respect to content. According to Dennett, intentionality is a more fundamental phenomenon than consciousness since only intentional systems can be described as conscious. Any theory of consciousness, therefore, must be built upon the foundation of a solid theory of intentional content.[9]

As stated earlier, *Consciousness Explained* has sometimes been interpreted as a betrayal, or turnaround, in relation to Dennett's earlier writings on intentional systems (Akins, 1996, p. 184; Sedivy, 1995, p. 481). Dennett, however, maintains that *Consciousness Explained* is the complimentary volume to *The Intentional Stance* (which fills out the explanation of content that Dennett gave in *Content and Consciousness).*[10] With *Consciousness Explained*, Dennett attempts to "take the next step" and rely more extensively on design stance terminology in order to explain what lies behind our experience of being a conscious subject.

My argument, contrary to the critics mentioned above, is that in attempting to provide a theory of consciousness, Dennett does more or less what he claims he intended to do all along: he first provides a theory of content and then a theory of consciousness,

---

[9] Dennett claims that most philosophers see things the other way around – i.e., intentionality is seen to be dependent on consciousness which is considered the fundamental phenomenon (1995, p. 236).

[10] In other words, *The Intentional Stance* (1978) expands on Dennett's early ideas on content, while *Consciousness Explained* (1991a) provides the complementary explanation of consciousness (Dennett, 1998, p. 355).

with the second theory based on the first.[11] In tracing back through his earlier work, it can be seen that Dennett's approach to consciousness is, in fact, closely dependent on the ideas that he has developed over the years in relation to the different stances that can be taken when it comes to describing an intentional system. Dennett, then, has not exactly reneged on his commitment to the intentional stance in *Consciousness Explained*. Keeping in mind the agenda he outlined in "Intentional Systems," – that theory-builders should always move in the direction of using the design stance in their explanations – his eagerness to "move ahead" and take a design stance in *Consciousness Explained* is not all that surprising. Dennett is operating on the supposition that both his audience, and the problem itself, are now ready for a design stance explanation.

3.5.1 The general claim

In *Consciousness Explained*, Dennett makes the claim that it is the simultaneous, or parallel, processing of a multitude of very sophisticated and complexly-related representations that lies behind our *sense of* being a conscious self.[12] The conscious experience of ourselves as subjects operating in an objective world is essentially tied up with our ability to somehow "process information" about the world around us by means of a complex control-system that allows for recursive self-representation (Dennett, 1991a, p. 310). In other words, part of what our representational system keeps track of is the relationship between our particular physical body and the environment in which it is located.

---

[11] See Dennett, 1998, p. 355 for a discussion of how he has proceeded as planned in his published works.

[12] Note, again, the similarity between Dennett's functionalist account of the sense of "self" and Van Gulick's (described in the previous chapter).

Dennett maintains that underlying our sense of being a conscious and subjective self, there lies nothing more nor less than the operation of a type of very complex intentional system described in earlier publications. In part III of "Intentional Systems," Dennett makes the claim that it is for the subclass of intentional systems that have language and can communicate that we need to address the issue of consciousness. "The appreciation of meanings – their discrimination and delectation – is central to our vision of consciousness" (Dennett, 1995, p. 237). Likewise, in "Real Patterns," Dennett emphasizes that it is our linguistic capability that is central to our sense of being a conscious subject capable of thinking thoughts about the external world (Dennett, 1991b, p. 45).

Dennett has already acknowledged that the content of our beliefs, etc. cannot be traced back to specific mental events and/or internal representations (as in some Fodorian language of thought scenario). The content of our beliefs is not real – at least not in the sense that it corresponds (in any verifiable manner) to anything taking place in our brains. Likewise, the "stream of consciousness"[13] which we experience, has no reality in the physical world since neither the "stream", nor the self that experiences it, can be traced back in any legitimate way to physical brain processes. Dennett's claim is that the sense we have of being a conscious subject capable of entertaining beliefs and desires is nothing more than an illusion. The difference in his 1991 approach to the illusory nature of thought and self is that, in *Consciousness Explained,* he proposes to give us some design-level details to help us understand what lies behind these illusions.

[13] In describing this "stream of conscious experience," Dennett refers to the "meandering sequence of conscious mental contents famously depicted by James Joyce in his novels" (Dennett , 1991a, p. 214).

3.5.2 The intentional stance resurfaces

In *Consciousness Explained,* the intentional system is presented using specific design terminology (e.g., parallel processing of a complex control system capable of recursive self-representation). However, Dennett doesn't – in fact, isn't able to – escape the need for intentional interpretation when it comes to explaining consciousness. In fact, an important part of his theory of consciousness corresponds to his previous claim that in order to understand the conscious thoughts of another entity, it is necessary to take an intentional stance so as to "make sense of" their behaviour. In *Consciousness Explained,* the behaviour in question consists of a verbal report of conscious thoughts and sensations. What is strange, however, is that when it comes to collecting data on the subjective thoughts and experiences of a particular intentional system, Dennett now makes the claim that he can take an intentional stance in relation to the system's verbal output while "never abandoning the methodological scruples of science" (Dennett, 1991a, p. 72). His proposal for capturing the phenomenal experience of a conscious "subject" is as follows. Make multiple recordings of the conversation you have with the subject and have transcripts prepared by three different stenographers in order to ensure that the data you accumulate remain (relatively) immune to bias and over-interpretation (Dennett, 1991a, p. 75).

But something seems to be wrong here. Dennett's suggestion that such a transcript would contain "valid" data misses the point that he has previously (and consistently) made with respect to intentional interpretation.[14] Dennett has always insisted that any particular interpretation of behaviour – verbal or otherwise – can never be verified/confirmed using empirically sound methods. The suggestion that there is an

[14] As discussed above, he makes this claim in both *Content and Consciousness* and *Brainstorms.*

objective way to go about collecting data on subjective experience which will ensure that the data in question have any kind of empirical integrity is misplaced here. If phenomenology is not "among the data of science," (Dennett, 1991a, p. 71) he should not try to convince us that this particular form of intentional interpretation has any scientific credibility, especially since what we are ultimately told to do with this transcript is to treat it as a "work of fiction" (Dennett, 1991a, p. 79). The reason he should not is that, by his own theory, the data collected consist of nothing more than the transcription of the fictional beliefs and desires of an illusory conscious self.

Dennett's attempt to put a different "spin" on intentional interpretation in this case is based on his decision in *Consciousness Explained* to "forge ahead" with a design stance approach to consciousness. He simply has trouble reconciling what he calls his "empirical theory of mind" with the fictional data it must rely on. But, surely, this – the problematic relation of abstract (or fictional) content to physical brain matter – is the crux of the matter! It is in order to resolve this relation (of mind and brain) that Dennett writes the book to begin with. His attempt to legitimize, or "clean up", the intentional stance here can be interpreted in two ways. To begin with, it can be seen to highlight the fact that Dennett's approach to the three stances he has defined in his earlier work is unstable. Secondly, it seems to indicate that Dennett is, after all, not entirely satisfied with his empirical theory of mind.

Another important point to consider here is this. As Dennett admits, there is no way of knowing whether the subjects describing their conscious thoughts and experiences are actually having these experiences, or are simply a talking zombies. Given Dennett's approach, any intentional system that has a design complex enough to produce the kind

of output (language) described above must be considered to be conscious "in the fullest sense" (Dennett, 1991a, p. 221). For Dennett, the notion of a conscious self is simply another fictional notion which is useful when trying to understand highly sophisticated intentional systems – whether these consist of a particular software/hardware configuration or of flesh and blood. As many critics have pointed out, however, this appears to be more of a denial, rather than an explanation, of phenomenal consciousness. Dennett has long held the opinion that abstract notions such as beliefs and desires and first-person experiences are not reducible to physical brain states. From his point of view, therefore, the only scientifically respectable way of dealing with them is elimination. This raises the issue of (and confusion about) whether Dennett should count himself as an eliminativist or instrumentalist when it comes to intentional mental states.

3.5.3 Teleological function and hardware/software considerations

I have been arguing that Dennett's explanation of subjective, or phenomenal, consciousness does not reject, but rather incorporates (in a somewhat strange order) all of the theories and ideas he has been developing since *Content and Consciousness* in 1969. For example, as discussed above, Dennett's original views on how and when to ascribe content play a major role in his explanation of consciousness. Likewise, Dennett's 1991 model of consciousness can be seen to rely heavily on his earlier views with respect to how intentional systems gradually develop a teleological functional design.

According to Dennett's theory, systems which have evolved to the point where they can accommodate language, are able to pick up an already invented and largely debugged system of habits (e.g., the alphabet, the wearing of clothes) and modify this as required.

This process results in the creation of what Dennett refers to as a "virtual machine" (i.e., software) which runs on the parallel network architecture of a brain (Dennett, 1991a, p. 214). It is the processing of this virtual machine – described by Dennett as a continuous stream of self-probings – which provides the intentional system with the sense that it is a subjective self who has first-person access to its own thoughts and sensations.

For Dennett, then, the sense you have of being a real self is simply an illusion created by the complex and highly-evolved representational system which keeps "you" informed on an ongoing basis about the current state of your environment and your body and the connection between the two. You believe you are a self (with a history and an individual set of proclivities, and so on) but beliefs – as Dennett has told us from the beginning – don't really exist.

In *Consciousness Explained,* Dennett clearly moves on from the notion of teleological design as a means to and end (i.e., in which a general notion of a functional design is used to facilitate the intentional interpretation of what is assumed to be rational behaviour). Here, a more fully-developed design stance has become an end in itself. Dennett appears to be saying that we now have enough scientific knowledge to come up with a design description that is robust enough to fill in some of the gaps that the intentional stance leaves behind.

3.5.4 The Irony of Multiple Drafts

It is during his discussion of the multiple drafts model of consciousness that Dennett gets down to the specific technical details of how the virtual machine might (or might not) operate on the hardware of our physical brain. In the first chapter of the section entitled

"An Empirical Theory of the Mind," Dennett introduces the multiple drafts model of consciousness as a scientifically legitimate alternative to the misguided approach of the Cartesian materialist.[15] In describing his proposed model of consciousness, he states:

> All perceptual operations, and indeed all operations of thought and action, are accomplished by multitrack processes of interpretation and elaboration that occur over hundreds of milliseconds, during which time various additions, incorporations, ememdations, and overwriting of content can occur, in various orders. (Dennett, 1991a, p. 185)

In other words, according to Dennett, there is no central location to which the "information"[16] resulting from one or more perceptual detections "is sent" for the purpose of "re-presentation." The likelihood of a particular perceptual discrimination becoming conscious, as well as the way (e.g., the temporal order) in which it becomes conscious (assuming it does become conscious) depends entirely on what else is going on in the brain (in parallel) at that moment. Multiple, and parallel, "drafts" relating to a particular experience are all potentially available and, says Dennett, no particular one of them can be singled out as canonical. He concludes, therefore, that there can be no definitive version of "what it is like to be" in a particular mental state.

Multiple drafts has sometimes been criticized as attempting to reduce subjective conscious experience to the workings of the functional design, or hardware/software configuration, which Dennett proposes. I maintain that, on the contrary, Dennett would not allow for the reduction of content and/or conscious experience. Clearly, his message

---

[15] According to Dennett, Cartesian materialism comprises a faulty set of concepts about the conscious mind. For example, he says, the Cartesian materialist defines the conscious mind as a kind of "locus of subjectivity" (Dennett, 1991a, p. 255) – some kind of central observer, or point of view, which must reside in a particular location in the brain which he equates with the Cartesian theatre.

[16] I have enclosed certain terms used in this description of multiple drafts in quotation marks to point to the fact that Dennett's explanation (although supposedly given in design terms) must still rely heavily on metaphor to get his point across.

all along has been that our beliefs and subjective experiences can never be reduced to specific brain states or, for that matter, to specific functional modules/locations or sub-modules in a given design. In fact, Dennett has been quite consistent over the years in his claim that the content we ascribe to our mental states has no ontological reality whatsoever. An observer can interpret our behaviour according to certain reliable patterns of rationality which, although governed by design, never reduce to specific functional and/or brain states. To repeat: they do not refer at all. That is what the intentional stance tells us. This makes it seem somewhat contradictory and very confusing, therefore, that Dennett attempts to take a design stance in order to explain what he has described as entirely illusory.

There is a certain irony at work in Dennett's explanation of the multiple drafts model. In attempting to talk about conscious experience using a design stance, he has run directly into a substantial problem – the exact same problem that he (in previous writings) described as requiring an intentional stance solution. His basic claim in relation to multiple drafts is that what your experiences are about cannot be referred back to any specific time (of occurrence) or location in the brain since neither the experience nor what it seems to be about has any ontological reality. In following Dennett's description of multiple drafts, the reader is forced to conclude that the ongoing simultaneous, parallel and competitive processing which occurs in the (pandemonium) model is so complex that any attempt to take a design stance in order to determine what is being experienced when, is futile. In this sense, multiple drafts ends up making a strong argument in favour of the intentional stance. The only conclusion it is sensible to draw from the design stance approach used in *Consciousness Explained* is that if you want to get an idea of what it is like for a conscious

entity (with language capability) to experience the world, your best bet is to take an intentional stance towards the verbal report of the subject in question.

In spite of Dennett's claim in *Consciousness Explained* that he is presenting an empirical theory of mind, it is clear, in reading carefully through this book, that no such claim can be made. As Akins puts it, Dennett is a philosopher who is "trying to float two separate projects that drift in different directions" (Akins, 1996, p. 189). Dennett the philosopher (author of the intentional stance) and Dennett the would-be engineer (Dennett, 1998, p. 356) (proponent of the design stance) seem to be uneasy companions throughout this book. In the end, the multiple drafts model is unable to make any clear connection between subjective consciousness and the processing of the virtual machine. This is a destructive rather than a constructive project (Sedivy, 1995, p. 455) – i.e., multiple drafts tells us what phenomenal consciousness *isn't* but it falls well short of providing a constructive empirical model of what it *is*.

In "Intentional Systems," Dennett made the claim that intentional interpretation of a system's behaviour requires the taking out of a "loan" on intelligence. In other words, you explain a system in intentional terms only in the case that the system's design is too complex for you to come up with any concrete sort of explanation. In *Consciousness Explained*, Dennett indicates that he is now ready to pay back the loans (of intelligence) that he has taken out over the last decades by providing the reader with a fully scientific account of what it means to be conscious. However, the fact that the explanation of his empirical model of mind must rely on terms such as "drafts" and "pre-publication editing" can only lead to the conclusion that it has backfired when it comes to any sort of loan payback.

3.6 Conclusion

More than one critic has made the claim that Dennett's empirical model of mind is unable to deliver when it comes to providing a viable design stance explanation of conscious experience. In "Ships in the Night," Akins takes Dennett to task for failing to provide the details of how his computational model of consciousness might actually produce phenomenology. Likewise, Sedivy, in "Ignoring Ryle and Co.," rightly points out that Dennett "does not provide an adequate case for the identification of conscious mental episodes with functional organizations of brain states" (1995, p. 458). As he nears the end of his explanation in *Consciousness Explained*, however, Dennett appears to anticipate these very criticisms. In spite of the emphatic tone he takes during much of the book, he begins to admit that his proposed model of consciousness may not have, in fact, provided a truly scientific explanation of how conscious mind and physical brain relate. He acknowledges that he has simply replaced "one family of metaphors and images with another" (Dennett, 1991a, p. 455). When it comes to understanding what lies behind our conscious experience of the world, is the idea of a "virtual machine" translating "multiple drafts" any more accurate and/or helpful than the idea of a self who is somehow able to make sense of, and respond rationally to, the world it experiences? Dennett, obviously, thinks that it is but even he admits that we are still in the "metaphor and hand-waving stage" (Dennett, 1991a, p. 275).

My argument, then, is that Dennett's explanation of consciousness does not represent any betrayal, or turnaround, when it comes to the agenda he laid out in his earlier work on intentional systems, and that his way of viewing the connection between mind and

brain has remained consistent throughout the many books and articles he has published. The problems that rise to the surface in *Consciousness Explained* in relation to design-stance explanations of fictional entities – such as beliefs, pains, and selves – were evident back in Dennett's 1969 discussion of the limitations of intentional-level interpretation of systems.

Dennett's commitment to a behaviourist/verificationist approach to mind requires that he must always keep a certain "metaphoric" distance from the physical brain states that are presumably responsible for our thoughts. But herein lies another (and perhaps more serious) problem. Dennett himself writes that in basing a model of consciousness on a theory of content, it is first necessary to ensure that the theory of content is sound. Dennett's theory of content, however, never explains how it is possible for us to entertain thoughts about objects and events in the world. It simply provides a way of getting around the fact that we have no (legitimate) explanation of the intentional nature of mind.

Dennett asks, "Are there mental treasures that cannot be purchased with intentional coin?" (Dennett, 1978, p. 16). These words imply that an account of intentional content is all that is needed in order to come up with a legitimate explanation of our conscious subjective experiences. Although Dennett apparently feels that his explanation of intentionality is robust enough to warrant his moving on to an account of consciousness, I maintain that his explanation is not stable (or problem-free) to the extent that it can provide a solid base for an account of the conscious properties of mind. In spite of the significant efforts on the part of functionalists such as Dennett and Van Gulick, a comprehensive and/or believable explanation of intentionality has yet to be established. Until it has, we must be extremely circumspect when it comes to using any "explanation" of content to provide an account of our conscious subjective experiences of the world.

Chapter 4

Natural Reasons – Problems in Fred Dretske's Representational Theory of Mind

To most of us, a folk-psychological explanation of behavior – in which our beliefs and desires are seen as having causal efficacy in relation to our behavior – seems exceedingly obvious. For example, it just seems to make sense that I open the door because I believe there is someone behind it and I wish to see this person. To many philosophers involved in the philosophy of mind, however, this common-sense view of what causes us to behave in one way rather than another is highly naive and likely erroneous. It is certainly true that the job of explaining exactly how our mental states relate to the neurophysiological workings of our brains has proved to be extremely challenging.

In the previous chapters, I have taken a look at two materialist philosophers (Robert Van Gulick and Daniel Dennett) who, having developed an account of how the intentional, representational mind relates to the physical brain, have gone on to attempt a solution to the "problem of consciousness" based on this account. As discussed in chapter 2, Van Gulick takes a "more of the same" approach. His claim is that if a representational system possesses the prerequisite level of sophistication in the form of complexly-defined and interrelated representational states, conscious awareness of self versus world will result. I argued in chapter 2, however, that Van Gulick must fill in

certain gaps in his representational account before he can use it as a base on which to build a model of consciousness.

In chapter 3, I tried to show how Daniel Dennett's model of mind moves progressively in the direction of a commitment to eliminativism. I discussed how Dennett's more recent attempt to move on from his well-known "intentional stance" to a "design stance" position in relation to first-person conscious experiences was unsuccessful since Dennett's "empirical" model of mind does nothing so much as highlight the problems inherent in any attempt to locate or describe the physical events that supposedly underlie a particular conscious thought. My argument was that multiple drafts, in fact, works to validate Dennett's original argument in favour of taking an intentional stance and forces him even further in the direction of eliminativism with respect to beliefs, desires, and selves.

Fred Dretske, on the other hand, sees no reason to be so "skittish about belief" (Dretske, 1988b, p. 511). He chides instrumentalists such as Dennett by referring to the words of J. L. Austin who stated that "it would be silly to hedge one's realism about dreams, numbers, and thoughts simply because they lack the properties of ordinary dry goods" (Dretske, 1988b, p. 511). In his writings, Dretske has worked diligently to do what philosophers such as Dennett describe as entirely misguided – he has attempted to "naturalize" the mind. According to Dretske, the right kind of representational model of mind can provide a an entirely natural account of how the beliefs and desires, etc. that comprise our mental states relate to the physical workings of our brain. For Dretske, your belief (e.g., that there is beer in the fridge) and your desire (e.g., for a cold drink) are

somehow able to influence/cause the sequence of physical events in your brain which eventually result in your going to the fridge to get that beer. Dretske, of course, is careful to stipulate that it is not the (abstract) belief, itself, that causes a given physical event. Rather, it is the way in which your brain is able to *represent* the fact that there is beer in the fridge that results in your heading to the kitchen.

In his 1995 book, *Naturalizing the Mind*, Dretske presents a model of conscious experience which is based on his earlier work on natural representational systems. In this, his work follows the same pattern of development which I have argued is apparent in the work of both Van Gulick and Dennett. In order to determine whether Dretske's approach to representational systems is able to explain the connection between our brain processes and our thoughts about (and subjective experiences of) the world, I want to begin by looking at his 1981 *Knowledge and the Flow of Information*.

4.1 Dretske's 1981 Information Theoretic Model of Intentional Systems

In his autobiographical entry in Guttenplan's *A Companion to the Philosophy of Mind*, Dretske explains how his philosophical focus evolved over the years from the study of epistemology to philosophy of mind and acknowledges that the information theoretic account he presented in the 1981 book *Knowledge and the Flow of Information* – although adequate as a theory of knowledge – was somewhat "short on details" (Dretske, 1995a, p. 262) when it came to explaining how beliefs could actually *cause* behavior. Nevertheless, many of the ideas which Dretske proposes in 1981 retain a strong influence in his subsequent explanation of the intentional and conscious nature of representational systems. For this reason, it is helpful to look at his description of how physical structures

in the brain develop the capability of transmitting information in such a way that the system in question is caused to behave in one way rather than another.

In "Meaning and Belief" – part III of *Knowledge and the Flow of Information* – Dretske attempts to explain what is required, in terms of the *coding* of information, in order to transform an information processing system into a complex cognitive system capable of manufacturing sophisticated (i.e., higher-order) intentional *structures* out of lower-order informational states. His explanation centres on the transformation of what he refers to as "analog" to "digitalized" information. According to Dretske, certain physical structures in the brain "carry" information in analog form which is related to some object or event in the external world. These analog structures are the result of the incoming information-bearing signals that are created during the early stages of perceptual processing (Dretske, 1981, p. 181).

The defining feature of this so-called analog information is that it is all-inclusive. In other words, analog information includes a theoretically unlimited amount of nomically and/or analytically-related information about the sensory event in question (Dretske, 1981, p. 178). For example, on perceiving a red square, the resulting analog information structure would comprise a vast and comprehensive amount of complexly-nested information related to the square in question such as its size, orientation, location, colour, as well as the fact that it is also a parallelogram, a rectangle, and so on.

All of this information is of no use to the system, however, until it has undergone a certain "restructuring." According to Dretske, in order for the information to be efficacious in the production of some result or action, a certain element(s) within the analog

representation must be "digitalized" to provide the system with access to a structure's "semantic content" (Dretske, 1981, p. 177). Only the digitalized element(s) in an information structure has what Dretske refers to as "semantic content" – i.e., information encoded in digital form. Only information encoded in this way can be used in the eventual production of the intentional states (e.g., beliefs) which are, according to Dretske, causally efficacious in relation to the system's behavior.[17]

Note that the use of the phrase "semantic content" is somewhat misleading here. Although it sometimes sounds like he is claiming that these digitalized elements actually *contain* semantic information, Dretske is usually careful to emphasize that it is the way in which information is encoded by the system that provides the hookup, or connection, between a specific digital element and a particular state of affairs in the world. However, as we will see, no amount of careful wording in relation to how a "semantic" element (associated with a digitalized information structure) can act to represent a particular state of affairs can eliminate the problems inherent in Dretske's account..

Dretske's basic claim in 1981 is that a certain kind of system, S, has the ability to focus on the digitalized element in an information structure, thereby "screening out" all of the other nested components which remain encoded in analog form. The question arises as to just *when* and exactly *how* particular elements in an information structure get selected for digitalization. According to Dretske, it is during the learning process that a given element is selected for digital encoding. It is learning that is responsible for converting

[17] Note that behavior here must be understood as something over and above the physical movements produced by the system in question. Dretske's argument is that it is intentional states which are somehow responsible for the production of behavior. The problem, of course, is how to show the connection between the brain states which cause physical movements and the intentional mental states which allow us to interpret physical movement as a coherent action/behavior.

neural states into specific structures which have the function of representing a particular state of affairs (Dretske, 1995a, p. 261). S, then, develops the ability to digitalize a given element (i.e., to create a semantic structure) through the training and feedback which occur during learning (Dretske, 1981, p. 193). The development of these neurological structures (configurations) takes place by means of repeated exposure to the objects and events of the external world. In order for S to develop the concept of *animal*, for example, it is necessary for S to be exposed to many occurrences of animals and non-animals. Once a conceptual structure with the appropriate "semantic content" is in place, subsequent instances, or tokens, of this established type activate, or trigger, certain behaviors. But how accurate is this process? Although Dretske's model of the development of concepts makes an effort to accommodate false belief, the issue of misrepresentation is one that brings to the fore some disconcerting problems. As Fodor points out in "Semantic, Wisconsin Style," "causal theories have trouble distinguishing the conditions for *representation* from the conditions for *truth*. This trouble is intrinsic; the conditions that causal theories impose on representation are such that, when they're satisfied, *mis*representation cannot, by that very fact, occur" (Fodor, 1984, p. 234).

This is not the only problem with the 1981 account. Dretske's discussion of the role of learning in the establishment of concepts and the processing of beliefs offers an explanation of how mental states might relate to our behavior, but the details of exactly how our brains are able to accomplish this feat seem to be missing. Dretske makes the claim that the setting of some kind of "internal switch" is what determines that one component of an information structure will be digitalized rather than another and that the

selection is based on what S "needs to know" in order to continue to process information in a manner that will result (either now or eventually) in some form of appropriate behavior (Dretske, 1981, p.181). But references to the setting of switches bring up the nasty question of "who" or what is setting the switch (or, indeed, *how* it is set). In other words, who or what is making the judgement that one element rather than another is to be selected for digitalization). It appears that some kind of inner-understander-of-meaning must be involved here.

In describing the way in which his theory of intentionality has evolved over the years (Dretske, 1995a, pp. 259-265), Dretske admits that his 1981 account of mind was weakened by several inherent (and unsolved) problems and acknowledges that his early writings on intentional systems failed to provide a detailed explanation of the way in which mental states could be said to be identified with (but not *reduced* to) the neurological activity of the brain. According to Dretske, this was because in his 1981 account of intentionality, the term "information" was used in such a way that it was describing the nomic dependencies and relationships that existed *between* particular physical and mental events and not with the events themselves. Dretske acknowledges that, without a naturalistic theory of belief, the mental states under discussion in 1981 were faced with the threat of being entirely epiphenomenal (Dretske, 1995a, p. 260). In other words, he was never able to get past the theorizing stage in his description of mind or to provide a truly "naturalized" account of intentionality.

In his subsequent writings, Dretske backs off from the information theoretic terminology he relied on in *Knowledge and the Flow of Information*. In addition, he attempts

to develop a much more detailed explanation of how references to objects and events in the external world are actually incorporated into the development, and subsequent activation, of neural structures in the brain. I will argue in the sections which following, however, that Dretske is never quite able to eradicate the problems that underlie his 1981 investigation into the nature of mind.

4.2 Dretske's Representational Theory of Mind

Dretske claims that the purpose of his 1988 book, *Explaining Behavior: Reasons in a World of Causes*, was to fill in the gaps in his earlier model of the intentional mind by providing a *naturalized* account of intentionality (Dretske, 1995a, p. 263). His first move in this undertaking was to redefine behavior in a way that would facilitate his causal story.

4.2.1 *Behavior: a Process, not a Product*

The standard functionalist account of behavior as output which occurs as a result of the processing of input is flawed, Dretske says. This definition gives the impression that the cause of behavior is one and the same as the cause of output and confuses causal explanations of why we act in a particular way with causal explanations of the body's physical movements. This conflation of behavior and output misleads us into thinking that the cause of behavior is identical to the cause of output. In other words, we begin to confuse psychological explanation of behavior with neurobiological explanations of motor activity so that in the end "our *thinking* this and *wanting* that" are left with no job to do (Dretske, 1988a, p. 36).

Dretske emphasizes the point that behavior is a process and not a product. Behavior, he writes, is neither the internal cause nor the external effect but rather "the one producing the other" (Dretske, 1988a, p. 33). However, this claim requires a detailed account of the relationship between reasons and causes – a convincing explanation of how our beliefs and desires might influence the physical brain states which cause our behavior.

It is in his comprehensive account of different types of representational systems, that Dretske attempts to provide this explanation. In the sections which follow, I will take a critical look at various aspects of Dretske's representational account of intrinsic intentionality.

4.2.2 *The Role of Representation in the Development of Intentional Systems*

Dretske defines a representational system (RS) as a system "whose function it is to indicate how things stand with respect to some other object, condition, or magnitude" (Dretske, 1988a, p. 52). According to Dretske, there are three types of representational systems. Type I, which he refers to as a *conventional* representational system, has no intrinsic representational capability. It relies on intrinsically-intentional systems, such as a ourselves, 1) to select an indicator and, 2) to assign a function to this indicator. Examples include: maps, diagrams, letters (e.g., the letter *a* stands for certain sounds), as well as the use of arbitrary objects such as (say) nickels and dimes to represent some other thing(s) (e.g., the distribution of boys and girls in the room). The important thing to remember about a conventional RS, is that *we* select the indicator and *we* assign it its function or meaning.

A Type II RS is described by Dretske as a system in which natural signs (e.g., bird songs, fingerprints, tree rings, etc.) are assigned the job of indicating something specific. Unlike the arbitrary symbols (e.g., letters and coins) which are used to indicate in a Type I RS, natural signs, according to Dretske, are already capable of indicating that which they will be assigned the job of indicating. In other words, there must be a direct physical/causal relationship between the natural signs selected as indicators in a Type II RS and the information that they will be indicating.

Note that a natural sign can indicate a variety of things about the physical world. For example, a fuel gauge which tells us that we have a full tank of gas can also tell us that there is a large downward force on the bolts which hold the fuel tank to the car's frame (Dretske, 1988a, p. 59). In other words, in assigning the gauge the function of indicating the amount of gas in the tank, we *exploit* the physical/causal relationship between the gauge and the downward force (of the weight of the gas). Our role with respect to a Type II RS is to decide which of the many physical things such an indicator is capable of indicating should be assigned as its *function*. In the case of a Type II RS, the selection of function is based on the specific information needs of the intentional system in question. It is this exploitation of naturally-caused physical events by a genuine intentional system that makes a Type II RS "a curious blend of the conventional and natural" (Dretske, 1988a, p. 54).

4.2.3 *The Intrinsic Intentionality of Type III Representational Systems*

Having concluded that there are, in fact, instances of natural indication to be found in nature,[18] Dretske moves on to his discussion of a Type III RS which he refers to as intrinsic representational systems. Like Type II RSs, Type III RSs possess natural indicator capabilities. The difference between the two, however, is that the former have their representational functions assigned by Type III RSs (such as ourselves), while in the case of the latter, the function of a given representation is determined "naturally" by means of an ongoing learning process. In other words, through repeated exposure to the objects and events of the external world, a natural representational system S is able to "assign" a function to a particular representational process.

The main burden of Dretske's representational theory comes to light right here, at the point at which function is "assigned." In order to establish that a Type III RS is an *intrinsically* intentional system (i.e., that the assigning of a particular function to a particular indicator is a process which takes place in an entirely "natural" manner "intrinsic" to the RS itself), Dretske must provide us with a convincing account of how this might take place *without* reference to any unexplained intelligence (e.g., an homunculus).

In addition to possessing the *natural* ability to indicate how things stand in relation to objects and events in the external world, it is essential that an intentional system have the ability to misrepresent. Only if a system has the ability to make mistakes, "does it have, in its power to get things right, something approximating *meaning*" (Dretske, 1988a, p. 65). According to Dretske, misrepresentation depends on the connection between 1) *what* is

---

[18] See Olsen (forthcoming) for a discussion of why Dretske's argument in favour of natural indication fails (Chapter 4, pp. 70-74).

being represented and, 2) the *way* it is being represented. For example, if I see a furry fast-moving animal running toward me, I (as a Type III RS) may interpret the incoming signals generated by my sensory "machinery" and decide (correctly) that it is my neighbour's dog. On the other hand, I (as a Type III RS) may interpret the exact same incoming signals and decide (incorrectly) that a rabid wolf is running in my direction.

It is here, in the explanation of misrepresentation, that we come face to face again with "the chronic problem" – the same problem that was said to plague Dretske's 1981 account. As soon as there is talk of interpretation of signals, it seems that we refer to some kind of internal interpreting system at work inside the head. In order to get things right (or wrong), it seems like somebody or something must be performing some kind of judgment or comparison based on the information provided by the indicator element in question. Dretske works hard to reassure us that this is not, in fact, the case. He argues that our interpretive capabilities are entirely physical – they have developed because our brains, as a result of the evolutionary process, possess the plasticity required for learning (i.e., for the structuring and restructuring of specific neural configurations) (Dretske, 1988a, p. 104).

This claim sounds intuitive enough but what about the details? How exactly is it that the brain is capable of interpreting neural states/structures/configurations in order to provide us with contentful thoughts? Dretske reminds us that meaning is an abstraction and reassures us that he has no intention of trying to show how something abstract could have causal efficacy in relation to behavior. Meaning *itself* can have no causal role, he writes; rather, it is "the *fact that something has meaning*" (Dretske, 1988a, p. 80) – that a

specific semantic structure is encoded in such a way that it points to/represents a particular event or object – that is the causally relevant fact about that thing. It is only because a semantic structure is encoded in such a way that it points to/represents something *else* that it can be described as "having meaning."

Dretske claims that his worry in relation to his 1981 informational model of the intentional mind was that since he "traced the intentionality of cognitive states ... to the intentionality of information, to the modality inherent in the dependency relations constituting information," it left mental states open to the threat of epiphenomenalism since information itself had no causal or explanatory clout (Dretske, 1995a, p. 263).

The question I want to ask here is: has Dretske, in *Explaining Behavior*, provided the type of details that are needed to overcome the problems inherent in his information-theoretic account of mental states? In this latter account, Dretske takes the ideas he introduced in 1981 (e.g., the selection/recruiting of a specific element in an information structure, the role of learning, the connection between concepts and beliefs, the issue of misrepresentation, etc.) and attempts to provide a detailed explanation of the way in which our intentional states of mind connect to the neural states and/or structures in our brains. The provision of details does not ensure by itself, however, that the details are *correct*. In what follows, I will argue that Dretske's 1988 representational theory is unable to solve the problem of exactly how mental states are active in the causation of behavior.

4.3 The New (1988) Model of the Intentional Mind

To reiterate, Dretske's 1981 account of information processing seemed to leave our intentional states with no causal role in the production of behavior and, for Dretske,[19] this was (and is) an unacceptable conclusion. The challenge in 1988, then, was to provide an explanation that gave these mental states a real job in the production/development of neurological structures in the brain. Having acknowledged the threat of epiphenomenalism in relation to his earlier account of intentionality, Dretske is very careful, in *Explaining Behavior*, to provide a detailed description of the way in which a particular belief (e.g., that it is windy) might be connected to the neurological activity that comprises the first stages of a particular behavioral process.

Meaning, Dretske reminds us – unlike an event or structure – is not a spatio-temporal particular but, rather, an abstract entity and, as such, it cannot *itself* be a cause. His revised claim is that it is not meaning, but rather "a *thing's having meaning*" (Dretske, 1988a, p. 80) that can be said to cause a specific physical effect. The question is, then, what kind of *natural* thing can be described as having (or acquiring) meaning? The answer (for Dretske) is: an internal representation in a Type III RS – a representation that is the result of a Type III RS's making use of its own natural indicator capabilities. This exploitation takes place during the process of learning about, and interacting with, the external world.

In the chapter entitled "The Explanatory Role of Belief," Dretske provides his explanation of how C (an internal indicator) could cause M (some physical result) as a result of C's having the function of indicating F (some external condition). In order to

[19] Note that this is unlike Dennett, for example, who (usually) appears to be quite satisfied with an instrumentalist approach to content.

understand how (or *if*) this account is able to overcome the problem of epiphenomenalism by providing a believable explanation of the cause(s) of behavior), it is helpful to run through a specific example of how C's indicating F might be said to cause M.

Let's say that F (the external condition) is a very windy day and let M (a particular physical movement) equal the placing of your hand on your hat to keep it from flying off your head. Dretske's argument would be (I presume) that it is the fact that the internal indicator, C, carries information to the effect that there are strong currents of air blowing around your head, that *causes* you to raise your hand (M) in an effort to save your hat. The phrase "carries the information" does not mean that C contains the content of your belief (e.g., that it is windy). What C actually does here, says Dretske, is to indicate the fact that it is windy. In addition (and simultaneously?), C acts as an *effector switch* in the production of particular physical movements. In other words, in the neural configuration/structure that results in your hand's going to your head, there is a certain element C which acts to activate the physical movement in question. Both the information to which C refers (e.g., windy weather), as well as some sort of triggering mechanism which C possesses in relation to hand movement are established in the form of a particular neurological structure during the process of learning. Dretske states:

> Learning of the relevant kind is a process in which the dependencies, the correlations defining information, play the role of what I call structuring causes: they help reconstitute effector circuits and thereby contribute, causally, to any future behaviour that depends on these reconstituted circuits. (Dretske, 1995a, p. 264)

This description of how a natural indicator evolves (through learning) into a neural structure with two roles – in the sense that it comprises both a representation, as well as a sort of activation switch in relation to a particular physical movement – is inventive and

complexly-nested information which is nomically or analytically related to the object or event which was perceived. As discussed above, however, until at least one element in an analog structure is encoded in *digital* form, S can do nothing with it. It would be erroneous, therefore, to consider any kind of causal connection between analog information and the behavior/actions (as opposed to the movements) of the information system. All of this Dretske states quite clearly in his 1981 discussion.

Now let's look at what Dretske tells us about natural indicators in *Explaining Behavior*. His claim is that natural signs, or indicators, provide us with information about "a great many things" (Dretske, 1988a, p. 59). As with analog information, however, the "indicator" doesn't really indicate anything until S highlights, or focuses on, a particular element of information and assigns it the function of representing some specific thing. It seems, however, that Dretkse cannot allow S to highlight, focus on, and assign without *presupposing* that which his account is supposed to explain, namely, intentionality.

It can be argued that the "natural indicator capabilities" in a Type III RS are comparable to the "analog representations" in Dretske's information theoretic model – i.e., both are defined in terms of potentiality since, in both cases, information must be *interpreted* before it can be used. In both cases, it appears that some kind of "recruiting process" must take place in order to create a structure having a semantic element. Dretske writes (in 1981):

> The information embodied in this internal (analog) representation can now be digitalized in different ways. Depending on the position of the "internal switch" different semantic structures can be generated. With the switch in one position the system digitalizes the information that *t* is a square. That is, a structure is produced having this as its semantic content. (Dretske, 1981, p. 181)

Dretske maintains that his 1981 theory of mind talks about information only in terms of the nomic relations between particular physical events. However, the passage quoted above – with its references to *embodied information, internal representations, internal switches* and *structures* – does appear to attempt a description of how it would be possible for information about the external world to connect to the physical workings of the brain. But, to reiterate, the act of recruiting, the setting of switches, and the interpretation of information seem to require that which Dretske is in the process of explaining. In other words, his *explanation* of intrinsic intentionality appears to rely on the *assumption* of intrinsic intentionality.

4.4.2 The Role of Learning

According to Dretske, it is during the process of learning that the natural indicator capabilities in a Type III RS are exploited in order to recruit a given natural indicator for the purpose of representing a specific item of information. But in Dretske's 1981 account of mind, he describes, more or less, this same process when he writes: "An internal structure develops (during learning) as a system's way of completely digitalizing information about, say, the *F*-ness of things" (Dretske, 1981, p. 201). This description, although it doesn't refer to neural circuitry per se, does appear to describe the same process – a process in which an internal structure representing the F-ness of things is established through learning.

In addition, both accounts make the claim that the intentional system in question has (due to its particular evolutionary history) evolved in such a way that its neural states have the requisite level of "plasticity" for learning (Dretske, 1981, p. 187 and 1988a, p. 104).

Internal structures (or, in 1988, neural configurations) have the ability to be modified by means of repeated exposure to objects and events in the external world.

4.4.3 Concepts and Beliefs

Dretske's description of how concepts are established and subsequently used as a sort of template in the assessment of beliefs is strikingly similar in the two accounts. In 1981, he describes how a particular internal state evolves "which is selectively sensitive to the information that s is F" (Dretske, 1981, p. 193). Once this structure is established, through the process of learning, it is used to determine whether any subsequent tokens which are "triggered" by sensory events match the established type, or concept.

The question immediately arises, however, as to *when* such a learning process can be said to begin or end. As Fodor points out in his article "Semantics, Wisconsin Style," any attempt to draw a distinct line between before and after in relation to a learning process is bound to run into a serious problem (Fodor, 1984, p. 241) namely, the problem of who, or what, decides when the learning is complete and the concept (e.g., s is F) is ready to be used to decipher subsequent tokens s is F. Unfortunately, the same problems are apparent in Dretske's 1988 description of concept formation.

In *Explaining Behavior*, Dretske describes a representational system as being "selectively sensitive" (Dretske, 1988a, p. 97) to the presence of F. In addition, his explanation of the interaction between structuring and triggering causes appears to be very closely related to his earlier explanation of concepts and beliefs. In fact, Dretske's discussion of the distinction between the structuring and triggering causes of behavior can be overlaid quite easily on his 1981 explanation of the way in which a connection is

established between information structures – which are developed over time through the process of learning – and instances of token structures which are somehow compared to these pre-existing concepts. However, whether he is using terms such as concepts and beliefs, or structuring and triggering causes, the same problem remains. Some kind of unexplained intelligence appears to be involved in the judgement of when a concept, or structuring cause, is robust enough to take on the role of being selectively sensitive to all subsequent instances of F.

4.5 Are the New Details Sufficient to Defeat the Threat of Epiphenomenalism?

To be fair to Dretske, his 1988 representational theory of mind is, in many ways, very impressive.[20] As I argued above, however, it is an account which is based on many of the original ideas which he presented in *Knowledge and the Flow of Information*, and these ideas – irrespective of the terminology used in their presentation – are inherently problematic. I now want to take a closer look at the two problems which were mentioned earlier in connection with Dretske's 1981 model of intentionality in order to determine to what degree they have been eliminated by his 1988 account.

4.5.1 The Problem of Intrinsic Intentionality

I am arguing that the most formidable problem that Dretske's account is faced with is the chronic reappearance of some kind of unexplained intelligence. In 1981 this was the problem of who or what actually makes the judgement about which element in an

[20] Although he argues that Dretske's representational theory is fatally flawed, Olsen describes it as "the most formidable orthodox naturalist position" currently available (forthcoming, p. 67).

information structure is to be digitalized. It is difficult to make sense of a process in which certain elements of analog information are selected for digitalization without referring to some kind of intelligent entity in the system – to something that is able to determine what is relevant in a given analog structure. Dretske, of course, claims that this process does not require any kind of internal intentional entity (e.g., homunculus). Our brains are able to construct these structures "by themselves, in some natural way, either (in the case of the senses) from their selectional history or (in the case of thought) from individual learning" (Dretske, 1995a, p. 261).[21] As noted above, however, the details of how this natural intentional system is able to operate seem to be missing.

In 1988, the same problem – of how a natural representational system is able to harness an F-indicator to the appropriate effector mechanisms – reappears. Dretske's claim is that it is through the reinforcement of a particular behavior which occurs in certain conditions, that internal indicators of these conditions are *recruited* as causes of the output in question. But who or what does the recruiting? Dretske admits that just "*how* they are recruited by this process may be (and to me is) a complete mystery" (Dretske, 1988a, p. 98). In his discussion of recruitment, Dretske appears to fall back on the weak and circular claim that "[s]ince this learning *does* occur, the recruitment *must* take place" (Dretske, 1988a, p. 98). This claim, however, is just as vacuous as the claim that our beliefs *must* cause behavior, because it is possible to describe the latter in terms of the former.

Just as it did in 1981, the homunculus problem reappears in relation to Dretske's 1988 description of the role of the learning process in the formation of representations. For

[21] As was discussed above, however, the issue of learning is itself problematic with respect to when a given concept might be successfully established since we are left with the question as to who or what might be able to determine how and when a concept is to be established.

example, even if we ignore the problem of how internal indicators are recruited, we are faced with a similar problem in trying to understand who or what it is that is able to make the judgment that is required during the comparison of beliefs and concepts or (in 1988) triggering and structuring causes. The occurrence of an "unexplained explainers"[22] in Dretkse's 1981 account acted to undermine his information-theoretic model of mind. I intend to argue that the very same problem invalidates the representationalist account he provides in *Explaining Behavior*.

It can be seen that Dretske's entire representationalist model of mind hinges on the viability of this theory of natural indication (Olsen, forthcoming, p. 83). His account clearly relies on the confusing assumption that "there is something, in nature (not merely in the minds that struggle to comprehend nature), some objective observer-independent fact or set of facts, that forms the basis of one thing's meaning or indicating something about another" (Dretske, 1988a, p. 58).

Although *Explaining Behavior* provides us with a substantially more detailed account of the possible connection between our intentional states and our physical brain states, I maintain that Dretske has been unable to completely eradicate the problematic fact that his explanation relies on some gratuitous form of unexplained intelligence. In addition – and in relation to the homunculus problem – Dretske's account of representation has been criticized (in particular by Jerry Fodor) as being unable to overcome what is referred to as the disjunction problem (Dretske, 1995a, p. 262).

---

[22] This is a term Terence Horgan uses in his 1993 article "From Supervenience to Superdupervenience," to refer to those terms which are used in the explanation a given model of mind which are, themselves, unexplained.

4.5.2 The Disjunction Problem

In his article "Semantics, Wisconsin Style," Jerry Fodor offers his opinion as to why it is so difficult to come up with a robust causal account of intentionality. As he puts it, "causal theories have trouble saying how a symbol could be tokened and still be false" (Fodor, 1984, p. 236) since, in a causal theory, if a representation (R) is caused by (S) – an object or event in the world – then S must obtain (i.e., S must be true). In other words, according to Fodor, "there is, of course, no such thing as mis*information* on Dretske's sort of story" (Fodor, 1984, p. 239). Whatever gets represented must obtain. So for example, take the representation R which covaries with S, your family pet. In the case of a causal theory such as Dretske's, R is said to be caused by S. Suppose, however, that one dark night you go out to look for this pet and because you can't see very well a 'wild' tokening of R occurs – e.g., you mistake an old tricycle (T) for your pet. According to Fodor, if wild tokenings of R are possible, then the nomic dependence of R upon C is imperfect. What R represents is neither S nor T exclusively but rather the disjunction (S v T). R covaries not with the family pet or with the tricycle but rather with either of these conditions (and likely many more as well) (Fodor, 1984, p. 240). "Disjunction is the problem of distinguishing the *misapplication* of a concept . . . from the *correct* application of the disjunctive concept" (Rey, 1995, p. 191). For example, "a representation that covaries with horses and is *misapplied* to cows on a dark night is a representation that could be taken to covary with *horses or cows on dark nights*" (Rey, 1995, p. 191). The disjunction problem appears to wipe out the possibility of misrepresentation, and thereby undermine any causal theory of representation.

Dretske's way of handling this problem is to stipulate that it is only during the learning process that the brain is able to develop concepts (i.e., create the appropriate internal structures) that will subsequently be used to flag token occurrences of R. In other words, it is during learning that the correlations that define what R is to represent are established (Fodor, 1984, p. 241). According to this account, it is necessary to "first get the concept right" before it can be used to determine whether any subsequent tokens provide an appropriate match.

However, as discussed above, Dretske's explanation of the role of the learning process appears to rely on some form of unexplained intelligence. Otherwise, how is a given system able to determine exactly when a concept is fully established? This stipulation – which requires the definition of when a given learning period begins and ends – is, unfortunately, entirely artificial. It requires that Dretske make the assumption that learning takes place according to "a privileged set of 'typical' or 'ideal' circumstances" (Papineau, 1995, p. 226). As David Papineau points out, there doesn't seem to be any non-question-begging way of identifying such ideal circumstances other than as those "where people form *true* beliefs" (Papineau, 1995, p. 226).

In order to see how difficult it would be to ascertain when and how a concept might get established, let's look a little more closely at Dretske's explanation of how an internal indicator, C, becomes a representation of F in virtue of the control duties it takes on in relation to M (some physical movement). As discussed above, Dretske describes learning as that process in which the correlations defining information play the role of structuring causes. In other words, "they help reconstitute effector circuits and thereby contribute,

causally, to any future behavior that depends on these reconstituted circuits" (Dretske, 1995a, p. 264). But as has been frequently pointed out, it is not at all clear from Dretske's account how it is possible to determine the actual sequence of events required in order to establish a legitimate concept against which subsequent belief tokens can be compared.

Take, for example, the case in which I hold on to my hat on a windy day. According to Dretske's account, the first time I hold down my hat with my hand, certain circuits are constituted in such a way that an internal indicator, C, is established which both flags the fact that "there is wind," and activates the process required to generate the hat-saving behavior. But before my actions can be said to be causally connected to my belief (that it is windy) and my desire (to keep my hat on my head), a structure, or belief, which has been developed during a specific learning period must be in place. This seems to imply that I perform my hat-holding actions "mindlessly" for the first few times that the wind blows. Exactly how many exposures to wind and blown-away hats do I need before I have finished learning and have a "legitimate" belief to work with? Or in Dretske's terms, how many exposures to winds that will (and winds that won't) blow my hat off are required before C and its power to activate M can be said to be a structuring cause? The fact that questions such as these that can't be answered leads to the conclusion that Dretske's discussion of the connection between structuring and triggering causes must be somewhat off the mark.

This confusion – about the relationship between structuring and triggering causes – is even more serious than it seems at first. Dretske's claim that you can't have intentional states prior to the establishment of learned concepts leaves open the possibility of a class

of humanoid beings who are entirely lacking in intentionality. For example, given his model of learning, Dretske would have to claim that if a physical duplicate of you should suddenly (and miraculously) materialize, this biological twin's physical movements could not be referred to as behavior (or actions) since these movements would have no meaning because they would not have been triggered by any intentional state(s). For example, if you deliberately raise your arm to frighten away a pesky fly, your action can be said to have a purpose – a purpose, or meaning, which has been established over time through learning. If your bio-double raises his or her arm, however, he or she cannot be said to be shooing away a fly, but rather just moving his or her arm. Dennett, in *Dretske and his Critics*, points to (what he sees as) the absurdity of this notion when he asks: How long, one wonders, should "acquiring the requisite extrinsic relations take? . . . How many flies must buzz around the head of a bio-double before he can start shooing them" (Dennett, 1991c, p. 125)?

This thought experiment is good for more than just another of Dennett's quips, however. It points to the serious gap that occurs in Dretske's description of the establishment of, and interaction between, structuring and triggering causes. We know that Dretske defines learning as the establishment of particular brain structures that result from the system's interaction with the external world. It is just not clear, however, how much (and what kind of) interaction is required before a reliable concept is actually in place and ready to provide an accurate type for subsequent belief tokens. But if Dretske fails to explain the learning process adequately, then the disjunction problem remains unsolved since, without a viable theory of learning, Dretske has lost his way of keeping

this problem at bay: a theory of mind simply cannot get along without a consistent account of how false belief (and other "misrepresentational") states can occur.

In "Semantics, Wisconsin Style," Fodor also makes reference to the first problem I discussed above (i.e., the homunculus problem). He claims that even if it were possible to determine when learning begins and ends in relation to a particular concept, Dretske's account is seriously flawed. This is because the judgements that seem to be required in the establishment of concepts take us right back to the original, and most serious, problem that underlies both of Dretske's accounts – the problem of who, or what, is at work defining, and making judgements in relation to, the structuring and triggering causes of behavior.[23]

These two problems – the problem of intrinsic intentionality and the related disjunction problem – were originally articulated in relation to Dretske's 1981 information theoretic model of mind. However, I am arguing that neither problem has been successfully resolved in Dretske's 1988 account of representational systems. In spite of using a more acceptable terminology in the provision of a much more detailed account of the relationship between reasons and causes,[24] Dretske's 1988 representational model of mind fails to resolve the basic problems inherent in his earlier explanation of intentional systems. Such a model cannot, therefore, be expected to provide a solid base for the development of a representational theory of consciousness.

---

[23] Fodor refers to this unexplained intelligence in Dretske's account of learning as "the Teacher's pedagogical intentions" (Fodor, 1984, p. 242).

[24] For criticisms of Dretske's account of reasons and causes in his "component" view of action, see Olsen (forthcoming), section 8.2.

4.6 Representation and Consciousness

After *Explaining Behavior*, Dretske went on to provide a representational account of sense experience, a project he describes as being a "tougher nut" (Dretske, 1995a, p. 263) than his investigation into the causal efficacy of belief. Dretske's plan was to establish a robust explanation of representational systems that would act as a base for his explanation of conscious experience. Even back in his early information theoretic account, Dretske sometimes made reference to two kinds of representations: those that are established by means of evolutionary design, and those that are established through the learning process. The latter are those representations involved in the causation of behavior, while the former relate to the sensory experiences which act as fodder in the grist of the intentional mill.

As Olsen points out, any natural theory of representation must entail an explanation of the role of consciousness in the establishment of internal representations since it appears that "meaning is established *only* through the 'meaning-giving' function of consciousness" (Olsen, forthcoming, p. 67). In fact, even in 1981, there are frequent references to the role that consciousness must play in the processing of information. For example, there seem to be significant similarities between Dretske's description of digitalized information and Ned Block's definition of access consciousness. According to Block, a state is access conscious "if its content is . . . freely available as a premise in reasoning; and if its content is available for the rational control of action and speech" (Block, 1993, p. 182). From Dretske's description of analog and digital representations, we can see that the information in the former is inaccessible since we have no conscious access to it until the system filters out most of it and focuses on only a specific element. This distinction – between straight sensory input and the conceptual processing of these incoming signals – has been the basis

sensory input and the conceptual processing of these incoming signals – has been the basis of Dretske's very earliest work. *Seeing and Knowing* (1969), for example, focussed on the difference between seeing and believing (i.e., the perception of some object or event and the conceptual processing of this perception).

It is not surprising, therefore, that in *Naturalizing the Mind* (1995b), Dretske approaches the issue of conscious experience by using two different kinds of representations: systemic representations$_s$, whose indicator functions are "built-in" by means of the evolutionary process, and representations$_a$, whose indicator functions are "acquired" (hence, the 'a' subscript) during the learning process. The latter, according to Dretske, act to mediate, or make sense of, the vast number of systemic representations that result from the processing of sensory input.

When I began this chapter, it was my intention to discuss Dretske's representational theory of consciousness in terms of the problems that arise in relation to the connection between experience and thought – between our sensory experiences and the conceptual framework which processes them. In investigating Dretske's representational account of intentionality, however, it became apparent that (as with Van Gulick and Dennett) the problems inherent in his theory would act to invalidate any representationalist account of consciousness. As Seager points out, a representational theory of consciousness is "hostage to the fortunes of its underlying general theory of representation" (Seager, 1997, pp. 93-4). As I have argued above, Dretske's theory of representation appears to depend on some kind of unexplained intelligence in its explanation of natural intentional systems. Such an

This conclusion (in relation to Dretske's work) is, in fact, the conclusion of my thesis overall. I maintain that the problems that arise in the three representational accounts of consciousness which I have discussed in previous chapters, can always be traced back to problems inherent in the original theory of representation on which they are based. In the final chapter, I want to take a look at why it is so difficult to come up with an credible naturalistic theory of how our mental states hook up to the neurophysiological workings of our brain.

Chapter 5

The Instability of Nonreductive Materialism

(And What to Do About It)

In the preceding chapters, I have looked at three contemporary philosophers who, having given somewhat different functionalist accounts of the connection between intentional states of mind and physical brain states, have then gone on to attempt an explanation of consciousness based on their respective representational models. In examining each author's evolving account of intentionality, I made the claim that, in each case, the explanation of consciousness could be traced back to their original ideas with respect to how mental states such as beliefs and desires relate to brain states. The rationale in all three cases seems to have been – as Daniel Dennett puts it – "first a theory of content or intentionality – a phenomenon more fundamental than consciousness – and then, building on that foundation, a theory of consciousness" (Dennett, 1998, p. 355).

I maintain, however, that Van Gulick, Dennett, and Dretske have jumped the gun when it comes to claiming to have provided any sort of robust explanation of mental states since, in each case, their functionalist models of intentionality are weakened by an instability which forces a return to the problems of earlier models of mind such as eliminative behaviorism and/or identity theory. My conclusion is, therefore, that none of the three accounts was viable in the sense of being able to provide a solid base for an explanation of consciousness.

As discussed earlier, the problem of consciousness is often seen as being more difficult than the issue of intentionality which is widely thought to be explainable as some kind of representational system operating within the context of one of a variety of functionalist accounts of mind.[1] Nevertheless, and admitting a certain degree of extrapolation, my argument is that there is no functionalist/representationalist account of mind problem-free enough to warrant the claim that it thoroughly explains what lies behind the intentional nature of our mind. In fact, in progressing through the works cited for this thesis, I began to form the opinion that it is not just the problem of *consciousness* (i.e., as some free-standing and separate mental entity) that cries out for explanation, but rather the problem of *content* (i.e., the way in which meaning relates to the physical world) that continues to provide us with a very difficult (and still very much unresolved) problem – the mind/body problem.[2]

Dretske, Van Gulick, Dennett, and others propose that the notion of consciousness can best be understood in the context of the connection between consciousness and the intentional nature of our thoughts. I agree with all three of them in this respect. It appears very likely that a discussion of one "problem" entails an examination of the other. Hilary Putnam writes:

> . . . how plausible is it that one should be able to reduce (hypothetical) "laws" involving the notion of *consciousness* without becoming involved in "reducing" the propositional attitudes? The concept of consciousness (certainly the concept of consciousness that is relevant for *epistemology*) is the concept of *availability to thought*.

---

[1] For example, Block in "Troubles with Functionalism," maintains that functionalism can handle intentionality and only runs into problems when it comes to dealing with the qualitative aspect of our experiences (Block, 1978). Dretske, in his entry in Guttenplan's *A Companion to the Philosophy of Mind*, refers to the explanation of experience (as opposed to beliefs, for example) as "a tougher nut" (Dretske, 1995a, p. 263).

[2] My point here is simply that when it comes to solving the "problem" of how mind relates to body, the issue of intentionality provides just as nasty a roadblock as does the issue of phenomenal consciousness.

> Once again, it seems that either we do not know what theory it is that we are speaking of "reducing," or else the theory includes a substantive portion of our talk of propositional attitudes. (Putnam, 1994, p. 481)

In spite of my agreement with the view that there is a strong connection between intentionality and our conscious experience of the world, however, my claim in this thesis is that none of the three intentionalist models of mind I have examined is up to the task of explaining the nature of this connection. In the sections which follow, I want to examine what I have referred to as the *instability* inherent in the functionalist model of mind in relation to a discussion of the general viability of nonreductive materialism.

5.1 The Instability of Nonreductive Materialism

In "From Supervenience to Superdupervenience," Terence Horgan outlines the difference between those philosophers of mind who believe that nonreductive materialism *can* provide a viable explanation of mind (e.g., Davidson, Fodor, Van Gulick, Horgan, etc.), and those (e.g., Churchland and Kim) who do not. As outlined in chapter 1, reductive materialism, or identity theory, came about as a reaction to behaviorism which was seen to be problematic because it required the denial of mind and mental states. Reductive materialism provided a way back to the mind by means of type-type reduction which claimed that a given mental event was entirely reducible to a specific physical (brain) event. This version of physicalism, however, ran into serious epistemological problems when it became increasingly obvious that the chances of coming up with an explanation that showed strict identity between mental events and brain states was highly unlikely (and likely impossible).

Nonreductive materialism was the reaction (and, it was hoped, the solution) to these difficulties. Its goal was to provide a materialistically-sound explanation of the connection between mental processes and brain processes without making the claim that the former can be (strictly) reduced to the latter.[3] Currently, the most widely accepted nonreductive materialist approach to mind is functionalism. Functionalism, as discussed above, attempts to define higher-order mental properties in such a way that the connection between these properties and the physical events on which they are said supervene can be explained in physical terms (Horgan, 1993, p. 579). Just how this explanation should play out, however, has created a great deal of discussion and dispute. In "Supervenience and Superdupervenience," Horgan makes the claim that there are good reasons for being skeptical about the viability of the functionalist model of mind (Horgan, 1993, p. 579). Following Horgan's lead, one of the central claims of this thesis has been that functionalism has failed to eliminate the chronic problems inherent in previous explanations of mind – the very problems, in fact, that it was supposed to resolve. In the sections which follow, I want to take another look at these chronic problems in relation to each of the three models of mind (Van Gulick's, Dennett's, and Dretske's) discussed in the previous chapters.

[3] In "Nonreductive Materialism and Mental Causation," Ausonio Marras describes nonreductive materialism as the thesis that: "psychology is not reducible to physical theory in the classic sense of 'reduction,' according to which a we reduce a theory to another theory by deriving the laws of the former from the laws of the latter via 'bridge principles' ..." (Marras, 1994, p. 465).

5.2 Van Gulick's Gap: How One Problem Leads to Another

Van Gulick's representationalist account of mind provides a good example of the troubles that functionalism runs into with respect to liberalism versus chauvinism as outlined in Block's 1978 article "Troubles with Functionalism." As discussed in chapter 2, Van Gulick – in order to avoid becoming too liberal when designating a system as one which can be described in functionalist terms – proposes the use of a teleological restraint. He claims that only systems which behave in an adaptive manner with respect to their specific environment can be described using functionalist terminology. These systems, therefore, must be seen to be performing the functions that they were designed (by an evolutionary process) to perform. However, as Block points out, the solution to the problem of liberalism generally brings to the fore the problem of human chauvinism. In stipulating the teleological design restraint, it is likely that the functionalist account will become overly chauvinistic since systems which are not designed through a similar evolutionary process cannot be described in functionalist terms.

The uneasy balance between liberalism and chauvinism provides a first glimpse of the instability I am claiming underlies the functionalist approach. Van Gulick can be seen as opting for chauvinism (as the lesser of two evils), thereby resolving the issue and eliminating the problem of instability. However, in the process of providing the details of his "chauvinistic" functionalist account, Van Gulick runs into another unstable situation. His homuncular functionalist model is described in terms of a set of hierarchically-nested (and progressively less intelligent) homunculi which are used to explain how abstract intentional entities such as beliefs might relate to our physical brain processes. The *decompositional* strategy of homunucular functionalism provides (according to Van Gulick

and others) an explanation of how it is possible for physical systems to possess intentional capabilities since it is said to eliminate the problem of having unexplained instances of intelligence. There is, however, a high price to pay for this strategy.

In describing the burden of the functionalist philosopher, Horgan claims that the goal must be to give a "tractable specification" in non-intentional and non-mental vocabulary of the sufficient and necessary conditions for the instantiation of mental properties (Horgan, 1993, p. 579). Those philosophers who promote homuncular functionalism (e.g., Dennett, Van Gulick, Lycan, etc.) maintain that it does just that. However, as discussed in chapter 2, this solution to the problem of unexplained intentionality is drawn back (whether unwillingly or not) in the direction of a reductive-type account. For example, in Van Gulick's discussion of hierarchical systems, mental states are said to *decompose* eventually to the point where the activities of the very lowest-level (dumbest) homunculi can be described *in purely physical and/or hardware terms.* In spite of the claim that homuncular functionalism is merely an abstract model of mind, the terms used by Van Gulick (and others) appear to be based on a belief in the *possibility* of the reduction of abstract to physical, at least in the sense of a theoretical identification in which a notion in one science is "reduced" to a notion in a different science.[4]

Functionalists maintain that they are not promoting reduction since their theory of mind is simply that – a theoretical model. My claim, however, is that functionalism (and, in particular, homuncular functionalism) can't help but be pulled in a direction in which

---

[4] The example of the notion of light being reduced to the notion of electromagnetic radiation is frequently used to provide an example of this kind of theoretical identification. Putnam claims that any claim of theoretical identification "stands or falls with the possibility of showing that the approximate truth of the laws of the former science can be *derived* from the laws of the later science (the more "basic" one) with the aid of the proposed system of theoretical identifications . . ." (Putnam, 1994, p. 479).

a more concrete explanation is required of how the three levels of mind – physical, functional and intentional – interact. Otherwise, the functionalist account of intentional mental states remains just as powerless when it comes to defeating Cartesian worries (e.g., unexplained instances of intelligence) as the physicalist version of mind whose problems it was designed to remedy.

In "From Supervenience to Superdupervenience," Horgan describes functionalism as one of the "recent so-called 'naturalizing' projects, in philosophy of mind ..." (Horgan, 1993, p. 579). His claim in this article is that any such naturalizing project cannot help but be reductive in a certain sense. Even if it eschews type-type reduction, it would seem that the a functionalist model must provide an explanation of inter-level connection which is robust enough to discount all counterexamples. But as can be seen by taking a look at the three functionalist models of mind I have discussed above, this is not the case. None of the three accounts is problem-free and there is no indication that one set of problems (Van Gulick's or Dretske's for example) is any more desirable than the others.

Van Gulick has written several interesting articles dealing with this very criticism – i.e., that the functionalist account of mental states, because it is entirely abstract and theoretical in nature, can never provide any sort of robust model of how mind and brain fit together. In "Who's in Charge Here?" he argues against the criticism that functionalism – since it fails to provide any sort of verifiable account of the causal efficacy of mental states – is unable to eliminate the threat of epiphenomenalism. Van Gulick maintains that when it comes to explaining the causal connection between intentional states and behavior, we are setting ourselves up for failure by setting our standards of explanation too high. He points out that, in fact, "none of the properties of the special sciences are

causally potent" (Van Gulick, 1995a, p. 249). Why, then, should psychophysical properties be expected to "meet a higher standard for causal potency than biochemical, geological, or optical properties?" (Van Gulick, 1995a, p. 249). Although the events and objects picked out by these special sciences (e.g., geology) are entirely composed of physical parts, says Van Gulick, the causal powers of these events and objects are not determined *solely* by the physical properties they possess and the laws of physics, but also according to the *organization* of the physical parts within a whole – an organization which is defined, or *interpreted*, by the predicates of the special science in question. In other words, it can be said that physical events are determined by the laws of physics together with *"initial boundary conditions"* (Van Gulick, 1995a, p. 250).

In taking this approach, Van Gulick demonstrates a strong allegiance to (and belief in the viability of) nonreductive materialism.[5] He suggests that it is necessary to give up the notion that physical/causal explanations are somehow more valid than special science explanations. His intention here is to remove the aura of "special status" associated with physical properties and provide mental properties with some kind of "different but equal" status for their role in the causation of behaviour. He writes:

> We have two models of the world which cannot be reduced in the sense that there are no well-ordered complete translation functions from one to the other – a gap which results in part because of the ways their respective concepts are anchored in our specific discriminative and cognitive capacities. (Van Gulick, 1995a, p. 255)

[5] Likewise, in the article "What Would Count as Explaining Consciousness?," Van Gulick makes the claim that the standards that are set in terms of what might constitute a valid explanation of the conscious mind are simply too high. In determining what is required to explain consciousness, Van Gulick says, there is no need to worry about the issue of logical and/or nomic sufficiency. He suggests that a predictive model is perfectly adequate when it comes to explaining the qualitative aspect of our experiences (Van Gulick, 1995b, p. 72). In his approach to the problem of consciousness, it is possible to see (once again) Van Gulick's allegiance to a very flexible form of nonreductive materialism.

This rationalization of the gap between the physical and the mental is not new, of course. It can be seen to fit in right along side of (for example) Dennett's advice with respect to taking an intentional stance in the interpretation of intentional systems. Van Gulick, however, isn't plagued by the same eliminativist tendencies as Dennett and so his position on nonreductive materialism is certainly more stable than Dennett's. However, if we go along with Van Gulick in "loosening up" on the requirements for the explanation of the causal role of mental states, we face the risk falling back into the acceptance of some form of property dualism and the Cartesian worries that this entails. Why is this so? According to Kim, any functionalist account that is unable to show a clear physical, or causal, connection between mental states and physical behavior leaves the former threatened by epiphenomenalism since by the principle of explanatory exclusion there cannot be two or more independent explanations of a single physical event (Baker, 1995, p. 490). If a functionalist account cannot provide a clear explanation of how mental properties are causally efficacious in the production of physical behavior, then these mental properties must be epiphenomenal or else the account in question appears to countenance the dualistic notion that our behavior has two separate causes!

Van Gulick's homuncular functionalism, then, appears to be faced with the prospect of maintaining an unsteady balance between reduction and property dualism. It was, however, the concerns associated with these very positions (i.e., identity theory and Cartesianism) that functionalism was supposed to eliminate!

5.3 A Dennettian Motto: If you can't solve the problem, eliminate it.

In many respects, Daniel Dennett's and Robert Van Gulick's models of the intentional mind are very similar: both use an homuncular account of functionalism with an emphasis on a teleological restraint. Dennett, however, is not willing to condone any sort of approach to mind that allows for a non-physical mental property to take part in any causal story. For this reason, he is always quick to describe such entities as selves, beliefs, and so on as (in a strong sense) illusory.

Of the three functionalist models of mind described in the preceding chapters, Dennett's is the most difficult to classify according to a commitment to reductive versus nonreductive materialism. As discussed above, Dennett himself sometimes appears to be uncertain about which stance (design or intentional) to take in the description of a given system. I argued above that in the end it is Dennett's eliminativist tendencies which define his approach to mind and that these tendencies can be traced back to his early description of mental states in which he makes the point that in ascribing "content" to intentional states, we are simply using a heuristic device in order to come up with an interpretation of the behaviour of the system in question. According to Dennett, our beliefs and desires have no ontological reality. Rather, talk about beliefs and desires can be used to help in the interpretation of the behaviour of systems which can be described in functionalist terms. I maintain that Dennett has remained true to the claim that "mind" is not part of the physical universe from 1969 to the present and that it is his strong commitment to this notion that requires an eliminativist approach to conscious mental states.

What can seen to be present in Dennett's account (and not in Van Gulick's) is a certain "angst" with respect to what degree of realism to hold in relation to the beliefs and

desires of folk psychology. Is it necessary to move in the direction of reductive materialism in order to explain the connection between the mental and the physical, or is an instrumentalist stance with respect intentional systems the only (or most practical approach) to take? Of the three philosophers I have discussed in this thesis, Dennett appears to be the most uncertain about the answer to this question.

As discussed above, Van Gulick has reconciled himself to the fact that the functionalist account he offers is simply a theoretical model which can be used to understand the nature of our mental states. He maintains that there is no reason to set the standards with respect to the explanation of mind so high that we are forced to admit defeat in the face of insoluble mystery. Van Gulick, then, takes Dennett's advice with respect to the intentional stance and is content with the notion that beliefs and desires cannot (and don't need to be) reduced. Dennett, however, doesn't seem able to stick with his own advice. The direction he takes in *Consciousness Explained* shows that he is not satisfied with an intentional interpretation of our mental states but must push on to what he refers to as an empirical theory of mind.

I concluded above that Van Gulick's representational model of mind fails to provide any sort of robust account of how abstract entities such as mental states relate to the physical structure that is said to underlie them and that, therefore – according to Kim's principle of explanatory exclusion (Baker, 1995, p. 490) – his account is threatened by Cartesian and/or epiphenomenal worries. Van Gulick's reply would likely be that if his account is not entirely robust, it is robust enough. It can be of some use in helping us to understand how mind and brain relate. Dennett, however, is not that easy to please. From the beginning, his allegiance to a fairly strong form of verificationism has kept him

committed to eliminativism with respect to mental entities such as beliefs and subjective experiences.[6]

In "From Supervenience to Superdupervenience," Horgan states that those who cannot accept nonreductive materialism as a viable metaphysical position, generally head (back) in one of two directions: reductive materialism or eliminative materialism (Horgan, 1993, p. 575). In Dennett's case it looks like he can't quite decide which of these two directions to take. He is eager to provide a reductive account of consciousness and yet, at the same time, he appears to acknowledge that a serious explanatory gap exists when it comes to explaining the connection between conscious mental states and the functionalist design he proposes. His solution, therefore, is to eliminate (i.e., treat as illusory) problematic entities such as beliefs and conscious selves .

Dennett, then, can be described as a paragon of instability and, as such, he provides a good example of the balancing act that is required in order to keep the functionalist account of mind afloat. It is important to keep in mind that Dennett's intentional stance was designed to accommodate his commitment to an eliminativist/behaviorist approach to mental states. In the end, however, the taking of a "stance stance"[7] with respect to the explanation of mind – since it entails the acknowledgement that there are multiple ways of interpreting the same system – comes too close (for Dennett's comfort) to the admission that there are two kinds of properties, physical and mental. His solution to this discomfort is to reconfirm his allegiance to an eliminativist description of mind and consciousness.

[6] Putnam describes Dennett's eliminativism as another instance of the phenomenon of "recoil" in philosophy when he writes; "It is, I suspect, just because consciousness and reference cannot be identified with a definite brain function (respectively, a definite physical relation) that Dennett is led to the denial of both subjective consciousness and objective reference" (Putnam, 1994, p. 476-7).

[7] See Dretske's (1988b) article detailing Dennett's attitude with respect to stances entitled "The stance stance."

These vacillations – between the elimination, and explanation, of mental states – demonstrate clearly the type of instability that I am claiming underlies the functionalist account.

5.4 Dretske's Alternative to Eliminativism: "Natural" Representation

Dretske's representational account of mind and consciousness refuses to give in to defeatist worries when it comes to explaining how mental states take a role in the causation of behavior. Dretske, as opposed to Dennett, is a full-fledged realist with respect to beliefs and, therefore, is committed to providing a detailed account of "how beliefs and desires – in virtue of their representational content, *not* their neural-physical properties – can cause, and causally explain, behavior and action" (Kim, 1991, pp. 52-3). In chapter 4, however, I argued that Dretske's representational account of intentionality runs into the same (or, at least, very similar) problems that led one functionalist (Van Gulick) to fall back on a very flexible version of nonreductive materialism and another (Dennett) to opt for an eliminativist approach to beliefs and qualitative experiences.

In "How Reasons Explain Behavior," Jaegwon Kim provides a critique of Dretske's account of intentionality which fits into the argument that I have been making in relation to the functionalist account of mind (i.e., that it is plagued by instability). Kim's discussion deals with "broadly metaphysical issues" (Kim, 1991, p. 53) and these are just the sort of issues (as opposed to the specific details of Dretske's account) which I want to look at here.

As discussed above, Kim maintains that the main roadblock when it comes to explaining the causal role of the intentional properties of mental states is related to the problem of causal-explanatory exclusion (Kim, 1991, p. 57). To put it in very simple terms,

this is the problem of showing how non-physical properties are able take part in a physical causal process. If every physical event has a physical cause, then any mental properties associated with this event appear to be entirely epiphenomenal since to claim that a given event has *two* causes is to succumb to a form of *overdetermination*, or dualism. It is Kim's claim that it is this threat of overdetermination that lies behind the instability of the functionalist account. When faced with this threat, the functionalist appears to have two choices. He can eliminate the problem by retreating to some form of eliminativist behaviorism (in which mental entities such as beliefs, etc. are declared to be illusory). Alternatively, the functionalist is forced to loosen up on the formal definition of physicalism – what Kim refers to as "the rejection of causal-explanatory closure of the physical domain" (Kim, 1991, p. 56). As we saw with Van Gulick, however, this latter choice appears to head back in a direction which condones some form of Cartesian description of the connection between mind and body, and this, unfortunately, is what the functionalist is motivated to avoid at all costs.

In "Dretske on How Reasons Explain Behavior," Kim first refers to Dretske's method of getting around the problem of overdetermination as his "dual explanadum strategy" (Kim, 1991, p. 58). In other words, he claims that Dretske's account requires two explanations – one for what causes behavior, and one for what causes bodily movement. This presents a problem and, according to Kim, "*a successful execution of the strategy requires commitment to dualism*" (Kim, 1991, p. 59). But, Kim says, being a committed physicalist, Dretske can't really mean to support this strategy. Kim goes on to interpret Dretske's representationalist account of intentionality as rather a "*reductive account of content*" (Kim,

1991, p. 61). Kim, of course, has an agenda here and this is to claim that Dretske's account can be improved by relying on a truly reductive (i.e., strongly supervenient) approach.

Is it a question, then, of Dretske's just not giving *enough* of what is required (e.g., reduction) or is his approach basically unstable, as I am claiming all functionalist accounts of intentionality must be? Not surprisingly, I opt for the latter interpretation. Dretske, as a realist, is highly motivated to fight the threat of ephiphenomenalism with a naturalist account of mind but, in doing so, he seems to be (once again) hanging somewhere between reduction and Cartesianism. His account of the causal efficacy of intentional states takes him in the direction of reduction since it is based on talk of the reconstituting of effector switches by structuring causes, and so on. At the same time, his representational account appears to rely on some unexplained intelligence in order to make the claims that it does with respect to natural representational systems. The fact that a discussion of Dretske's representational model contains references to both reductionism and Cartesianism is a good indication that his account of mind (just like Van Gulick's and Dretske's) is basically unstable.

5.5 Looking at Nonreductive Materialism from a Different Angle

I have made the claim that the functionalist approach to mind is plagued by a certain instability which invariably results in it's falling back on some earlier explanation of mind. In the "Dewey Lectures,"[8] Hilary Putnam discusses the issue of the instability of certain philosophical positions from a broader perspective. He argues that philosophy since the

[8] The full title of these published lectures is "Sense, nonsense, and the senses: An inquiry into the powers of the human mind."

17th century seems to have kept itself in a chronic state of "recoil" (Putnam, 1994, p. 446) in which it has "oscillated between equally unworkable realisms and idealisms (Putnam, 1994, p. 488). This suggests that the currently popular model of mind is always motivated by, and therefore defined in terms of, a reaction to a previous (and now debunked) position. For example, and as described earlier, in reaction to the dualist threat, philosophers either eliminated all references to the mind with behaviourism or attempted to reduce mental states to physical states by means of the identity theory. Problems with these two approaches, however, led philosophers back in the direction of nonreductive materialism (what Putnam refers to as Cartesianism *cum* materialism). Nonreductive materialism, however, reopens the issue of Cartesian division which requires extreme measures such as eliminativism, and so on.

In the "Dewey Lectures," Putnam urges that we break this cycle of recoil by "examining the central metaphysical issue of realism" (Putnam, 1994, p. 446) and the motivation that lies behind the inevitable return to a realist account of mind. He makes the claim that when it comes to realism, taking an extreme position in either direction (i.e., the claim that objective reality exists in a state of absolute independence from our perception and experience of it, versus the claim that the world as we know it is nothing more than a product of our own mind) is not productive. Instead, Putnam encourages those involved in philosophy to work on redefining what the important metaphysical and epistemological issues are before coming up with solutions to problems that may, or may not, exist.

What, if anything, has Putnam's discussion of realism versus subjectivism got to do with my claim that the functionalist account of mind is, by definition, unstable? In the 1960s, Putnam attempted to identify propositional attitudes with computational states but

by the mid-1980s, he became convinced that the type of reduction he was proposing was not feasible (Putnam, 1994, p. 480). He recounts how his early formulations of internal realism were an unsatisfactory attempt to resolve the apparent antimony involved in the relating of mind and world.[9] Putnam describes how his idea of "functionalism" employed the notion of theoretical identification, or reduction, of "intentional talk to physical *cum* computational terms" (Putnam, 1994, p. 479). He acknowledges that it wasn't too long before he realized that the formal properties of computational states were entirely unlike the formal properties of psychological states and that, therefore, reduction (in the form of theoretical identification) was not a possibility.

In discussing the project of reducing intentional talk to physical/computational terms, Putnam is forced to acknowledge what he sees now as his own past mistakes. He warns that the attempt to define mind in functional terms is not "the straightforward scientific project it might seem at first blush to be, but a chimera" (Putnam, 1994, p. 479). No one, says Putnam, has any idea what a truly scientific psychological theory – one whose terms could be reduced to the terms of computer functionalism – might look like. He writes:

> One hears a lot of talk about 'cognitive science' nowadays, but one needs to distinguish between the putting forward of a scientific theory, or the flourishing of a scientific discipline with well-defined questions, from the proffering of promissory notes for possible theories that one does not even in principle know how to redeem. If I am right, the idea of a theoretical reduction in this case – the reduction of the entire body of psychology implicit in our ordinary practices of mental-state attribution to physics *cum* computer science – is without any clear content. One cannot make precise the unexplained notion of "identity" of "sense data" with "functionally characterizes [sic] states of the brain" with the aid of the concept of the reduction of one theory to another if one has no idea of the nature of the theory *to*

[9] He writes: "I still thought of the mind as a thing, and, hence, saw no recourse but just to identify it with the brain" (Putnam, 1994, p. 461).

> *which we are supposed to do the reducing* (and only a very problematic idea of *what theory we are supposed to reduce*). (Putnam, 1994, p. 480-1)

In discussing what he sees as the inadequacy of the functionalist account of mind, Putnam is not suggesting a retreat to identity theory, behaviorism, or dualism. On the contrary, his point is that must stop the pattern of recoil in relation to the most recent (problematic) theory of mind and look at the way in which the mind relates to the world in an entirely new way. Putnam's new way is what he refers to as *natural realism* which is based on the claim that our cognitive processes are in direct contact with the world at large. According to natural realism, we interact with the world directly – not indirectly by means of some sort of *interface* (e.g., functional system) which represents it.[10] Putnam blames the misconceptions of functionalism on what he refers to as "the mathematization of nature" (Putnam, 1994, p. 468)[11] His claim is that since the 17th century, it has come to be increasingly accepted that natural entities, and the connections between them, can only be described according to mathematical laws which are expressible in terms of algebra and calculus. He maintains that the push to describe nature in this manner erroneously implies that the "everyday descriptions" of the world we live in must be false (Putnam, 1994, p. 468).

It would be easy to see Putnam's reaction to functionalism as yet another example of recoil in the sense that natural realism appears to have sprung from the ashes of the unsatisfactory functionalist account of mind. Putnam, however, claims that his new way

---

[10] In this respect, Putnam's interpretation is very similar to that of McDowell who claims that the main problem with current materialist models of mind is the erroneous idea of an interface (representational or otherwise) which is said to exist between mind and world. See McDowell's *Mind and World* (1994).

[11] The reference he gives for this phrase is Husserl's *The Crisis of the European Sciences and Transcendental Phenomenology* (1970).

of looking at the mind is not intended to provide (yet another) alternative metaphysical account. Rather, it involves "seeing the *needlessness* and the *unintelligibility* of a picture that imposes an interface between ourselves and the world" (Putnam, 1994, p. 487).

Kim makes the claim that anyone who takes a realist attitude with respect to mental states is under a certain obligation to provide an account of how mental causation is possible (Kim, 1991, p. 52). But Kim is a philosopher for whom there is only one way to understand the connection between body and mind and this way entails a commitment to strict physical reduction and strong supervenience.

What is interesting is that Kim and Putnam (and Horgan and Dennett and Van Gulick, for that matter) are in agreement up to a certain point. They all acknowledge that a nonreductive materialist account of mind such as functionalism comes with a set of chronic and unresolved problems. Where they disagree, however, is in their reaction to these problems. Van Gulick, for example, would likely accept both Putnam's argument against the mathematization of nature and his subsequent claim that "abandoning 'identity theory' does not commit us to a form of dualism (Putnam, 1994, p. 483). At the same time, however, Van Gulick gives every appearance of being committed to his functionalist/representationalist account of mind.

The question to ask is: are the representationalist accounts of mind I have discussed in this thesis simply in need of adjustments and modifications, or is the entire functionalist undertaking simply wrong-headed as Putnam would have us belief? Van Gulick, and Dennett, and Dretske never claim to have resolved all the problems that come with the territory. On the other hand, they all seem optimistic that their (or some other functionalist's) next project will be the one that gets us where we want to go.

Putnam appears to counsel the abandonment of the functionalist project; but this, surely, would be seen as too drastic a step by those who have spent the last few decades refining their functionalist models. Furthermore, if current day philosophers of mind did give up on the functionalist project (as Putnam, McDowell, and others seem to suggest they should), what options are left for them?

5.6 "Philosophers are better at questions than answers" (Dennett, 1996, p. vii)

Putnam's conclusion in the "Dewey lectures" is that "what has weight in our lives should also have weight in philosophy" (1994, p. 517). I believe it would be wrong, however, to interpret these words as a rejection, or a discounting of, the connection that has developed during the last century between philosophy of mind and scientific disciplines such as psychology and/or cognitive science. After all, in the year 2000, "what has weight in our lives" can't help but be influenced by the results of the ongoing scientific investigation of the natural world which has been taking place since the time of the Ancient Greek philosophers.

McDowell states that, as philosophers, our aim should be "not to answer sceptical questions, but to begin to see how it might be intellectually respectable to ignore them, to treat them as unreal, in the way that common sense has always wanted to" (McDowell, 1994, p. 113). I maintain, however, that the role of common sense is not to discourage us from *asking* questions but, rather, to point us back in the right direction when our questions get bogged down in circularity.

In this thesis, I have attempted to show that the functionalist account of mind, at least in its present state, is not up to the task of explaining the connection between mind

and matter. My conclusion is *not*, as Putnam and McDowell seem to suggest, that the functionalist project should be abandoned. On the contrary, it is the questions and problems inherent in the functionalist accounts of mind provided by Van Gulick, Dennett, and Dretske that help to remind us of the role that philosophy could play in a very non-philosophic age.

According to my interpretation, Putnam (who rejects the functionalist model of mind) and Dennett (who embraces it) end up saying the same thing. But Dennett (as he often does) perhaps says it better: "Finding better questions to ask, and breaking old habits and traditions of asking, is a very difficult part of the grand human project of understanding ourselves and our world (1996, p. vii). It is the work of philosophers such as Dennett and Putnam that helps to show us the direction we should take next in the "grand human project" to which Dennett refers.

Works Cited

Akins K.A. and Winger, S. 1996. "Ships in the Night: Churchland and Ramachandran on Dennett's Theory of Consciousness," chapter 8 in *Perception*, (ed., Akins, K.A.). New York: Oxford University Press.

Baker, L. R. 1995. "Propositional Attitudes." In Guttenplan, ed., *A Companion to the Philosophy of Mind.* Cambridge, MA: Basil Blackwell. .

Bechtel, William. 1988. *Philosophy of Mind: An Overview of Cognitive Science.* Lawrence Erlbaum Associates: Hillsdale, New Jersey.

Block, Ned. 1978. "Troubles with Functionalism," from *Readings in Philosophy of Psychology.* Cambridge, MA.: Harvard University Press.

Block, Ned. 1993. Review of *Consciousness Explained* in *The Journal of Philosophy.* 15:4, pp. 188-193.

Block, Ned. 1995. "Functionalism (2)." In Guttenplan, ed., *A Companion to Philosophy of Mind.* Cambridge, MA: Basil Blackwell. pp. 323-332.

Byrne, Alex. 1995. "Behaviourism." In Guttenplan, ed., *A Companion to Philosophy of Mind.* Cambridge, MA: Basil Blackwell. pp. 132-140.

Campbell, Keith. 1984. *Body and Mind.* Notre Dame, Indiana: University of Notre Dame.

Davidson, Donald. 1970. "Mental Events." In D. Davidson, *Essays on Actions and Events.* Oxford: Claredon Press.

Davidson, Donald. 1995. "Davidson, Donald." In Guttenplan, ed., *A Companion to Philosophy of Mind.* Cambridge, MA: Basil Blackwell. pp. 231-236.

Davies, Martin & Humphreys, Glyn. W. (eds.). 1993. *Consciousness: Psychological and Philosophical Essays.* Cambridge, MA: Basil Blackwell.

Dennett, Daniel C. 1969. *Content and Consciousness.* London: Routledge & Kegan Paul.

Dennett, Daniel C. 1978. *Brainstorms: Philosophical Essays on Mind and Psychology.* Cambridge, MA: Bradford Books/The MIT Press.

Dennett, Daniel C. 1987. *The Intentional Stance.* Cambridge, MA: Bradford Books/The MIT Press.

Dennett, Daniel C. 1991a. *Consciousness Explained*. Toronto: Little, Brown & Co.

Dennett, Daniel C. 1991b. "Real Patterns," *Journal of Philosophy*, 87 (1) pp. 27-51.

Dennett, Daniel C. 1991c. "Ways of Establishing Harmony." In McLaughlin, Brian, ed., *Dretske and His Critics*. Cambridge, MA: Basil Blackwell.

Dennett, Daniel C. 1995. "Dennett, Daniel C." In Guttenplan, ed., *A Companion to Philosophy of Mind*. Cambridge, MA: Basil Blackwell. pp. 236-244.

Dennett, Daniel C. 1996. *Kinds of Minds: Towards and Understanding of Consciousness*. New York: Basic Books.

Dennett, Daniel C. 1998. *Brainchildren: Essays on Designing Minds*. Cambridge, MA: Bradford Books/The MIT Press.

Dretske, Fred. 1969. *Seeing and Knowing*. University of Chicago Press.

Dretske, Fred. 1981. *Knowledge and the Flow of Information*. Tucson: University of Arizona Press.

Dretske, Fred. 1988a. *Explaining Behavior: Reasons in a World of Causes*. Cambridge, MA: A Bradford Book/The MIT Press.

Dretske, Fred. 1988b. "The Stance Stance." *Behavioral and Brain Sciences*. 11:3. 511-512.

Dretske, Fred. 1995a. "Dretske, Fred." In Guttenplan, ed., *A Companion to the Philosophy of Mind*. Cambridge, MA: Basil Blackwell. pp. 259-264.

Dretske, Fred. 1995b. *Naturalizing the Mind*. Cambridge, MA: A Bradford Book/The MIT Press.

Fodor, Jerry, A. 1975. *The Language of Thought*. New York: Thomas Y. Crowell.

Fodor, Jerry, A. 1984. "Semantics, Wisconsin Style." *Synthese* 59: 231-250.

Guttenplan, Samuel (ed.) 1995. *A Companion to the Philosophy of Mind*. Cambridge, MA: Basil Blackwell. pp. 3-107.

Horgan, Terence. 1993. "From Supervenience to Superdupervenience: Meeting the Demands of a Material World," *Mind*. Vol. 102. 408. 555-586.

Kim, Jaegwon. 1991. "Dretske on How Reasons Explain Behavior." In McLaughlin, Brian, ed., *Dretske and His Critics*. Cambridge, MA: Basil Blackwell.

Kim, Jaegwon. 1996. *Philosophy of Mind.* Boulder, Colorado: Westview Press.

Levine, Joseph 1983. "Materialism and Qualia: The Explanatory Gap," *Pacific Philosophical Quarterly.* Vol. 64. pp. 354-361.

Lycan, William. 1995. "Functionalism (1)." In Guttenplan, ed., *A Companion to Philosophy of Mind.* Cambridge, MA: Basil Blackwell. pp. 317-323.

Marras, Ausonio. 1994. "Nonreductive Materialism and Mental Causation." *Canadian Journal of Philosophy.* 24:3. 465-494.

McDowell, John. 1994. *Mind and World.* Cambridge, MA: Harvard University Press.

McGinn, Colin 1991. *The Problem of Consciousness.* Cambridge, MA: Basil Blackwell.

Nagel, Thomas. 1995. *Other Minds: Critical Essays, 1969-1994.* New York: Oxford University Press.

Olsen, Christopher. forthcoming. *The Janus Character of Mind/Brain: Mental Representation, Cognition and Consciousness.*

Papineau, David. 1995. "Content (2)." In Guttenplan, ed., *A Companion to the Philosophy of Mind.* Cambridge, MA: Basil Blackwell. pp. 225-230.

Peacocke, Christopher. 1995. "Content (1)." In Guttenplan, ed., *A Companion to Philosophy of Mind.* Cambridge, MA: Basil Blackwell. pp. 219-225.

Place, U. T. 1956/62. "Is Consciousness a Brain Process?" In V.C. Chappell, ed., *The Philosophy of Mind.* Englewood Cliffs, NJ: Prentice-Hall.

Putnam, Hilary. 1988. *Representation and Reality.* Cambridge, MA: A Bradford Book/The MIT Press.

Putnam, Hilary. 1994. "Sense, nonsense, and the senses: an inquiry into the powers of the human mind." *The Journal of Philosophy.* XCI:9 445-517.

Rey, Georges. 1995. "Concepts." In Guttenplan, ed., *A Companion to the Philosophy of Mind.* Cambridge, MA: Basil Blackwell. pp. 185-193.

Rosenthal, David. 1995 "Identity Theories," In Guttenplan, ed., *A Companion to Philosophy of Mind.* Cambridge, MA: Basil Blackwell. pp. 348-355.

Seager, William. 1997. "Critical Notice of Dretske's *Naturalizing the Mind*". *Canadian Journal of Philosophy.* 27:1, pp. 83-109.

Searle, John. 1992. *The Rediscovery of the Mind*. Cambridge, MA: The MIT Press.

Searle, John. 1995. "Intentionality (1)." In Guttenplan, ed., *A Companion to Philosophy of Mind*. Cambridge, MA: Basil Blackwell. pp. 379-386.

Searle, John. 1997. *The Mystery of Consciousness*. New York: A New York Review Book.

Sedivy, Sonia. 1995. Consciousness Explained: Ignoring Ryle and Co.," *Canadian Journal of Philosophy*. 25 (3) pp. 455-483.

Smolensky, Paul. 1995. "Computational Models of Mind," In Guttenplan, ed., *A Companion to Philosophy of Mind*. Cambridge, MA: Basil Blackwell. pp. 176-185.

Van Gulick, Robert. 1982. "Mental Representation – a Functionalist View," *Pacific Philosophical Quarterly*. Vol. 63 pp. 3-20.

Van Gulick, Robert. 1988a. "Consciousness, intrinsic intentionality, and self-understanding machines," Chapter 4 in *Consciousness in Contemporary Science*. A. J. Marcel and E. Bisiach (eds.). Oxford: Clarendon Press.

Van Gulick, Robert. 1988b. "A Functionalist Plea for Self-Consciousness," The Philosophical Review. Vol. XCVII, No. 2.

Van Gulick, Robert. 1993. "Understanding the Phenomenal Mind: Are We All Just Armadillos?" Chapter 7 in *Consciousness: Psychological and Philosophical Essays*. M. Davies and G. W. Humphreys (eds.). Cambridge, MA: Basil Blackwell.

Van Gulick, Robert. 1995a. "Who's in Charge Here? And Who's Doing All the Work?" In J. Heil and A. Mele, eds., *Mental Causation*. Oxford, UK: Clarendon Press.

Van Gulick, Robert. 1995b. "What Would Count as Explaining Consciousness?" In Thomas Metzinger, ed., *Conscious Experience*. Schoningh: Imprint Academic.