Saving Swampman's Mind
–
Physical Intentionality Without History

By
Peter Kelly

Submitted to
Central European University
Department of Philosophy

In partial fulfilment of the requirements for the degree of Master of Arts

Supervisor: Professor Katalin Farkas

Budapest, Hungary
2014

Table of Contents

Introduction

How do you think about? This is not the grammatically incomplete question it may initially seem: our thoughts are about things other than themselves (including other thoughts); our beliefs are about states of the world; our sensory interactions with the world are mediated by representations that are directed toward those things they represent. Our mental lives have an inherent 'aboutness' that is difficult to explain. This difficulty is brought to the fore if you believe, as many do, that the mental is in fact physical; that it is at least realised or produced by the brain – or even identical with it, and entirely reducible to physical matter. To hold either of these positions is to be a physicalist, and being a physicalist makes the problem of the aboutness of the mental – also known as the problem of intentionality – particularly difficult to tract.

Why is intentionality a problem for the physicalist in particular? Physicalism is the thesis that everything is physical. This includes the mental, and all of its accompanying intentional states; beliefs, desires, representations are all examples of intentional states, since they are all directed towards or about something other than themselves. Combining these two ideas – that certain of our mental states are intentional in nature, and that everything is physical – leads immediately to the problem: how can purely physical matter be *about* anything at all?

Clearly this text is about something (hopefully exactly what will become clear). But this aboutness requires you, the reader, to be an interpreting agent; the shapes and lines on this page are not intrinsically about anything and instead rely on a shared understanding of what we, in our linguistic community take the shapes and lines to represent. When it comes to mentality, its intentionality, and physicalism, we do not have the luxury of appealing to an interpreter whose role is to provide meaning. If physicalism is correct – particularly if reductive physicalism is correct – then your mind – the interpreter – is nothing more than a collection of mindless processes occurring in your brain. In seeking to explain intentionality physicalistically, we must build that interpreter up from nothing more than the mindless constituents of the physical world.

Of course, nobody is required to accept any form of physicalism, and there are plenty of alternative positions on offer. But for this thesis I will be assuming a reductive physicalism, a type-identity theory of the mental. I will be assuming throughout that there is only physical matter making up our reality, and that all of our mental states and processes can be reduced to – are identical with – our brain (or bodily) states and processes. There are abundant reasons to accept this physicalist viewpoint, including the fact that this view is a good fit with science – one that respects the fact that our most profound discoveries in recent years have been produced by science, implicitly assuming a form of physicalism. I appreciate that this is certainly no knock down argument in favour of my assumptions, but the goal of this thesis is to explore, in depth, a particular problem for physicalist theories of intentionality, a problem that only arises if we assume the truth of physicalism.

As a result, physicalism does not need to be true for what follows to be (I hope) a worthy exercise.

Physicalist theories of intentionality have tended in recent years to seek to explain the aboutness of the mental in terms of evolution and evolutionary function. This move allows philosophers to get a handle on the basic problem of intentionality: if mental content is about something, then it seems to have a certain purpose, or teleology, to it. Evolution via natural selection gives the physicalist a scientifically respectable method of pointing to purpose in the natural world. In what follows I will be following suit by adopting a teleological approach to mental content and intentionality.

The first chapter will describe in some detail two of the leading teleological theories. I will point out those aspects that I think prove useful, and those aspects that I think either do not work or are unnecessary. This chapter will also take a look at a competing theory of mental content that has proven rather influential, in particular because it provides a novel solution to a problem that its author believes to plague all teleological theories. The ultimate purpose of this chapter will be to justify my choice of teleological theory, which I will then use as a framework for answering a particularly tricky problem for these types of theories.

Chapter two will describe this problem in detail. The Swampman objection, as it is known, provides us with a tricky choice: deny a strong intuition about the mentality of a seemingly conscious creature, or deny the plausibility of any theory of content that appeals to teleology for purpose. Contrary to most

teleological theorists, I will, in this thesis, refuse to deny the intuitions in this thought experiment. At the same time, however, I believe that a certain teleological theory can provide us with a solid framework for understanding intentionality – even if it requires some reconstruction in order to understand intentionality in the Swampman thought experiment.

Finally, in chapter three I will spell out my proposed solution to the Swampman experiment, using the framework discussed in previous chapters. This approach will evaluate some of the key terms of my chosen theory and attempt to rework the account given therein so that it does not rely on the constraints that undermine its ability to account for the Swampman objection. My method will largely involve pressing the idea that a long tail of evolutionary history is not necessary for physical processes to have functionality. If this is successful, it should provide us with a way to understand intentionality in purely physicalist terms. Furthermore, even if some of the concepts prove more difficult to reformulate than others, this work should serve to illuminate the way that certain commitments of the teleosemantic theories serve to constrain their explanatory power – and perhaps unnecessarily so – when it comes to dealing with objections such as Swampman.

Chapter One

Teleosemantics

As mentioned above, a promising set of theories for dealing with the problem of physical intentionality are those that appeal to teleology. Here I will look at two such theories: Fred Dretske's information –theoretic account and Ruth Millikan's Biosemantics, taking a detour via the disjunction problem and Jerry Fodor's attempts to overcome it. Both teleological theories, in attempting to give an account of intentionality, rightly focus on the ability of intentional systems to *mis*represent. This is because any theory of representation and intentionality must be able to accommodate the fact that we can get things *wrong* in our thinking: we can believe falsehoods, desire fabrications and misperceive our environment. As we will see, a major part of the teleosemantic project is providing a way to understand this possibility of error, this capacity to misrepresent.

1.1 Dretske's Information-Theoretic Account

The first step in Dretske's approach to intentionality is to notice that there exist natural signs in nature that signal the presence of other phenomena.[1] Certain features of the world reliably provide information about other features of the world through causal co-variance; by nomologically co-varying with what they

1 Dretske, Fred (1981). Knowledge and the Flow of Information. MIT Press.

indicate, natural signs indicate the presence of these phenomena. For example, the presence of smoke naturally indicates the presence of combustion of some sort; the presence of a certain type of spots can reliably indicate the co-presence of measles. Natural signs cannot *mis*indicate, however. While it is possible to have spots in the absence of measles, this is not a case of mis-indication or misrepresentation. The particular type of spots that appear as a result of measles can only appear in the presence of measles. Other, differently caused spots can appear, but these are not natural signs of measles, but of their specific causes. Natural signs cannot get things wrong, because their presence *nomologically* co-varies with the presence of their causes: if they did not then they would not be considered natural signs. Because we are dealing with *causal* co-variance, the effect is a sure sign of its cause; if it is not, then we do not have causal co-variance and we are not dealing with natural signs.

Representations are different. Since we are trying to give an account of representation, and any account of representation must also be an account of misrepresentation, appealing to causal co-variation can only be a first step. With natural signs – which have the ability to reliably co-vary with features of the environment – we have a method for transmitting information. In order provide room for error in the picture, we need some normative apparatus, a way of describing the use of transmitted information as being either correct or incorrect.

To do this Dretske appeals to the notion of function. If it is the function of a system to indicate the presence of a feature in the environment, then this

functionality can presumably go wrong. We can see that in the case of artefacts. An example that Dretske uses is the petrol gauge on a car: this (human-designed) artefact has the function of representing the level of fuel in the car's fuel tank. But it can go wrong. If the gauge is hooked up to the car incorrectly, or it is being used in extreme-cold conditions that cause its machinery to freeze, the gauge can misrepresent the level of fuel present in the fuel tank – it can be wrong about that which it indicates. Despite the fuel gauge causally co-varying with whatever it is connected to – as with natural signs – the gauge allows for error precisely because it has a specific function to perform set by the designers of the gauge. If it is not able to carry out this function due to operating in conditions that are relevantly different to those within which it was designed to operate, then the fuel gauge will misrepresent the level of fuel in the car.

It is because the fuel gauge has a definite function that it is able to accommodate misrepresentation, or cases of error. This function is built into the device by the design of humans, by *intentional agents*. As a result, any representational capacity possessed by the gauge depends upon its being assigned functions. To bring this point out clearly Dretske discusses the example of the 'Twin Tercel': a car that mysteriously materialises in a scrap yard. This car will have a fuel gauge, but without intentional designers bestowing it with a definite function we cannot say that it represents or misrepresents anything.[2]

2 Dretske, F. (1996). Absent qualia. *Mind & Language*, *11*(1), 78-85.
I will return to this case later, since I do not think that Dretske is entirely correct in his assessment of the Twin Tercel's fuel gauge. But for now it serves as a good illustration of the supposed need for designed functions for representation to occur.

Returning to natural signs, for Dretske's project to get off the ground we need a way to naturally assign functions to information carrying natural signs. Functions, as we have seen, require design or purpose. Fortunately, there is a powerful, method in nature through which design and purpose naturally occur, and it is the method by which we – naturally representing, intentional systems – came into being.

Evolution by natural selection produces functions without any need for a higher-level, purposive designer. It is perfectly natural to say that living things, and the parts thereof, have functions: hearts have the function to pump blood, livers have the function of processing toxins, certain gut bacteria have the function of consuming toxins harmful to their hosts in a symbiotic relationship therewith. Through a process of selection and environmental filtration, certain processes attain the status of functions through contributing to an organism's adaptivity in its environment. Those certain beneficial processes are traits of the organism, and those organisms lacking such traits and will be more likely to be selected against than those organisms with these adaptive traits.

Translating this notion to the mind, Dretske claims that certain features of our mental apparatus, those capable of indicating, have attained the function to do so. This is directly analogous to the petrol gauge in Dretske's car: in that case the gauge's ability to indicate was assigned the function of representing by its designers; in our case certain of our physical processes act as natural signs of our environment, and because this contributes to our evolutionary fitness, natural selection has assigned these processes the function of representing our environment. Our visual system, for example, has the

function of representing the external world visually because the processes that constitute its ability to indicate via the causal co-variation of its parts and the environment contributed to our evolutionary success. As a result, there are optimal conditions under which those processes can operate: they will be best suited to conditions similar to those in which they were selected for in our evolutionary past. If those processes operate in sub-optimal conditions, then there is the possibility of error – resulting in misrepresentation.

One illustration that Dretske offers involves ocean-dwelling bacteria that live near one of the Earth's poles. These bacteria possess magnetosomes: tiny magnets that indicate the presence of magnetic fields, which when activated cause the bacteria to move in their direction. In optimal conditions (i.e. those conditions under which the bacteria were subject to selection pressures that resulted in the magnetosomes' functionality), the direction of a magnetic field is also the direction of the oxygen-starved water in which the bacteria flourish. According to Dretske, these magnetosomes have the teleo-function of representing good conditions for life. The bacteria's magnetosomes indicate the direction of magnetic fields, and natural selection filtered out those bacteria lacking in this ability. As such, the ability to indicate attained the status of a function to represent suitable environs for the bacteria. As with the car's petrol gauge, the ability to indicate was assigned the representational function by its designer; only in the case of the bacteria, that function was assigned by Mother Nature.

As for the bacteria's ability to misrepresent, this comes down to the magnetosomes operating under the wrong conditions. If, for example, a magnet is held near these bacteria, then the magnetosomes will still indicate

the direction of a magnetic field but its evolutionarily assigned function to represent suitable living conditions will be (potentially fatally) disrupted. By holding a magnet overhead we are removing the optimal conditions for the magnetosome's functioning; the bacteria were not subject to selection pressures in the presence of any magnetic fields other than those produced by the Earth's poles.[3] In these sub-optimal conditions the magnetosomes will still indicate magnetic fields, but they will *misrepresent* suitable conditions for life: it is the magnetosomes' function to represent suitable conditions under optimal conditions, and under non-optimal conditions they will malfunction and misrepresent.

A problem for Dretske's account is that if representations can be reduced to information then there seems to be no way to determine the content of a representation. For example, we have no way of determining if the bacteria discussed above are representing good conditions for life, or magnetic fields, or oxygen-free water, or the Earth's pole – and so on through any number of possible distal sources of information that the indicating magnetosomes are subject to. If the magnetosomes were selected for their ability to indicate, and this selection resulted in the indications having the function to represent, then there should be some determinate object of representation that is the reason for the bacteria's success throughout its selection history. [4]

3 This is an idealization: presumably throughout its evolutionary history the bacteria was subject to other sources of magnetism, but this simplification amply illustrates Dretske's main point.

4 Fodor, Jerry A. (1990). Information and representation. In Philip P. Hanson (ed.), *Information, Language and Cognition*. University of British Columbia Press.

Dretske was aware of this issue, and attempted to get around it by introducing learning conditions for an organism and more complexity to the signalling system in an organism. Suppose that it is the function of an organism to detect (and move towards for the purposes of survival) conditions *C*. It does this via two signaling systems that work by detecting two different signs, *f1* and *f2*, of *C* – say, for example, magnetism and smell. Only one of these is needed to put the organism into state *S.* Now, when the organism goes into state *S* it does not naturally mean that the organism has detected via *f1*, since it could equally go into the same state via *f2*, and vice-versa. According to Dretske we can now say that *S* indicates *C*, and not *f1* or *f2*. When we hold a magnet near our imagined bacterium (newly replete with olfactory capabilities) and cause it to enter state *S* we are no longer puzzled as to whether it is misrepresenting the presence of *C* or representing the presence of a magnetic field. It is misrepresenting *C*.

But the problem hasn't really gone away. Although we want to say that *S*'s function is to detect *C*, we still have to somehow account for *S* naturally meaning *C* and not the disjunction *either f1 or f2.* It will always be possible, or so it seems, to describe the function of *S* as being to represent *C* or the disjunction of its causes. This is taken to be a big problem for a naturalized account of representation, since we do not want it to be the case that individuating possible representations comes down to the issue of how *we* choose to describe the situation.

Dretske responds to this by appealing to associated learning. Roughly, any new stimulus can become a conditioned stimulus through learning, so the

thought is that repeated exposure to a particular *C* will bring with it new signs: new features of *C* will come to affect *S*, and as a result, over time, the nature of *S* will change given the different stimuli to which it responds. What will stay the same is that it will always indicate *C*. It will *invariably* indicate *C* despite their being an ever increasing disjunct of possible stimuli responsible for *S*. As such, the system has a definite object of representation: the invariant *C*.

However, it is difficult to see what exactly it is about the supposed object of a representation *C* that makes it a more likely candidate than the possible endless disjuncts. While it might be intuitively appealing to say that *surely* it is the time-invariant object and not the ever-changing list of disjuncts that is represented, there is nothing in Dretske's account that shows why this *must* be so without appealing to the very notion of purpose of which he is seeking to give a naturalistic account.

The above problem of multiple distal causes, most notably championed by Jerry Fodor, is part of a family of problems that come under the heading of the disjunction problem. In the next section we will look into the disjunction problem in more depth, and discuss Fodor's own response to the objection: his Asymmetric Dependence Theory of Content.

1.2 The Disjunction Problem and Fodor's Asymmetric Dependence Theory

The disjunction problem was forcefully applied to teleosemantic theories such as Dretske's by Jerry Fodor in his *A Theory of Content I: The Problem.*[5] Essentially, the disjunction problem is the following: a mental token (in our case, a representation) can have any number of distal causes and as such can carry information about any number of these distal causes. However, according to Dretske, representation should be reducible to information, and as such the meaning of a token should be reduced to any of the distal causes that the token carries information about. The problematic result for teleosemantics is that the meaning of a token (or the content of a representation) will thereby be analysed into a disjunct of all of the potential distal causes that the token carries information about. A representation of a dog could be caused by a dog or a dog-shape (a cat at night, say), and the theory provided by Dretske has no way to adjudicate as to whether the content of the representation is *dog* or *dog-shape* or *cat-at-night* or *dog or dog-shape or cat-at-night.* It is a live possibility that any of these is the content of the representation, even under optimal conditions.

As a result, Fodor thinks that any appeal to teleology in determining content is doomed to failure, as is any attempt to reduce meaning to information. If meaning is supposed to reduce to information, then it seems we have a real problem in determining the meaning of a token (i.e. the content of a representation) when that meaning and the information it carries (its distal cause) do not align. A token's meaning should be invariant, despite the fact that the token can be caused by (carries information about) any number of distal causes.

5 Fodor, J. A. (1990). *A theory of content and other essays*. The MIT press.

Enter Fodor's own Asymmetrical Dependence Theory (ADT). According to ADT, content fixing relies upon an asymmetrical dependence between tokens that are inappropriately caused (e.g. a "dog" token caused by a cat at night) and tokens that are appropriately caused (e.g. a "dog" token caused by a dog). As the disjunction problem notes, a dog and a cat at night both cause "dog" tokens, and teleological theories of content have no non-arbitrary way of picking out the correct description of content; but according to Fodor, the token "dog" means dog because cat-at-night-caused "dog" tokens depend for their content upon dog-caused "dog" tokens, and *not* vice-versa. In the case of an accurate representation, the meaning of a token derives from a lawful connection between the token and the world as-is (e.g. "dog" and dog); in the case of misrepresentation (e.g. "dog" caused by cat-at-night), the token's meaning is entirely dependent upon this other non-dependent, lawful connection.

ADT claims that a "dog" token means dog because under 'normal' conditions it is caused by a dog, and under 'non-normal' conditions its tokening depends upon it having been caused by a dog in the past. For Fodor, however, there are no 'normal' nor 'non-normal' conditions: in *any* situation, if a "dog" token is caused by a dog then meaning and information align. Similarly, in *any* situation, if a "dog" token is caused by something other than a dog - if it carries information about, say, a cat at night - then meaning and information do not align and, as such, the token's meaning "dog" relies upon its having aligned with dog-caused information in the past.

With the ADT, Fodor aims to provide a naturalistic conception of intentionality without resort to optimal conditions or talk of functions. The theory is designed seemingly from the ground-up to circumnavigate the disjunction problem, and it mostly manages this. Unfortunately for Fodor, however, it is not clear that his theory can work without helping itself to the very intentional idiom that physicalistic/naturalistic theories of content are trying so hard to avoid; and neither is it able to produce the correct results for meaning without making reference to those very optimal conditions he hoped to banish.

Let's start with an example of ADT's reliance upon optimal conditions. Let's say that you are looking out your window and are representing a dog because you are, in fact, looking at a dog. According to ADT, this representation is about a dog because of the lawful connection between "dog" tokens and dogs. However, unbeknownst to you, this morning I slipped a chemical compound of my own design into your morning coffee. Being interested in tricks of perception, I designed this compound such that the only thing it does is to trick your perceptual system into tokening "dog" representations in response to cats, and vice-versa. Now, because of my drug, you are representing a cat as a dog in a completely lawful way: it is the case that, under the present conditions of intoxication, there is a lawful connection between cats in the world and your tokening "dog" representations. Similarly, there is a lawlike connection between dogs in the world and your tokening "cat". Given this state of affairs, it seems like your "dog" representation can mean either dog or cat, since either distal causes will produce your "dog" token. Importantly, however, it is *not* the case that the "dog"-caused-by-cat token *means* dog because it asymmetrically depends on the lawful connection

between dog and “dog” as ADT claims: there exists a lawful connection between your “dog” representation and the cat outside your window, and your “cat” representation and a dog outside your window. There is no dog/”dog” or cat/”cat” law present for either to asymmetrically depend on. Something is wrong: given your intoxicated state you are tokening “dog” in response to cats in a lawful way (and vice-versa), and ADT has no way to tell us why.

What is wrong in this case? Clearly the drug is affecting your normal, optimal functioning. Without the drug in your coffee, you would represent the cat as a cat; your “cat” token would *mean* cat. But the drug has caused you to token “dog” – and “dog” should still *mean* dog despite the lawful connection between “dog” and cat. What ADT needs, it seems, are reference to optimal conditions. Under optimal conditions (in this instance, conditions that don’t involve you staring out the window hallucinating random dogs), a cat would cause you to token “cat” because of the lawful connection between cats and “cat” tokens. If a cat caused you to token “dog”, it would be because of a lawful connection between dogs and “dog” tokens upon which your “dog”-in-response-to-cat tokening asymmetrically depends – as ADT predicts. But without reference to optimal conditions – without building in ceteris paribus laws that explicitly allow for mistakes when things are not going as should be – ADT makes incorrect predictions: in this case your “dog” token will mean dog or cat. As a result, ADT cannot be a correct theory of intentionality without the optimal conditions that Fodor denies.

Furthermore, ADT cannot be a correct physicalist theory of intentionality, since Fodor has failed to specify the mechanism by which his theory picks out

the foundational semantic connections upon which misrepresentations' meanings rely. As pointed out by Gibson[6], the only lawlike connection we can really point to between "dog" tokenings and dogs is the connection between "dog" tokens and *looking like* a dog. Under optimal conditions, the property in question that is playing a causal role between dogs and "dog" tokens is the property of looking like a dog: if a cat causes "dog", then we are in no better position saying that it is because this relation asymmetrically depends on dogs causing "dog" than we are if we say that dogs cause "dog" by asymmetrically depending on the lawful connection between a cat that looks like a dog and "dog".

So what is it that makes the fact that dogs cause "dog" the foundational law to be used in all other explanations? The only way we have to get a grip on this is the fact that "dog" means *dog* and not *cat*. But this is a semantic explanation. The very thing we are trying to explain is required in order to make sense of the asymmetry on which Fodor depends. Dretske's proposal is able to account for why the dog/"dog" law is primary to all others: under optimal conditions, "dog" carries information about dogs in virtue of it being caused by dogs and not cats. In the case above, the optimal conditions will make reference to you not being on a drug that systematically changes your response to the external world; this is because the conditions under which your capacity to indicate and represent features of your environment evolved did not include any such drug. Fodor's theory, on the other hand, seems to

6 Gibson, M. (1996). Asymmetric dependencies, ideal conditions, and meaning. *Philosophical Psychology*, *9*(2), 235-259.
Gibson does not use the dog/cat example, but I will stick with it here for the sake of consistency.

rely upon the fact that “dog” independently means dog in order to explain why all other tokenings of “dog” asymmetrically depend on dogs causing “dog.

The ADT, in fact, does hold up as a method to find out the meaning of a term, so long as we build in other conditions: namely, optimal or normal conditions. Once we have done this, we can see why it is the fact that “dog” means dog and not cat: under optimal conditions “dog” carries information about dogs, and under non-optimal conditions it will carry information about things that look like dogs (or more, depending on whether I’ve been sneaking my drug into your coffee). Without these normal conditions there is no reason to promote one lawful connection over the others, whether it be the cat-viewed-while-intoxicated/”dog” connection or looks-like-a-dog/”dog” or even dog/”dog”. As a result, ADT is potentially a useful tool for picking out the meaning of a token (or the content of a representation), but it does not explain why the token means what it does, the meaning of the token being explanatorily prior to the asymmetric dependence through which ADT operates.

In the next section I will look at one other teleological theory of mental content: biosemantics. Posited by Ruth Milikan, Biosemantics is similar to Dretske’s proposal in that it uses the concept of evolutionarily-defined functions in order to bring a normative aspect to bear on the physical processes that take place within us as representing systems, thereby allowing for misrepresentation to occur. Biosemantics is also able to avoid the disjunction problem, by splitting representational systems into two parts: a producer and a consumer.

1.3 Biosemantics

Biosemantics uses the concept of indication pretty much as Dretske describes it. To give an example in terms of human beings, there are, through causal connection, features of our nervous systems that co-vary with features of the environment. These are said to indicate features of our environment and, as discussed above in the section on Dretske, there is no such thing as mis-indication. As with Dretske, biosemantics gets a handle on error and misrepresentation by appealing to the notion of functions – Milikan calls these 'proper functions' – and optimal conditions – which Milikan calls 'normal conditions'. In the biosemantic account of intentionality the concepts that provide the majority of the explanatory power are those of proper functions, normal conditions, and the producers and consumers of representations.

Proper functions are those effects that have allowed an organism's ancestors to reproduce or flourish and have accounted for its selection throughout its evolutionary history. To find out the proper function of some organism, we must look at those instances where the function has succeeded in accounting for the success of the organism in question. In order to find out the function of the human heart, for example, we should look at well functioning human hearts – not diseased or otherwise malfunctioning hearts, but those hearts for which their functions contribute to their host's on-going survival and reproductive success through the generations. In order to establish if a process counts towards the organism's success (thereby becoming a proper function), we have to look at the past successes of those processes in allowing the organism to survive and reproduce. As such, Millikan holds that

proper functions must be viewed relative to the past selection success of the organisms to which they belong, and not the functioning object's current dispositions.

Let's return to the example of the human heart. The human heart has function *F* (pumping blood) because, in its ancestors, it was *F* along with the conditions in which selection pressures acted upon it that accounted for its survival and ultimate reproduction.[7] According to Millikan we cannot just look at the dispositions of a single heart in isolation to learn its function, since for all we know this could be a malfunctioning heart. It must be within the context of an evolutionary history that we determine a thing's function. In fact (and relevantly to the third chapter of this thesis) Millikan explicitly states that nature could accidentally produce "items that (freakishly) have the exact form of existent hearts but that are not hearts (because their history is wrong)".[8]

I will be returning to this point later, but a quick digression is required to challenge Millikan on this point. The term 'proper function' as she has defined it makes explicit reference to the history of selection of the possessor of that function. As such, something cannot have the function to pump blood unless it has a history of doing so throughout the generations. As quoted above, a heart that appears along with its host through a freak of nature and performs the process of pumping blood – and in doing so keeps its owner alive – does not have the function of doing so. This is because we have no way of knowing that this process is occurring correctly; of knowing that by doing so it is

7 Obviously a heart does not reproduce, but the organism for which it pumps blood. Still, I trust the idea is clear enough.
8 Millikan, R. G. (1984). *Language, thought, and other biological categories: New foundations for realism*. MIT press. p. 29.

contributing to the organism's continued survival and potential for reproduction. But this seems to be a very strong constraint. Presumably if we witnessed this 'freakish' heart in operation, and saw that the blood pumping process was contributing to the organism's survival, we would be strongly inclined to attribute to it the function of pumping blood.

Millikan's claim that "having a proper function depends on the *history* of the device that has it, and not upon its form of dispositions"[9] is true in the sense that without seeing the device in action – by just looking at the form of its dispositions – we cannot tell its functions; but if we were to see the device in action and see what processes contribute to its continued operation then I think we have a very real way to attribute to it various functions. Of course, the problem is that, with no other examples of this kind to compare these processes to and thereby judge its relative successes and failures, it will be difficult to know for sure if the processes we are tempted to label functions are actually the functions that it *should* be performing. This is an important point, considering the weight that is placed on the normative nature of functions in accounting for cases of error. I will be returning to this later.

To recap: proper functions, according to Millikan, are those effects which have in the past contributed to an organism's ancestor's ability to reproduce. As for normal conditions, the next key term in the biosemantic account, it is necessary to look first at the notion of a normal explanation. A normal explanation is an explanation of how an organism is able to perform its proper function, making reference to the structure of the functioning device in

9 Ibid.

question, and the normal conditions under which, historically, it was able to perform this function. These normal conditions essential to providing an explanation of proper function, then, are those conditions under which the functioning device effectively performed its function.

To return to the bacteria example provided by Dretske, the bacteria's magnetosomes have the proper function of signalling the presence of oxygen-free water. The conditions under which this function can occur successfully are those conditions in which selection pressures have operated in the bacteria's evolutionary past. The sudden switching of the Earth's poles, or the presence of a nearby magnet are not, then, normal conditions for the magnetosome's proper function.

Again, Millikan's requirement that normal conditions be viewed in a purely historical light is something I will challenge later in this paper in more depth, but for now it is worth pointing out that, just as with the concept of the proper functions of a freakish, spontaneously appearing heart, I do think that there is a sense in which we can identify the normal conditions for its proper functioning. Those conditions will be the conditions in which the heart is able to contribute to the organism's survival. Again, however, just as with my dispute over proper functions, I will need to account for the fact that we just don't know if the conditions we would be willing to attribute to the freakish heart as normal *are* actually the normal conditions, since we lack the historical record with which to compare these conditions to those of the heart's ancestors. Again, I will return to this later.

Normal conditions are highly important to this theory, since it is normal conditions that partly determine the content of a representation. Normal conditions are those under which a representation, as used in accordance with a consumer device's proper function, maps onto the world in a systematic way. It is worth noting that the nature of this mapping can be different among producer/consumer chains, and even among multiple consumers using the same produced representations What is important is that the producer produces representations in a way that the consumer can read them appropriately; the mapping of representation to world (under normal conditions) can be transformed in an indefinite number of ways so long as the consumer can read them. So in the case of a visual representation, the normal conditions under which a representation can be used appropriately for the proper functioning of the representational system are those in which the representation matches or corresponds with the world in an appropriate way. Under normal conditions the representation as used by the consumer part of a representational system will map to the represented in a suitable way for the consumer part to carry out its proper function.

It is Millikan's account of the producers and consumers of a representation that gets the biosemantic theory around the varieties of the disjunction problem discussed above. Millikan uses the following as an example of a representational system at work between different creatures: when a beaver splashes its tail on the water, it does so in order to signal danger. Other beavers, upon seeing and hearing this splash, instinctively dive underwater to the safety of their home. In this scenario, the splashing beaver is the producer of the representation. The consumers are the beavers that use that

representation *as* a representation of danger, and act accordingly. This avoidance behaviour is the proper function of the consumers in using that representation *as* a representation of danger, and normal conditions are those in which the use of this representation of danger maps to the world in the appropriate way; under normal conditions the splash, when used by the consumer function as a sign of danger in accordance with the consumer's proper function, actually maps to real danger, and is used as such.

Of course, this means that normal conditions can rarely obtain: Beavers are skittish creatures, producing warning splashes often in the absence of danger. In such cases, the representation will be used by the consumer beavers as representations of danger (resulting in avoidance behaviour), in accordance with their proper function. But because the representation does not map to the world correctly (i.e. there is no danger), these are not normal conditions. Because the normal conditions (conditions under which there is actually danger causing the production of the representations) fix the content of the representation as used by the consumer, in cases where there is no danger the representations still have the content of "danger"; the difference is that when there is no actual danger these representations actually misrepresent.

As for cognitive systems, it is not entirely clear what the producers and consumers of representations are. An approximation would be something like the following: the retina in the human visual system produces representations as it indicates features of the world. Higher-level areas of the visual system and the motor system will consume these representations in accordance with their proper functions, and the normal conditions will be those in which there is

a mapping between the representation and the represented. We can clearly see how it would be the proper function of parts of the motor system in humans to undertake avoidance manoeuvres in response to a representation of a falling tree. In this case, the visual system (whether it be the retina or some part of the visual cortex) is the producer of the representation, and the motor system is the consumer.

Millikan's proposal differs in a number of ways from Dretske's. For a start, representation is not a dyadic relation between the represented and the indicator function. By splitting the representational system into two parts, creating a relation between represented, producer and consumer, Millikan shifts the burden of semantics onto the proper function of the consumer: if the representation is used *thusly*, then it means *X.* If this is occurring in normal conditions, then *X* actually obtains in the world. The representation's meaning does not arise from the natural meaning of the indicated as in Dretske; it arises from the use to which it is put. In fact natural meaning as Dretske makes use of it drops out of the picture entirely. Due to the fact that the producer can transform the representation in any way, just so long as the consumer can interpret it, the representation does not need to correspond in a natural meaning sense to the represented. There does need to be a mapping rule, but that can take any form so long as the producer and consumer devices are speaking the same language.

The reason this way of structuring the representational system avoids the two varieties of the disjunction problem discussed above is because there are clear normal conditions that should tell us the actual content of a

representation as used by the consumer. The normal conditions in the previous example would be an actual falling tree, since it is by consuming this representation as such that a human being would be able to survive and reproduce, passing the proper functioning of its motor system in consuming representations as such down through the generations.

Returning to the example from the start of this chapter, if you see a dog or a cat at night, according to biosemantics the content of your representation will be “dog”, because that is how the representation is being used. If the distal object is actually a cat, then the consumer part of the system has not carried out its proper function properly, because of the fact that the normal conditions for its proper functioning are not present. Under normal conditions, the distal object would be a dog, since the consumer part is treating this representation as one of “dog”; perhaps it is causing the entertaining of beliefs about dogs, or it is causing avoidance behaviour due to a fear of dogs. If the representation as it is being consumed and used as a “dog” representation actually fails to map onto the world, then these are not normal conditions. Again, the proper function of the consuming system, along with the normal conditions for its functioning, are what determines the content of a representation.

Summary

In this chapter I have looked at various attempts to provide a naturalistic account of intentionality. Fred Dretske’s informational theoretic account got the ball rolling by making use of natural signs in the environment, and their ability to indicate features other than themselves through causal co-variation.

By appealing to the notion of evolutionarily defined function, the theory was able to extend the notion of indication to representation and its opposite: misrepresentation. This theory is not sufficiently able to handle the disjunction problem, however, and its appeal to optimal conditions is not precise enough to determine the exact nature of representational content.

Fodor's Asymmetric Dependence Theory was specifically designed to combat the disjunction problem, and it does provide us with a nice way to identify the content of a representation when disjuncts are possible. ADT does not provide us with a good enough understanding of the mechanism by which representation can occur, however, and it helps itself to the very intentional idiom that we must avoid. Furthermore, it is not clear that his theory has any chance of working without appeal to the very optimal conditions that Fodor attempted to discard.

Biosemantics, I believe, is able to solve the disjunction problem. By splitting the representation system into producers and consumers, the theory can determine the content of a representation when there is a possible disjunct. Millikan's formulation of both proper functions and normal conditions in terms of history, and only in terms of history, is something I would like to dispute in the following sections of this thesis. However, biosemantics appears to be successful as an account of how representing systems such as human beings came into being; and it is likewise successful as an account of intentionality that starts from a purely physicalist basis and builds up from there.

In the next section I will look in detail at a thought experiment that is particularly problematic for any teleological theory of mental content. The Swampman objection puts pressure on any theory that appeals to the notion of prior evolutionary or learning history in providing a basis for physical intentionality. In describing the objection in detail I hope to bring to light the ways in which it challenges such theories, and then in chapter three I intend to explore ways in which the biosemantic account can be understood so as to support our intuitions about the Swampman case while still providing a physicalistic account of intentionality.

Chapter Two

Swampman

Having looked at the most promising teleological theories, it is now time to raise an important objection. This thought experiment, also known as the Swampman objection, is so effective against any theory that relies upon the prior history of an organism precisely because it presents a seemingly conscious being while subtracting the possibility of any history whatsoever from the picture. In this chapter I will describe the thought experiment in some detail, and present the responses that teleosemanticists have tended to provide. I will argue that while the thought experiment seems to force any such theory towards the non-intuitive position of denying Swampman any representational states, we should be wary of such a move, since doing so will have consequences for how we think of conscious states in general, given the physicalist assumptions adopted in this thesis.

2.1 Davidson's Double

The Swampman objection was first described by Donald Davidson in his paper *Knowing One's Own Mind.*[10] Davidson actually only discusses Swampman very briefly in the context of understanding content externalism

10 Davidson, D. (1987). Knowing one's own mind. In *Proceedings and addresses of the American Philosophical Association.* 441-458.

and first- and third-person knowledge of mental states. Despite this very short discussion, Swampman has gone on to be a particularly sturdy stick with which to beat the teleosemanticist, and the features that make this the case are all present in Davidson's brief description.

The thought experiment runs as follows: One day, while Davidson is going for a walk by a nearby swamp, lightning strikes a tree beside him, reducing him to ashes while simultaneously creating a molecule-for-molecule Davidson replica. This replica, the Swampman, looks and acts just as Davidson would: he drives home to 'his' house, has conversations with 'his' wife, writes philosophy exactly as Davidson would, and so on. Swampman is indistinguishable from Davidson (at least, at the moment he appears, anyway) down to the microscopic level. From the outside, at least, Swampman is identical to Davidson; there is no difference between the two. However, as Davidson writes,

> "But there *is* a difference. My replica can't recognize my friends; it can't *re*cognize anything, since it never cognized anything in the first place. It can't know my friends' names (though of course it seems to), it can't remember my house. It can't mean what I do by the word "house", for example, since the sound "house" it makes was not learned in a context that would give it the right meaning – or any meaning at all. Indeed, I don't see how my replica can be said to mean anything by the sounds it makes, nor have any thoughts."[11]

11 Ibid., p. 531. (Emphasis Davidson's).

The problem for Swampman is that he just doesn't have the causal history that is supposedly necessary for his thoughts to refer to anything; Swampman can't mean anything by his utterances and can't refer to anything with his thoughts. This, of course, is the problem of intentionality resurfacing for any physicalist theory that relies upon causal history: Swampman sure *seems* conscious, *seems* like he is be representing the environment as he moves through it, *seems* like his actions are driven by the usual beliefs and desires that motivate us all – after all, if Swampman is successful in heading home, the best explanation that we could offer is that he believes that it is *his* home and desires that he get there.

Witnessing Swampman negotiating his environment produces the intuitive feeling that he must be conscious, must be representing his environment, and must have intentional states. Adopting a reductive physicalist viewpoint, if Swampman is physically exactly the same as Davidson, then they should share all of their properties – including those that are mental. Since Swampman *is* identical to Davidson, I want to hold on to the intuition that his mental content is identical to that of Davidson (at least in the first moment). But this intuition is exactly what the teleosemantic view should reject. If intentional states are a matter of evolutionarily derived functions and the appropriate conditions under which they occur, then the bare minimum we should require for intentionality is an evolutionary history that produces those functions and establish those conditions. Swampman lacks exactly that history, and as a result we should probably reject our intuitions. In the next few sections I will describe how Dretske and Millikan have done just that. I will

then go on to provide reasons for why I don't think this is the best route to take if we are to maintain a physicalist outlook, and then in chapter three I will outline ways in which we can save both our intuitions about Swampman (provided you share the intuition that he does have intentional states) and the teleosemantic explanation of intentionality in actual cases.

2.2 Dretske, Swampman, and Twin Tercel

Dretske, in addressing the Swampman thought experiment, provides a version of his own that is more in keeping with his prior discussion of representational artefacts. Dretske imagines that lightning has struck a junkyard and produced a molecule-for-molecule replica of his car, a Toyota Tercel. Except, in the case of 'Twin Tercel', the petrol gauge is broken. Consequently, Dretske asks, is it really *broken?* Given that a petrol gauge's ability to represent is a so-called 'derived' representation, in that it relies for its status as a representational system upon its designers, how can we comment on whether or not Twin Tercel's gauge is producing misrepresentations (as it would be if it were truly broken)? Dretske argues that there is just no way for us to comment on this, since the fuel gauge does not have a function in the first place; the Twin Tercel's petrol gauge does not have a function precisely because it does not have a designer to attribute to it any functionality.

Now, Dretske is arguing for his own representational theory of conscious experience – that the 'what it is like' of all of our conscious experience is representational in nature. As a result, he sees cases such as Twin Tercel and Swampman as providing an internalist challenge to his larger project. I

am not arguing here for or against representationalism, but what is interesting for the current discussion is that Dretske bites the bullet on Swampman in the same way as he does for Twin Tercel.

Dretske's Twin Tercel case is designed to show how unreliable our intuitions can be when considering these sorts of fantastical cases, particularly when features such as resemblance and spatial placement are kept constant. As he writes,

> "We are, for instance, influenced by a striking resemblance in appearance and placement of parts. Yet, no one thinks that because my doorstop looks like your paperweight, and happens to be placed on papers (thus holding them down), that it *is*, therefore, a paperweight… Yet, when asked to render judgments about more complex objects—automobiles, for example—we blithely ignore the fact that the resemblance in both appearance and placement is (by hypothesis) completely fortuitous and, thus, irrelevant to determining the function of parts."[12]

According to Dretske, this is exactly what is happening with Swampman. Our intuitions are skewed because the placement and resemblance of the parts that make up Swampman are identical to those in Davidson. But the resemblance between the two is merely fortuitous, and therefore irrelevant to their actual functionality. Since Dretske's theory requires either a designer or

12 Dretske, F. I. (1997). *Naturalizing the mind*. MIT Press. p. 146

an evolutionary history to provide something with definite functions, Swampman cannot have intentional states, no matter how much his parts are identical to Davidson's. Intentional states require structures that have the evolved function of representing, and Swampman lacks the history necessary for providing those functions.

Here, I think, Dretske gets it wrong in terms of the difference between something having a definite function, and that thing functioning as something in virtue of the processes it facilitates. As I briefly mentioned in chapter two, if we were to see a freakish heart that had no causal history, but that heart was in actual fact in the process of pumping blood, keeping its host alive, then we would be strongly inclined to attribute function to it in virtue of the way it appears to be functioning. In the case of the doorstop acting as a paperweight, mentioned by Dretske above: no, it is not a paperweight. But it certainly is functioning as one, given that its sturdiness and weight are responsible for holding the papers down and stopping them from blowing away. As for the Twin Tercel's petrol gauge: no, it does not have a function. But it also isn't *functioning*, since it is not connected in the appropriate way to the petrol tank. If we were to start the engine and look at the pistons firing away, we might be confused about attributing to them the function of moving the crankshaft, since they had no designer to attribute that function to them, but we would be strongly inclined to say that they are functioning in that way when they actually move.

The placement of parts is what makes it the case that something undergoes or facilitates the physical processes that it does. If the placement of Twin

Tercel's pistons is such that they will move in response to fuel and air being pumped into the engine, then the placement of those parts *really does* make a difference in the engine's functioning. Dretske thinks that our intuitions are skewed because we neglect to consider what it takes for something to be a petrol gauge: assigned function. But the Twin Tercel's gauge is by stipulation not functioning in any way: once we start the Twin Tercel, the gauge does not move, while the pistons do. And the explanation for this is to do with the placement of the parts, not the lack of a designer.

Furthermore, there is an important difference between the Twin Tercel case and Swampman. Swampman is, again by stipulation, negotiating his environment just as Davidson would have done; and the very systems that allow him to do this – his perceptual apparatus, his motor control systems, and so on – are the systems that most physicalists would claim are responsible for his ability to represent. Our intuitions in the Swampman case rely upon the fact that Swampman is surviving in his environment, and to do so as convincingly as he is able to intuitively seems to be a result of his representing that environment such as to be able to negotiate it effectively. If the Twin Tercel was driving along the road like the normal Tercel normally does (which, presumably, it would be perfectly able to do), would we look at the properly functioning engine and refuse to attribute the pistons the function of pumping fuel despite it clearly doing so? Again, I would argue that there is a distinction to be made between a thing having an assigned function, and a thing functioning in a certain way in virtue of its physical makeup and the processes thereby possible.

Dretske thinks that we should reject what he calls the 'internalist intuition' when it comes to Swampman. He claims that our intuitions are pushed in the wrong way as a result of the resemblance between Swampman and Davidson, and their identical physical structures. Without prior history, Swampman just cannot represent because he lacks any structure with the assigned function to do so. We cannot trust our intuitions on this point, Dretske thinks, so even though it seems very much like Swampman will have similar mental states as Davidson, this intuition is best ignored. As I have explained, I do not think that this is a conclusion we need to accept. Next I will look at Ruth Millikan's response to Swampman, in which she bites the same bullet as Dretske, but for different reasons.

2.3 Biosemantics and Swampkinds

As with Dretske's account of intentionality, Millikan's biosemantic account predicts that Swampman will have no intentional states. Like Dretske, Millikan argues that Swampman cannot have the functions and conditions that are essential to her account of intentionality, since he lacks the requisite causal history. In chapter three I aim to give an account of conditions and functions that can apply to Swampman, despite Dretske and Millikan's misgivings.

According to Millikan, not only does Swampman lack the required history to possess intentional states, but he is also of a completely different biological kind from human beings. Intentional language such as *belief* and *desire* operates over human beings, and the mental states in question – the actual beliefs and desires – are formed in the same biological way that human

beings are. Since Swampman is not formed in this same way, and is not of the same kind as humans, then it is wrong to apply these concepts to him.[13]

In discussing the difference between Swampman and human beings, Millikan makes the distinction between 'natural kinds' and 'real kinds'. Natural kinds are those of the sort proposed by Putnam, such as water.[14] Natural kinds are "classes over which strict laws can be run."[15] Their nature is constituted by microphysical structure. Biological organisms – species – are not of this nature; they cannot be the subjects of strict natural laws. But they are 'real kinds': "*Real kinds* I define as groups over which a variety of relatively reliable inductions can successfully be run *not accidentally but for good reason."*[16] Among the subjects of these possible inductions are included all psychological theorizing, including intentional states. Swampman, however, is not of the same real kind as the rest of us human beings. Real kinds are individuated, at least in part, by their ontogeny – the development of the organism since birth – and phylogeny – the evolutionary history of the organism. Swampman and members of the human species clearly have very different ontogenies and phylogenies. If Swampman is not of the same real kind as human beings then, Millikan argues, we have no business in attributing to him our psychological states.

13 Millikan, R. G. (1996). On swampkinds. *Mind & Language*, *11*(1), 103-117.
14 Putnam, Hilary (1975). The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science*, 7, 131-193.
15 Millikan, R. G. (1996). On swampkinds. *Mind & Language*, *11*(1), 103-117. p. 107. Emphasis mine.
16 Ibid., p. 108.

But, as Millikan herself points out, Swampman is a member of the same real kind as Davidson, at least briefly.[17]

> "…Swampman's outer nature, his more superficial and easily observable nature, is like Davidson's too. It would be possible to run successful inferences from any of the superficial properties of Davidson to parallel properties for Swampman, and *vice versa,* and all this is for a very good reason."[18]

What are these superficial properties? Presumably they include the properties that we, as third-party observers, can see. And these will include all of the micro-physical properties of the pair's brains and bodies: all of the causally efficacious goings-on in Davidson's and Swampman's brains will be identical and potentially subject to the same inferences, for the first moment that Swampman appears, at least.

But let's suppose that Davidson was to survive the lightning strike. Suppose, as well, that rather than creating Swampman here on Earth, Swampman is instead created on a Twin Earth that is identical to this one (including the constitution of its watery stuff, its history, its micro-physical constitution – everything is identical), killing Twin Earth Davidson in the process. If Swampman materialized at the exact same spot as Davidson such that both Swampman (on Twin Earth) and Davidson (on Earth) get the same sensory input as one another at exactly the same time – their dispositions are lined up

17 For Millikan, being a member of a real kind is not transitive; although Davidson and Swampman are (briefly) members of the same kind, this does not mean that other human beings are of the same real kind as Swampman, even if they share this with Davidson.

18 Ibid., p. 108. (Emphasis in original).

and responding to the same inputs and outputs – then, since the two are micro-physically identical in a micro-physically identical world, their actions would line up identically also. After all, Davidson stipulated that Swampman does go home and continue work on his philosophy. So in our Twin Earth example we should expect that Davidson and Swampman both go home in identical ways and work on their philosophy in identical ways, in virtue of them having the exact same micro-physical constitutions, subject to identical laws and identical environments.

Such a story gains support from Millikan when she says "Davidson and Swampman are (very likely) the only members there are of a certain Putnam-style natural kind, defined by possession of a certain very exact inner constitution."[19] If this is the case, then the same laws will apply to them, such that if they were in identical situations (Earth and Twin Earth, in this example) then the same predictions could be made about their behaviour.

Just as before, the question is whether or not we can attribute intentional states to Swampman. And as a physicalist, I think that we have good reason to do so, or else run the risk of condemning Davidson's intentional states to the status of mere epiphenomena. The reason I think this is due to the causal exclusion argument.[20] According to this argument, if there is a sufficient physical cause P for a physical effect *P**, any supervenient mental cause *M* will be epiphenomenal. This is because *P*'s sufficiency for *P** means that *M* will be an over-determining cause of *P**. If you subscribe to the view that

19 Ibid., p.109.
20 Kim, J. (1998). Mind in a Physical World, Cambridge: Cambridge University Press. The causal exclusion argument is not unanimously agreed upon within philosophy. I am assuming its truth here, in line with the reductive physicalism I adopt in this thesis.

having more than one sufficient cause for a physical effect makes at least one of these causes an over-determining cause, and the universe does not feature systematic over-determination in this way, then the causal exclusion argument has some force.

The causal exclusion argument can be applied to our Swampman case involving Twin Earth. If, as Millikan would have it, Swampman has no intentional states – since intentional states must supervene not only on the physical constitution, but also on history and other extrinsic relations – and if Swampman, on Twin Earth, is subject to the same causal processes as Davidson on Earth, then all of Swampman's behavioural effects can be explained purely through the physical causes from Swampman's internal constitution. But since Davidson will also undergo the exact same processes as Swampman does, only on Earth, then Davidson's internal physical constitution serves as sufficient case for all of his behavioural effects. This means that the intentional states that Davidson has but Swampman lacks are causally over-determining Davidson's behaviours. If we take the causal exclusion argument seriously, then this is a real problem for the view that Swampman has no intentional states.

2.4 Is Swampman Even A Possibility?

Setting aside whether Swampman is the sort of being to which we can attribute mental states – which we had better do, since Swampman has remained a live objection despite Millikan's counter-arguments – there is also the question as to whether Swampman is even possible. Millikan presents a

couple of amusing arguments to the effect that Swampman would not be possible from a biologist's or a physicist's perspective – at least, not without strange results that would undermine the whole point of the thought experiment.

If Millikan is correct, and Swampman is impossible, it makes little difference to the arguments I have been, and will be, making here – in fact, I would welcome the result. This might seem like an odd statement – most of this thesis is concerned with Swampman, after all. But my purpose here is to show that, were Swampman real (and thus logically possible), he would pose no particular problem for the physicalist. If Swampman was real then I argue that we must attribute to him intentional states; and if we saw him moving around, negotiating the environment, and writing philosophy, then there is a clear way in which we could explain this. But these are all *ifs*. If Millikan is correct, then there is no need for the arguments I am making. But, since Swampman continues to be thought of as an effective argument against the best physicalist theories in town – teleosemantics – and if this requires that Swampman be at least logically possible, then the arguments I am making in this thesis can be used against those who continue to use Swampman as an objection regardless of Millikan's conclusions vis-à-vis Swampman's status as a possible entity.

This applies to other philosophers who advise against the use of Swampman in our theorizing. Dennett, for example, ridicules such thought experiments on the grounds that they throw our intuitions in wildly implausible directions. He thinks it ridiculous that we try to isolate one factor of a theory in such a way –

in this case, the representing organism's causal and evolutionary history – and argues that the results of such thought experiments, where the thought experiments deviate so fantastically from normal states of affairs, will always lead our intuitions in the wrong direction.[21] Neander, similarly, argues that we should not allow our intuitions to be swayed by Swampman, given its ludicrous improbability.[22] These are sentiments that I am sympathetic towards. But rather than adopt the tactic of these authors and ridiculing Swampman (something that is fairly justified), in this thesis I am instead electing to take Swampman seriously. I wish to argue that *if* Swampman were logically possible, and *if* he was able to act as Davidson first described him (negotiating the swamp and going home to his family, etc.), then we should ascribe to him intentional states; and these intentional states can be explained from within a physicalist framework without recourse to the intentional idiom. As I remarked at the beginning of this section, Swampman remains a live objection despite protestations from those that find the thought experiment ludicrous. But I think that the original objection can be answered using a similar tactic to that of Millikan and Dretske.

2.5 A Non-Fictional Swamp-Creature

Perhaps there is a way to maintain the force of the Swampman objection while avoiding the charge of irrelevance due to it being an entirely fictional example. The argument runs as follows:[23] presumably, for any evolved trait

21 Dennett, D. (1996). Cow* Sharks, Magnets, and Swampman. *Mind & Language*,*11*(1), 76-77.
22 Neander, Karen (1996). Swampman meets swampcow. *Mind and Language,* 11 (1), 118-29.
23 Peters, U. (2013). Teleosemantics, Swampman, and representationalism. *Grazer Philosophische Studien.*

there was the first mutation that formed the basis of that trait. The same should hold for whatever trait it is that confers the ability to represent on an organism. Presumably, as with all traits, the ability to represent would evolve gradually; but however gradually it does appear, there should be some point in its evolutionary history at which a bare minimum amount of representation begins to occur. Call this first organism with the ability to represent very basic visual features C1. C1 will be able to represent – according to the teleological accounts of content, anyway – through some form of causal co-variation between one of its features and features of the environment. But the function of representing in C1 does not have a history of natural selection. It is the first time this function has been in play in evolutionary history.

At this point it is easy to just say that C1 does not have the ability to represent, since the trait(s) that account for its 'awareness' of the environment have not been subject to the pressures of natural selection. But if that trait is useful for C1's survival, and is responsible in part for C1's producing offspring, then, presumably, its offspring one, two, three generations down the line will have the ability to represent, since the ability to indicate features of the environment will have become functions to represent the environment. Let's call the first offspring that is actually able to represent, C2. C2 will be similar to C1 in terms of its inner constitution in all of the relevant ways, since it will have the same physical arrangement in its primitive visual system. What, then – other than the fact that a generation gap separates them – is the difference between C1 and C2?

Here we seem to have a case of two organisms that are relevantly similar in terms of their inner constitutions, and that demonstrate the same behavioural capacities (remember that it was C1's ability to indicate the environment that allowed it to reproduce), yet in one the trait that is responsible for its visual interaction with the environment has a prior evolutionary history, and in the other it does not. Obviously this is a terribly crude picture that is most probably wrong in terms of a biological explanation. But it is not logically impossible, as Dennett claims Swampman is. While this is certainly far from a solid argument against any of the above anti-Swampman positions, I think it does serve to give us good reason to take the Swampman-style objection seriously if we are to develop a proper physicalist picture of intentionality.

Summary

Davidson's original Swampman thought experiment has had a large impact on all teleological theories of mental content. The central thought serves to apply pressure to those theories of content that seek to explain intentionality by recourse to the notion of evolutionarily- or learning-derived functions, as well as the notion of optimal or normal conditions.

Dretske's account of intentionality leads him to bite the bullet and deny Swampman intentional states. He attempted to show how our intuitions could be lead astray by Swampman by providing his own, similar example of the Twin Tercel; an exact replica of his car, but with a 'broken' petrol gauge. It is obvious to Dretske that this petrol gauge can't *really* be broken, since it lacks any designers or past users that could have bestowed functions upon it –

much like Swampman and his supposed ability to represent his environment, form beliefs, desires and enter into other intentional states. As I have argued here, however, I do not think the parallel works well enough to discount the possibility of Swampman being able to represent the environment, and does not take into account the difference between an assigned function and a process that is acting like a function such that it is causally indistinguishable from an assigned function.

Millikan's Biosemantics is similarly forced to bite the bullet and deny representational abilities to Swampman. Furthermore, Millikan argues that Swampman should not even be considered as a candidate for intentional states, since he is not a member of the appropriate kind. I have argued that since Swampman is a member of the same natural kind as Davidson, at least in the moment that he is created, then it is possible to create a Twin Earth-style thought experiment in which attributing intentional states to Davidson but not Swampman results in a problem for the efficacy of Davidson's mental states.

I addressed the possibility that Swampman is not a candidate for empirical theorizing anyway, since he is nothing more than a fiction. While I have sympathies for this view, I think that trying to find a plausible answer to Davidson's original thought experiment is a task that should be undertaken, if only to provide a response to those that still raise Swampman as a viable objection to teleosemantic theories of mental content. Finally, I provided a plausible alternative version of the Swampman objection that seems possible

in a way that certain philosophers think that Swampman is not, in order to support this conviction.

In the next section I will provide a possible solution to the Swampman objection. I intend to do this by adapting the notions in use by Millikan of proper function and normal conditions. By doing this I hope to show that we can hold on to the most promising teleological account of mental content for explaining actual cases of evolved organisms, while respecting the strong intuition that, since Swampman is acting like Davidson, he is representing like Davidson.

Chapter Three

Saving Swampman's Mind

As discussed in chapter one, I think that biosemantics provides the best real-world account of intentionality, and is particularly well placed among the teleosemantic theories to deal with the disjunction problem. So, in the spirit of biosemantics, I propose that Swampman's mental content will be fixed by what his representations – as put to use by the consumer device in accordance with its proper function – map to under normal conditions. Given that Swampman lacks the evolutionary history that biosemantics requires, this chapter will largely be concerned with re-conceptualising the notions of proper function and normal conditions.

3.1 A Brief Note on Dispositions

In what follows I will be assuming that dispositions are real states of the objects that possess them. At least some of the properties of matter are dispositional, with a great many potential ways in which they can be brought to manifest. By looking inside Swampman and noting that the states of his physical make-up are capable of bringing about various manifestations given the appropriate stimulus, we are in a good position to see how Biosemantics is still able to get a hold on Swampman in terms of both the proper functions

of his mental states, and the normal conditions under which they optimally operate.

3.2 Swamp functions

Swampman is a molecule-for-molecule replica of Davidson. As such, his physical matter will be identical, and the properties thereof will likewise be identical. This all means that his dispositions will also be identical. The receptors in his eyes will react in the same way to light, causing a chain of cause and effect down his optic nerve, to his visual area, and on to the different stages of cognitive processing (or what would be cognitive processing in Davidson's case). All of this will result in the same reactions from Swampman to, for instance, a tree nearly falling on him in the swamp or any number of sensory cues. Of course, I want to argue that Swampman moves out from under a falling tree because some of his mental states represent the incoming visual signal as a falling tree. The question is how he can represent as so described. To adopt the Biosemantic account of intentionality, Swampman has no evolutionarily selected function to act as the consumer of a representation, and for the same reason there are no normal conditions under which he can use representations correctly.

The proper functions of a representing organism are a result of natural selection: it is this selection that explains how the processes in the organism came to be functions. But to speak of functions under a purely reductive physicalist picture is to speak elliptically: there are no 'real' functions in nature (much like there are no 'real purposes), only processes that consistently

produce the same effects from certain inputs. Functions are just processes that continue to exist due to the beneficial effects they produce. In the case of Swampman we have the exact same processing occurring had Davidson been in his place. This is because Swampman has all of the same physical dispositions as Davidson. Dispositions are inherently forward facing; they do not require their history in order to manifest – even if, under non-fantastical circumstances, that very history is the explanation for their existence. The processes that begin with light hitting Swampman's eye and result in him behaving in some specific way are still functioning in the same way as they would have in Davidson, we just do not have an evolutionary justification for calling the processes that make up this behaviour proper functions. What we have is the functioning of physical matter in a certain way, a way that is constrained by the physical make-up of Swampman.

Swampman will still have 'producer' and 'consumer' parts of his representational system required by biosemantics. An example of a producer is the photoreceptor cells in his retina: given his dispositions these will still react to light and send a signal upstream. As for the consumers of representations – perhaps his motor system, or his belief-forming mechanisms – these will still physically and causally operate in the same way as they would in Davidson – there will be the same functioning. An element essential for the Biosemantic view is the relation between the representation and the represented; unlike the informational view of Dretske, which sees this relation as a dyadic one, Biosemantics instead takes the relation to be a (at least) triadic relation between the represented, the producer of the representation, and the consumer. Swampman, in virtue of his functioning

representational system – identical processes occurring in virtue of its dispositional makeup being identical to Davidson – will have all three of these components in play as he interacts with the world.

At this point a supporter of Biosemantics could point out that these components are not really producers and consumers, since they cannot have the function to be so, lacking the requisite history of selection. Just like with Dretske's Twin Tercel from earlier, Swampman's 'producers' and 'consumers' cannot be said to have functions. But I would argue that, due to Swampman's dispositional, physical makeup, the same matter that makes up Davidson's producer and consumer devices is functioning in Swampman. If Dretske's Twin Tercel's petrol gauge did happen to be wired up correctly (due entirely to the cosmically unlikely way it was put together), and it did properly co-vary with the amount of fuel in the tank, then we would be right to say that it was functioning as a petrol gauge, representing the amount of petrol in the tank. Of course, this requires us as the driver – the consumer of the representation – to treat the output of the petrol gauge as a representation of the amount of fuel in the tank. If, after a long drive into the desert, the gauge's connections came loose and it stopped indicating the amount of fuel in the tank, then, after having become used to it correctly representing the fuel, we could reasonably claim that the petrol gauge is no longer functioning properly – it has begun to misrepresent.

Here it can be objected that if the gauge had not been displaying the fuel amount correctly from the outset, and eventually began to represent correctly, then we would have no way of saying whether it was functioning correctly at

the start of the journey or later; whether it was ever representing or misrepresenting. Actually we can, but only because the driver is treating the gauge as working in a certain way. The driver can get far into his journey, notice the gauge's lack of movement, and then quite legitimately claim that it had always been broken, that it had never functioned correctly. Of course it had never functioned the way the driver had thought, so strictly speaking it had never malfunctioned. But the driver had treated it as if it were functioning in a certain manner, and as such the gauge attained a derived functionality that it then failed to serve.

Treated as one entity, the gauge-driver complex's continued 'survival' (i.e. both working together to continue the journey) requires that the consumer device treat the produced representations as if they were produced according to a mapping rule between the representations and the represented (more on mapping rules in the Swamp Conditions section below). The processes that constitute the fuel gauge's operation either benefit the gauge-driver complex or do not. If they do not, then this can be seen as a malfunction. This is the same for Swampman and his producer-consumer complex: the processes within will either contribute to survival or not. The fact that these processes have no selection history does not change the fact that they are either doing something useful, or they are not. A key point here is one of continued survival: given the local conditions of the world, there are clearly beneficial ways in which Swampman's producer and consumer devices will operate, via the same mindless processes that operate in Davidson. Functions are processes that reoccur because they are beneficial – fitness enhancing in the case of organisms – and identical processes will be identically fitness

enhancing regardless of their etiology. Given time, as Swampman begins to negotiate his swamp and eventually makes it home to work on 'his' philosophy, it will become more and more obvious whether his inner processes are beneficial or otherwise. If Swampman continues to survive – which, according to the original thought experiment he will – then we will clearly see that those processes, which in Davidson were obviously functions thanks to their evolutionary heritage, are also functioning in Swampman. The processes are functioning to keep him alive, and if they do otherwise then they are malfunctioning.

The point is one of functioning versus proper functions. Swampman may not have proper functions as Millikan defines them, because proper functions defined in this way require a long evolutionary history. But due to the physical capacities of Swampman and his dispositional makeup, the parts of him that in Davidson we would have called a visual system will be capable of functioning as a visual system in Swampman. Physical processes, with their various inputs and outputs, will continue to occur regardless of how they got there and how we label them. For example, if an alien visitor from a planet without wind happened across a wind-vane here on Earth, decided to make a copy of it, and stuck it on a roof, the alien's lack of treating it as a wind-vane would not alter the fact that this newly fashioned piece of matter is pointing in the direction of the wind. The alien wind-vane has no history, other than being made by a designer with no function in mind or knowledge of the physical processes of which the wind-vane will be capable. Historical explanations can tell us how an arrangement of matter came to function in a certain way, but a lack of such an explanation does not stop that matter from doing so.

Looking at the reasons for requiring that functions have a history of selection, we should ask what is the difference between first-generation tokens of an adaptation, and those type-identical ancestors that follow. If an organism's adaptiveness – its evolutionary fitness – is thanks to its physical makeup plus the appropriateness of this to its current environment, then there should be no difference between a first-generation trait that has occurred due to accidental mutation, and a seventh-generation trait that has proven adaptive in the face of natural selection. The only difference is that the seventh-generation token has type-identical ancestors, whereas the first does not; and having type-identical ancestors should not make a causally relevant difference.

But there are still problems with this account of functions. As Millikan writes,

> "Imagine a physiologist trying to study the liver or the eye without having any idea what its proper functions are-what it is supposed to do. Clearly his first job will be to try to find out what it is supposed to do, what it is for it to "work." Until he has formed some kind of hypothesis about this there is no way of proceeding to a study of how it works. There is no way of knowing even when it is working, let alone working right or well, and no way of distinguishing the Normally constituted and properly functioning samples of its kind from those that are malformed, diseased, or malfunctioning. Nor is there any way of proceeding to a study of how it works without knowing

> something about the surrounding conditions upon which it normally relies."[24]

Back to Swampman and Davidson. What if, due to a small difference between the two, Swampman's visual system is such that it appeared to be malfunctioning compared to Davidson? Perhaps, where Davidson would have seen a tree falling – putting him in a particular internal state – Swampman, in the exact same position, with the exact same inputs to his visual system, would be in a completely different mental state, one that we would not associate (in Davidson) with Davidson's seeing a tree or Swampman's seeming to see a tree. In Davidson such a difference could be described as a malfunction, but in the case of Swampman it is not so clear; this different internal state is just a result of functioning physical processes, the response of his visual system to the environment. The problem is that we, as observers, would only be licensed to 'assign' functions (i.e. label them as such from our third-person POV) based on the fact that Swampman bears a striking resemblance to Davidson. Strictly speaking, there would be no way for us to know that a certain part of him has the function of producing or consuming representations in certain ways, other than by reference to Davidson. In the case of possible differences in representing the falling tree, should this part of Swampman's retina have the proper function of producing representations to be read by the consumer device as a falling tree, or a standing one?

Without a prior history to separate proper functions – those useful for survival – from just any old functioning, Swampman could not misrepresent; his

24 Millikan, Ruth G. (1986). Thoughts without laws: Cognitive science with content. *Philosophical Review*, 95, 47-80. P.56.

representational system could not malfunction (if it has no proper function, then who is to say what a malfunction looks like?). If we were to watch him moving around his environment and were able to know the precise nature of his mental content, then we would have to say - no matter what that content is and how it correlates with the environment – that the processes that make up his representational system are just occurring, that they are functioning in some way. The best we can say about Swampman is that he can function in a near-limitless number of ways; whichever way his physical processes run, there will be functioning (of a sort). Of course, the processes that occur in Swampman will be pretty similar to those that would have occurred in Davidson (at least for a short while), so there will be the same processes occurring. But the difference between the two is that the processes in Davidson can be said to be functioning or malfunctioning, based upon his evolutionary history. In Swampman, the physical processes that occur *just are* functioning, with no way for them to malfunction, at least that we can make sense of.

This is obviously a problem for this thesis, since we want to be able to say that Swampman representing a falling tree as a standing one is a clear case of misrepresentation. The solution to this problem is to note, as I have been arguing, that the physical processes that constitute the proper functions of a producer and consumer device in Davidson will be occurring identically in Swampman. To see this more clearly, imagine that Swampman successfully manages to reproduce (this is not such a strange thought; after all, in the thought experiment Swampman is supposed to be indistinguishable from Davidson to family and friends). Does Swampman's progeny have intentional

states? Presumably the processes that are responsible for Swampman's avoidance of the falling tree that first day in the swamp contributed to his survival, proving to be an adaptive trait; have these processes yet become functions?

Perhaps it would take a few generations of swamp-children before we could reasonably label his inner processes as having definite functions. But with hindsight we might wonder about Swampman's inner processes. If a team of experts were to have a detailed understanding of Swampman's inner workings and behaviours, along with those of his children, it seems at least reasonable to assume that these experts would be able to, with hindsight, label certain processes in Swampman – those that contributed to his survival – as having certain functions. With such a benefit of hindsight, including being able to view Swampman's behaviours in the context of his offspring, it no longer is so clear-cut that Swampman does not have brain states (and other states) that can truly be said to have functions, even if these are functions (or functioning processes) that do not have the unnecessary etiological constraints imposed by Millikan.

The above is not to suggest that each of Swampman's processes require an evolutionary future in order for them to operate in ways that are beneficial to his survival and hence for them to be properly labelled as functions (although it requires that Swampman has a future in which his processes can function). Rather, I want to point out that the processes in the Swampman that goes on to reproduce will be type-identical to the same processes in the non-reproducing Swampman when he e.g. looks at a tree in a swamp; and if we

can with hindsight label the reproducing Swampman's processes as proper functions, and these are identical to the non-reproducing Swampman's processes in terms of their constituents and effects, then there is less of a reason to label one set of processes as functions and the other as not.

Since proper functions are those selected-for processes that contribute to the continued survival of an organism, then if Swampman's relevant inner processes (i.e. those constituting his producer and consumer devices) contribute to his successful negotiation of his environment in line with the way we would expect Davidson to negotiate his environment, we should continue to view these as functions in Swampman. The only difference is the lack of being selected for, but the processes are doing the same job; they are working with the same constituents and produce the same outputs. As such there seems little reason to deny that they are functioning in a certain manner and according to certain constraints (i.e. their contribution to Swampman's continued survival).

If we could observe Swampman's inner workings as he negotiated the swamp, we might look in his chest and see his swamp heart. Watching the processes occurring in this chunk of meat, processes that are responsible for the blood being pumped around his body, we would be justified in calling these functional; and if the processes stopped doing what they are doing, we would be justified in saying that they are malfunctioning. The same goes for his visuomotor system: as we watch Swampman negotiating his environment we would quickly see that his visuomotor processes are functioning to help with this task; and if they stopped functioning in this way and Swampman did

not avoid a falling tree, we would be justified in saying that they are malfunctioning. This will take time for us to be sure that this is the case, but after a while it will become more and more clear what the processes that make up Swampman are functioning for, and, depending on his level of success in negotiating his environment, whether they are ever malfunctioning.

Having time for Swampman's processes to occur is essential for us to know how they normally function. If when Swampman first appears he experiences a hallucination of a falling tree in the swamp instead of a veridical experience, then we will not be able to tell if this is a result of his properly functioning visual system or if it is a malfunction. Furthermore, if Swampman's hallucination leads Swampman to step out of the way of an actual falling tree (one that is visually inaccessible to him thanks to his hallucination), then the internal process responsible for the hallucination, it seems, actually contributed to his survival. So should we say that this process has the function of representing a non-existing tree? I would argue not. In Davidson, such a hallucination would be a malfunction brought about by his functioning representational system operating under non-normal conditions. When Swampman first appears there are no normal conditions under which his internal processes 'should' operate. But given time there will be. Given time Swampman will continue to successfully negotiate the swamp, and doing so will require that he is in general not hallucinating; what will be required is that his processes contribute to his survival under normal conditions.[25]

25 More on what it takes for Swampman to have normal conditions in the next section.

Given that Swampman does successfully negotiate his environment, there will be constantly increasing examples of his representational systems operating in ways that contribute to his continued survival success. As these build up it will become increasingly clear what his internal processes do and – importantly – what they are *for*. Once we see this in Swampman (which will require some time for the number of 'uses' of those processes to build) we will then be able to retroactively look at his initial hallucination and label it a malfunction. It will have the same content as a veridical experience would have had, since these are set by normal conditions, and a hallucination is the result of non-normal conditions. Again, however, time will need to pass during which Swampman can negotiate his environment for those normal conditions to be defined.

Given a way to understand Swampman's functions, we now need a way to account for his normal conditions if we are to explain his intentional states using the biosemantic framework.

3.3 Swamp conditions

What Swampman lacks are the normal conditions under which, as the consumer device uses representations, those representations map correctly to the represented. As a result, in order to provide an account of Swampman's ability to represent and misrepresent, we need to be able to specify normal conditions under which his functioning physical processes will either be successful or unsuccessful in producing and consuming representations as

representations of the represented, thereby facilitating Swampman's survival. With no normal conditions, there is no way to account for misrepresentation and no way of fixing the content of his intentional states.

Normal conditions are those in which the producer and consumer devices function in such a way as to increase their likelihood of being selected for by natural selection. But, strictly speaking, it is a mistake to say that functions are selected *for* via natural selection. Natural selection is selection *against.* As such, normal conditions are those conditions under which consuming a representation in a certain way did not historically result in a decreased probability of reproduction. Swampman clearly will not have normal conditions as described by Millikan when he materialises in that swamp. But there will be pressures on him analogous to those that occur during natural selection. Given the functioning systems that he has for detecting features of the world, there are clear ways in which these systems can go awry. In sharing his dispositional makeup with Davidson, his internal processes are geared towards producing and consuming representations in such a way as to facilitate his survival. After all: if they are that way in Davidson, they will be in Swampman, too.

If the consumer devices of his representational system do not use the representations in the correct way, then Swampman will be at a disadvantage. The local conditions in which Swampman finds himself will be either appropriate for his survival, meaning that he is in some sense adaptive to those conditions, or they will be hostile, in which case they are non-normal conditions and conditions in which his representational system will not

facilitate his survival. If he does not see the falling tree as a tree – if the representations produced by his visual system are not consumed *as* representations of a falling tree – then he will be crushed, and he will die.

As discussed above, given the same dispositional make-up as Davidson, and given the fact that these dispositions do not care about past history (we do, as an explanation, but the nature of his dispositions and their likelihood of manifesting do not), then in terms of Swampman's survival there are definite circumstances in which his representations, as consumed, will map to the represented – and circumstances in which they will not. The point is that *as soon as* Swampman begins interacting with the world, which will be instantaneously, there is a clear sense in which his representational system will be either adaptive to local conditions – i.e. they will be operating under normal conditions – or non-adaptive.

This notion of normal conditions is clearly different to that found in Millikan's writings. Millikan's normal conditions are necessarily a result of natural selection: they are those conditions under which an organism's processes functioned in such a way as to result in evolutionary success. Under the biosemantic account, if we take an organism with a properly functioning representational system and place it in non-normal conditions, then the functions at play in the representational system will malfunction. Under the approach I have discussed here, things are quite different. First of all, there is no way to really say whether a functioning process in Swampman is properly functioning or malfunctioning; it is just functioning (other than by comparing him directly to Davidson, of course). So in one set of conditions, C1,

Swampman will function in one particular way. In another set, C2, he will function in another way.

If we think of Davidson, let's say that C1 are his normal conditions; everything is functioning properly. In C2, on the other hand, let's imagine that, due to a difference in air density causing a difference in the refraction of light, Davidson's depth perception is affected causing him to see objects as being further away than they really are. In C2, Davidson is going to have a hard time avoiding a falling tree in the swamp. This is clearly a non-normal condition. The proper function of his consumer device is to consume representations so as to facilitate his survival. In C2 Davidson's representation of the environment will be off in terms of depth. Since these are non-normal conditions, and Davidson's consumed representations do not map to the environment such that they conform to a rule that adjusts for this refraction effect, Davidson's consumed representations will not map correctly to the represented.

Now back to Swampman. In C1, Swampman's visual system will function in a particular way; in C2, Swampman's visual system will function in a different way. Because Swampman lacks the history of natural selection that defined Davidson's functions, we cannot say that in C2 his visual system is malfunctioning – the best we can say is that it is functioning differently. Of course, being dispositionally identical to Davidson, the processes occurring (the functioning) will not use an appropriate mapping rule just as with Davidson in C2. Swampman will consume representations as-is, and act accordingly. He will fail at avoiding the falling tree, just as Davidson would

have. The important point is that, while we cannot (without some time having passed, and in retrospect) say that this way of functioning is malfunctioning (as we can with Davidson), we can clearly see how in different conditions Swampman will, thanks to his dispositional makeup, function in ways that are either good for his continued survival, or bad for it.

There are conditions under which the functioning of Swampman's representational system will involve his representations matching up with the represented; under which his chances of survival are increased; under which the chances of him being 'selected against' are reduced. As Swampman begins to interact with his environment, two things will become ever more apparent: 1) what his physical processes are actually doing (i.e. what their function is); and 2) whether or not the conditions he finds himself in are beneficial to his continued functioning.

The situations in which Swampman consumes representations appropriately will be situations in which the consumed representations accord to some mapping rule between the representations and the represented, allowing for continued functioning. Likewise, there will be situations in which representations are consumed such that they do not allow for continued functioning, where the mapping rule followed by the producer and consumer does not result in a mapping between the representations and the represented. These are cases of normal and non-normal conditions.

Given the above way of understanding Swampman's situation, I believe that we have a handle on both the proper functions and the normal conditions that

are key to the biosemantic account of intentionality. Although these two features of the biosemantic account are traditionally reliant upon natural selection, I do not think that in the case of Swampman that this is a necessary constraint. For normal organisms there is a reason why they have proper functions and normal conditions: natural selection. The explanation for Swampman's functioning is incredible, but then that is built right into the thought experiment. Swampman will be the same as Davidson in every respect. Davidson's inner processes result in intentional states. Swampman's identical processes will too – otherwise the intentionality in Davidson is not contributing in any causally interesting way to his survival. The difference is that in Swampman's case he will need to interact with his environment for some time before we can determine what his inner processes are doing, what they are for, how they function. They will be doing something, and if that something is contributing to his survival then those processes will be functions operating under normal conditions.

3.4 Swamp Brain

To repeat what was said at the beginning of this chapter: Swampman's mental content will be fixed by what his representations – as put to use by the consumer device in accordance with its proper function – map to under normal conditions. Having established that there are reasons for thinking that Swampman is not importantly different to Davidson in that regard, I would like now to look at another thought experiment – similar in form to Swampman – that produces some strange results given the above formulation.

If content is a matter of distal causes, proper functions, and normal conditions, what happens when we remove the usual distal causes? This time, instead of a person mysteriously appearing with no prior history, we are looking at the philosophically classic brain in a vat appearing with no prior history. The classic brain in a vat thought experiment is usually used in the context of sceptical arguments: how can we know that we are not merely brains in vats with all of our experiences pumped into us via computer simulation by profoundly technologically advanced scientists? While theoretically the fact that we are maybe swamp-brains in vats is perhaps a live possibility, the thought experiment is not being deployed here as a sceptical challenge. Instead, I want to see what swamp-brain in a vat (hereafter SBIV) will be representing, given our definition above.

If my arguments so far have been successful, then we don't need to worry too much about SBIV's producer and consumer devices and their proper functions. As for normal conditions, well, since SBIV's proper functions are those processes that contribute to its continued survival in the relevant ways, and since SBIV is being fed sensory inputs from a computer program in such a way as to mimic the real world (right down to what might harm and outright kill it), then those normal conditions will be conditions under which his properly consumed representations, in accordance with some rule, map faithfully to the represented.

Taking this together, SBIV's mental content will be fixed by what his representations – as put to use by the consumer device in accordance with its proper function – map to under normal conditions. SBIV's input is computer

code, and the distal causes of that input are computer code (imagine the code-as-input being like the light coming into the retina, with the distal cause being the code that simulates the tree that the light bounced from); furthermore, the fact that SBIV needs to survive in this perfectly rendered simulation means that it must avoid 'standing under' falling-tree-code, 'ingesting' poison-code, and so on. As a result, SBIV's consumer devices will be fulfilling their proper function if they use falling-tree-code representations to initiate avoidance behaviour; and the normal conditions will be those under which this is due to there being actual falling-tree-code to avoid. So the mental content of SBIV's intentional state in this case would be tree-code.

This is perhaps not the result that we might intuitively expect. After all, presumably SBIV's mental states will be qualitatively identical to yours or mine, given that the whole point of the brain in a vat experiment is that neither of us would know if we were that brain. But what else could we be representing if we were that brain? The scientists that are running the computer simulation are feeding in electrical signals, but those signals are designed so that they mimic absolutely the inputs and outputs of real-world brains. That the computer signals fed into a brain in a vat create indistinguishable qualitative experience is built right into the thought experiment to begin with, so the fact that representing tree-code is indistinguishable from actually representing a tree is to be expected, even if it seems like an odd result.

Summary

A strong intuition in the case of Swampman is that he would have the same internal life as Davidson. Biosemantics, as originally described, rejects this intuition thanks to its requirement of an organism's evolutionary history. In this chapter I have provided a way to keep biosemantics as a framework for intentionality while holding on to the strong intuition regarding Swampman. I have argued that this is possible if we reconceive of the notions of proper function and normal conditions such that they can be applied to first-generation organisms that have type-identical traits to their representation-capable ancestors. Defining the functions and conditions that contribute to Swampman's survival will require time during which he is able to negotiate his environment successfully. By seeing how the processes that make up his representational systems contribute to his survival, we will be able to see what they are doing and, in facilitating his survival, what they are for.

Once we have an idea of Swampman's functions and normal conditions, we can now return to how Swampman is capable of intentional states. A representational system (in our case Swampman) is able to represent because certain features of his physical make-up have the function to either produce of consume representations. If Swampman operates under normal conditions, then his representations, as used by his consumer device, will map to their represented. However, if Swampman is in non-normal conditions, then the possibility of error is increased. If he consumes representations somehow incorrectly due to these non-normal conditions, then Swampman will misrepresent. Swampman has intentional states because he has the functionality to possess them.

Conclusion

Intentionality has proven to be a difficult problem for any physicalist theory of the mind. The challenge has been to understand aboutness or purpose without using the intentional idiom, to build something capable of intentional thought from nothing but the physical constituents of the universe. It is perhaps not too surprising, then, that the best physicalist theories of intentionality have based themselves on an evolutionary framework, since evolution by natural selection is the only source of such purposiveness that we have found in the universe as described by the natural sciences.

As I have argued in this thesis, the best of these teleological theories is Ruth Millikan's biosemantics. Biosemantics, like other similar theories, makes use of the notions of evolutionarily derived function and the suitable conditions for their operation in order to explain intentionality. By recognising that representations are essentially useful to a representational organism, Millikan's account focuses on the consumers of representations and how they use the mental tokens in question. This approach avoids the disjunction problem as set out by Fodor and gives us a natural way of understanding representations as being part of a system that has evolved to use those representations.

The problem with Biosemantics (like other teleosemantic theories) is that it constrains itself too heavily through insistence on a long history for any representing agent. A reductive physicalism should require that two physically

identical beings be identical in all aspects, including their mental states. In the case of the Swampman thought experiment, because Swampman and Davidson are identical in their physical make-up, then they will share all of their dispositions. These dispositions make it the case that they will react in identical ways to external and internal stimuli, their bodies and minds undergoing identical physical processes. Biosemantics, through its requirement of history, unfortunately produces results in the Swampman case that do not seem acceptable to those who insist that Swampman should have the same internal mental life as Davidson, given his identical behaviour. From the perspective of reductive physicalism, I would argue, this result is unacceptable.

In this thesis I have isolated the key terms in the biosemantic theory of intentionality that give rise to this unacceptable conclusion. By noting that the physical processes that give rise to functionality are type-identical from the first generation to display them (such as the physical processes present in Swampman) to the nth-generation (provided they are the same types of processes), I have argued here that, as a creature operating in the same world with the same selection pressures as any other living organism, Swampman can be seen to share functionality with Davidson. Functionality, under the biosemantic account, results from processes that prove fitness enhancing, contributing to the organism in question's survival and reproductive chances. I have argued in this thesis that two identical processes should be identical in every respect – including their contribution to fitness. We won't recognise Swampman's inner processes as being beneficial to him immediately, but given time it will become more and more obvious what his

inner processes are doing, and what they are for. This purposiveness of his inner processes allows room for intentionality, even in a swamp-creature.

Secondly, by arguing that there are normal conditions under which those processes can operate, we can also have an understanding of the content fixing situations that are essential for the biosemantic account. Again, as with Swampman's functions, these normal conditions will not be immediately apparent when Swampman first appears. But as Swampman begins to interact with the world there will be clearer and clearer ways to categorise the conditions in which he operates as either normal conditions or otherwise.

Taken together, these ways of understanding functions and normal conditions allow us to attribute intentional states to Swampman while still holding on to a physicalist conception of intentionality. Both of these solutions require that Swampman has a future of negotiating his environment, something that is stipulated in the original thought experiment.

Explaining the functionality of our brains, and the resultant intentionality of our minds, requires evolution by natural selection. It is this historical process that sets up the physical make-up of our brains and bodies, with its dispositional nature, and the processes this facilitates. In the fantastical case of Swampman, on the other hand, we have a fantastical explanation: for him, all it took was a lucky strike of lightning.

References

Davidson, D. (1987). Knowing one's own mind. In *Proceedings and addresses of the American Philosophical Association.* 441-458.

Dennett, D. (1996). Cow-Sharks, Magnets, and Swampman. *Mind & Language*,*11*(1), 76-77.

Dretske, Fred (1981). Knowledge and the Flow of Information. MIT Press.

Dretske, F. (1996). Absent qualia. *Mind & Language*, *11*(1), 78-85.

Dretske, F. I. (1997). *Naturalizing the mind*. MIT Press

Fodor, Jerry A. (1990). Information and representation. In Philip P. Hanson (ed.), *Information, Language and Cognition*. University of British Columbia Press.

Fodor, J. A. (1990). *A theory of content and other essays*. The MIT press.

Gibson, M. (1996). Asymmetric dependencies, ideal conditions, and meaning.*Philosophical Psychology*, *9*(2), 235-259.

Kim, J. (1998). Mind in a Physical World, Cambridge: Cambridge University Press.

Millikan, R. G. (1984). *Language, thought, and other biological categories: New foundations for realism*. MIT press.

Millikan, Ruth G. (1986). Thoughts without laws: Cognitive science with content. *Philosophical Review*, 95, 47-80

Millikan, R. G. (1989). Biosemantics. *The Journal of Philosophy*, 281-297.

Millikan, R. G. (1996). On swampkinds. *Mind & Language*, *11*(1), 103-117.

Neander, Karen (1996). Swampman meets swampcow. *Mind and Language,* 11 (1), 118-29.

Peters, U. (2013). Teleosemantics, Swampman, and representationalism. *Grazer Philosophische Studien.*

Putnam, Hilary (1975). The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science,* 7, 131-193.