Success Semantics: Motivations, Problems, and Solutions

by

Meysam Shojaeenejad

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Arts

Department of Philosophy
University of Alberta

Abstract:

Inspired by Frank Ramsey's ideas on the relation between the truth of agents' beliefs and the success of their actions, J.T. Whyte developed the first fine-grained version of success semantics as a naturalistic account of the broad content of psychological mental states such as beliefs and desires (i.e. contents which depend not only on the agent's internal states, but also the external environment). The basic idea of success semantics is that what makes it the case that a state of an agent (e.g. a neural state) is a belief that P is that it tends to combine with the agent's desires to cause behaviour that would fulfill those desires if P. In this thesis I begin by discussing why we need a naturalistic theory of the content of mental states, and then evaluate the most well-known naturalistic accounts. On the basis of these evaluations I show how success semantics, as theory of broad content, fits into an ideal naturalistic theory of the content of mental states. Drawing on the work of Ned Block, I argue that success semantics yields the best account of broad content in a "two-factor" theory, the narrow content factor (i.e. the content that depends only upon the thinker's internal states) of which is accounted for by conceptual role semantics. However, Whyte's version of success semantics faces some problems. Some authors argue that the success of the actions that beliefs motivate under given circumstances is neither sufficient nor necessary for them to have such circumstances as their content. Some authors also charge the core idea of success semantics with vicious circularity, since it seeks to explain the content of beliefs in terms of that of desires, and Whyte's attempt to explain the content of certain "basic" desires without reference to the content of beliefs can seem to presuppose rather than explain their content. Finally, some authors question whether success semantics can explain the content of complex beliefs and desires, which depend for their influence on the content of the agent's other representations and seem far removed from basic desires and actions. In order to defend success

semantics against these objections, I make several contributions to current debates, offering a novel defense of some of Whyte's early ideas. First, I argue that, to solve the problems of insufficiency and non-necessity, we can return to Whyte's strategy of beginning by explaining the content of whole sets of beliefs influencing agents' conduct, and replace his view that accurate representations guarantee success with Bence Nanay's idea that they increase the probability of success independent of the accuracy of agents' other representations. Second, I argue that we can solve the problem of circularity by beginning with an explanation of the content of what I call "immediate desires," the content of which, unlike what Whyte identified as "basic desires," is independent of the content of the agent's other mental states. Finally, I argue that we can complete the solution of the problems of circularity and explaining the content of complex attitudes by adding to the basic idea of success semantics and my account of the content of immediate desires the idea that what for a state to be a non-immediate desire that O is for it to combine with the agent's beliefs to cause behaviour that would increase the probability of bringing about O if those beliefs were true. I use these ideas to propose a recursive success semantical account of beliefs and desires of arbitrary levels of complexity, explaining the content of each level of beliefs in terms that of desires of the same level, and explaining the content of each level of non-immediate desires in terms of that of beliefs of the level below it.

Dedication

This work is dedicated to Frank Ramsey (1903 – 1930).

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor Dr. Howard Nye for his constant support, motivation and guidance. It would not have been possible for me to take this work to this point without his magnificent support.

I would like to thank Professor Bernard Linsky for his support and guidance as my advisor through the process of this thesis, and also Professor Ingo Brigandt for his comments on my final work.

Finally, I would also like to show my gratitude for the financial support provided by the Department of Philosophy at the University of Alberta. It would have not been possible for me to work full time on this thesis without such support.

Table of Contents

Chapter One

Introduction

The topic of this thesis is an investigation of the prospects of *Success Semantics* as part of a naturalistic account of the intentionality or representational content of mental states. Success semantics, generally speaking, holds that what it is for mental states such as an agent's beliefs, desires and the like to represent something or have a certain content is for that content to play the right kind of role in explaining the success of the agent's actions. For example, according to most success semanticists, what makes it the case that some state of me is a belief that there is a desk in my room is that this state is disposed to combine with my desires to cause behaviour(s) that would fulfill those desires if there were a desk in my room (i.e. if this belief were true). For instance, the state will combine with my desire to have exactly one desk in my room to cause me to reject my friend's offer of a second desk for my room. In this thesis, I will try to demonstrate how this view can lead to a *naturalistic* or reductive theory of the intentionality or content of mental states, which explains what it is for brain states or mental states to represent the world in the non-intentional terms of more fundamental sciences, such as causation.

The objectives of this thesis can be expressed in three main steps. First, I will discuss the general motivations behind success semantics as a part of an ideal naturalistic theory of the content of mental states (in the second chapter). Second, I will investigate the major problems faced by success semantics, and discuss the main proposed solutions to these problems and debate about them (in the first part of the third chapter). I will show how we can find persuasive solutions for many of these problems by a wise combination of existing solutions (in the second part of the third chapter). Third, I will make a contribution to current debates about success semantics by offering a novel defense of some of J.T. Whyte's (1991) early ideas as a response to the main remaining problems (also in the second part of the third chapter).

One of the significant outstanding problems with success semantics is that its account of the content of beliefs and desires can seem to involve a serious vicious circularity. According to a general and rudimentary understanding of success semantics, what makes it the case that the

content of someone's *belief* is P (for example, there is a desk in my room) is that the belief will tend to cause the agent to perform actions which would fulfill her *desires* (e.g. her desire to having exactly one desk in her room) if P were true (e.g. to decline the offer of what would be a second desk). So far, the content of desires seems to be the basis for defining the content of beliefs. But, in a naturalistic account of intentionality, we need to explain the content of desires as well; i.e. in virtue of what does a desire count as having a particular fulfillment condition, or a state of affairs which will count as fulfilling the desire (e.g. what makes a state a desire to have exactly one desk in the room). Similarly, in a rudimentary understanding of success semantics, the content of desires would seem to be that a mental state is a desire for S if the state tends to cause me to do things that would bring about S if my beliefs were true. However, this idea seems viciously circular: in explaining the content of belief it appeals the content of desires, while in explaining the content of desire it requires the content of beliefs again.

J. T. Whyte (1991), in providing a solution to this problem, tries to provide a non-circular account of the content of a set of simple or immediate desires of an agent which does not rely upon the content of her beliefs. Roughly, what makes it the case that an immediate desire of an agent is a desire that S is that S would "satisfy" those desires in the sense that S would make the desires "go away" or stop influencing the agent's conduct, in a way that would reinforce a disposition of engaging in similar behaviour if the desire is manifest in the agent in similar circumstances. However, Whyte's solution in practice faces the problem that for the candidates he suggests as simple desires (e.g. desire for chocolate), they can be reinforcingly satisfied by states of affairs other their actual content (e.g. desire for chocolate can be reinforcing satisfied by having a fake chocolate which gives a taste, smell and crunch just similar to a real chocolate.) My innovation on Whyte's solution begins by contending that in any agent there is, in fact, a specific subset of desires with satisfaction conditions that can in fact be explained in a way that is completely independent of the content of her other intentional states. The most obvious such desires are immediate desires for the agents' *phenomenal states* – e.g. for chocolate-like taste (rather than chocolate per se). These immediate desires for tastes, feels, pleasures, pains, and so on are reinforcingly satisfied by the state of affairs of the agent being in these states – independent of whether she believes that she is. The second part of this novel Whytean account is to show how, for an initial set of an agent's beliefs, what makes them beliefs that P is that they are disposed to combine with these relatively simple desires to cause behaviour that would fulfil

those desires if P were true. Third, and finally, by giving the explanation of the content of an initial set of an agent's beliefs and desires in this way, we can then use them to explain the content of the rest of the agent's beliefs and desires via the notion that beliefs and desires combine with each other to cause behaviours which would fulfill the desires if the beliefs were true.

The focus of this thesis is to defend success semantics as a part of a naturalistic account of the intentionality of *psychological* mental states such as beliefs, desires, and intentions which play a significant role in explaining/predicting agents' behaviours without requiring any particular subjective aspects. Of course, I do not believe success semantics cannot be applied in explaining the intentionality of subjective experiences such as pain, or subjective experiences of seeing blue or orgasm. The reason to stick with psychological mental states here is simply its requiring less complexity than that needed to explain the intentional subjective states. Besides, I believe a naturalistic attempt to explain the content of subjective experience by success semantics will need to draw on an already existing success semantical account of the intentionality of psychological mental states (which is sufficiently fine-grained and sophisticated).

In the following parts of this introductory chapter, I will review some of the major motivations for giving a naturalistic account of the intentionality or content of mental states. This review will serve as a motivation for attempting to solve the philosophical problems faced by naturalistic theories of intentionality, which I will introduce in the second chapter. In addition to making the reader familiar with naturalistic attempts to explain the content of intentional states and solve the philosophical problems involved in doing so, my discussion in the second chapter will argue that success semantics offers the best explanation of mental states' *broad* contents (i.e. contents which depend not only on the agent's internal states, but also the external environment.)

In order to understand the main motivation behind giving a naturalistic account of the content of intentional states, I should explain its two main rivals. The first and most obvious alternative to the naturalistic theories is to hold that intentional states are a fundamental part of reality which requires no further explanation in non-intentional terms. In this regard, the most familiar ideas can be found presumably in *dualist* views claiming two separate realms of mental

and physical states, both equally fundamental and unexplainable in terms of each other (Descartes 1641; Brentano 1874).[1]

On the other side, however, the second rival to a naturalistic account of the content of intentional states would be the view that there really are no such mental states; e.g. all there is in the world are atoms, their movement, and their different configurations, to which mental states cannot be successfully reduced. The most familiar such view is known as *eliminative materialism*. Proponents of this view such as Churchland (1981) and Stich (1978) argue that intentional states are merely attributed by folk-psychology, and have no place in any *scientific* psychological account. They even reject that intentional states could be in some way built-up out of or explained in more fundamental non-mental terms; they think that (in our account of what exists) we must get rid of all of them and replace them by scientific psychological or neurological categories. An analogy to help understand this eliminative view could refer to the emergence of modern physics, chemistry, and biology; and how the categories of these disciplines were used to explain phenomena that had previously been explained in terms the workings of gods or spirits. Just like how the categories of modern scientific disciplines eliminated rather than gave a reduction-base for explanation in terms of gods and spirits, eliminative materialists think we must eliminate rather than find a reduction base for intentional states in our explanations of behaviour.

[1] Other views that take intentional mental states to be irreducible to physical states would include *idealist* and *neutral monist* approaches to them. Idealism (see e.g. Berkeley 1710) holds that mental states are the most fundamental states there are, and that physical states are to be explained in terms of or understood as built up out of them. Neutral monism argues that both mental and physical states must be accounted for in terms of some third kind of state, which would be neither mental nor physical as we currently understand them (see e.g. Chalmers 2015). Although these exotic views might sound promising for explaining the nature of subjective experiences or phenomenal consciousness (especially in light of the "hard problem" of trying to explain phenomenal states in physical terms), they do not seem particularly promising for explaining the nature of intentional states in general, which include states that can subconsciously direct agents' behaviors without having any subjective manifestation. Besides, neutral monism (or perhaps the most plausible version of idealism about intentional states) would eventually face a similar problem to that which naturalistic accounts face of how to explain intentional states in terms of neutral states (or, how agents' "macro intentional" states can be explained in terms of "micro-intentional" states of entities of which the agents are composed, which can be viewed as a variant of what Chalmers (2015) and others call "the combination problem"). Even dualist views, as I will discuss below, can face a similar problem in specifying fundamental dualistic psycho-physical laws. However the reasons mentioned above seem sufficient to give up these exotic views in the case of understanding intentional states in general, and to try to evaluate only dualism in this regard, which will be done into the main text.

In two following sections, I will delineate the reasons to reject both the dualist and the eliminativist approaches to intentional states. These will be the main motivations for giving a reductive or naturalistic account for the content of mental states.

1.1. Against Dualism and the Fundamentality of Intentional Mental States

In arguing against dualistic approaches, I start by the most commonly noted problem with this theory: the problem of the causal interaction of mind and matter. By holding that mental and physical states are two fundamentally distinct kinds of things, the dualist seems to need a persuasive explanation of how mental states and the body's physical states (including its neural states in particular) can interact with each other. Intuitively, we consider our beliefs, desires, and intentions as the causes of many of our behaviours towards the physical world. When we feel hungry, we might go to check the fridge for food because we believe there might be some food in there. Because we don't want to be burnt by fire, we refuse touching it. Similarly, physical interactions can cause us to have mental states. As we touch a shirt to test its quality, the fabric of the shirt via our sense of touch and chemical interactions among the nerve cells in our fingers and other neural cells in our brain, make us believe that the shirt is made of a high-quality material or not. But if beliefs and desires are fundamentally non-material states and our body and its environment are fundamentally material, how can these non-material states cause a movement in our material body and also be caused by material interactions of the environment with our body?

Before going deeper into the problem of causation for dualism, it is worth noting that causation could take place in a way that is radically different way from our intuitive understanding of it. To illustrate, we might intuitively consider physical contact between things to be necessary for them to exert causal influence on each other, while the causal story of quantum physics violates such intuitive conceptions.[2] Perhaps the most general way to talk about

[2] As an example for such unintuitive thesis in modern physics, I can refer to "action at a distance" in which physical contact is not taken necessary in cause and effect. See e.g. Berkovitz 2007.

causation among material entities themselves is to say that the physical world has certain fundamental laws, which can explain movement and changes in other properties of objects using concepts such as forces, masses, charges, and the like. A dualist can similarly attempt to understand causal interactions between the mental and the physical in terms of fundamental laws, positing fundamental psycho-physical laws in order to explain the interactions between non-material mental states and material (including neural) states. This can enable dualists to show how each of our physical and mental states can emerge by setting the parameters of each other, in a way that is no more metaphysically mysterious than fundamental physical causation.

Nevertheless, for the dualist view there are other sorts of problems beyond this potential mysteriousness of causal interactions between the two fundamentally distinct kinds of states. To explain these problems, it is useful to start by mentioning a distinction between *interactionist dualism* and various forms of *non-interactionist dualism.* While interactionist dualism supports the commonsense understanding of bi-directional causal interaction between the mental and the physical (i.e. mental states cause physical states and physical states cause mental states), non-interactionist dualism denies some part of this intuitive conception, so that sometimes both physical to mental and mental to physical interactions are rejected, or at least the mental to physical part is denied. Interactionist dualism, due to claiming that some physical states have non-physical causes, faces a dilemma regarding its potential violation of the principle of the *causal closure of the physical* (see Braddon-Mitchell and Jackson 2007: 14-19). According to this principle, all physical events have sufficient physical causes, and because our behaviours are physical events they would obey this rule as well. This principle seems extremely plausible in light of the overwhelming success of physical science. Although admittedly the prediction of all sort of physical phenomena via basic physical rules is still infeasible, there have been sufficient advances – especially in explaining how neural or chemical processes are based on some more fundamental physics – so that we can be confident in the existence of these physical causes. In other words, we already know enough about the physical explanation of neural events and the neural explanation of behaviour that we can be confident there will be no systematic anomaly in our attempt to explain our behaviour in physical terms indicating the operation of fundamentally non-physical causes, or any systematic violation of the predictions of purely physical science in explaining our behaviours that excludes them form the realm of what can be explained by the entities studied by fundamental physical science.

Now, the problem for interactionist dualism stems from its holding that non-physical mental states exert causal influence on our physical behaviours. Proponents of the theory must either (H1) deny the causal closure of the physical by holding that there are systematic deviations of our behaviours from the predictions of fundamental physics, or (H2) claim that there is some sort of systematic overdetermination of our behaviours by both sufficient physical and distinct sufficient non-physical mental causes. But neither of these two possibilities seems acceptable. In addition to making a very dubious empirical prediction about the failure of physical science, (H1) would involve maintaining that, although almost all of the physical world can be explained and predicted by some physical laws, the behaviours of agents are excluded in a way that makes them resist explanation in physical terms. Furthermore, the systematic overdetermination in (H2) not only makes for an unnecessarily complicated picture of the explanation of our behaviour, but also it seems to be desperately ad hoc. Of course, overdetermination can occur in cases where an event has more than one distinct cause, but in those cases the overdetermination is not systematic. To illustrate, suppose a person is shot and then dies from two bullets entering her heart at the same time, each of which could cause her death on its own. There is an overdeterminating explanation of this event, but because other events would have been different if one of the bullets had not been fired (e.g. the recoil of both of the shooters' weapons, the sound waves generated by both of the bullets, the shape of the impact wound, etc.) the overdetermination is not systematic. On the other hand, a systematic overdetermination of an event can occur by non-distinct causes in a sense that one cause is just a more detailed description of another. For example, a driver may stop at a stop light in virtue of its turning red and in virtue of its turning the particular shade of red it turns (e.g. crimson). Because the second cause is just a more detailed description of – and thus is not metaphysically distinct from – the first, such systematic overdetermination is entirely understandable. However, (H2)'s positing of overdetermination in explaining our behaviour that is both systematic and includes two completely metaphysically distinct causes appears to be a desperate attempt to retain dualism at the price of an extremely unparsimonious ontology. To illustrate, there were ancient views about the growth of plants and movement of clouds in the sky by spirits and deities, however, today we have a mechanistic scientific explanation of these phenomena. Although someone might think of the contemporary scientific explanations as a supplementary idea to those ancient views, due to the systematic over determination, the ancient views will turn out redundant so that we should

reject them. It is not right to explain the growth of plants and movement of clouds *both* in terms of biochemistry and atmospheric science *and* the plants' irreducible vital essence and the wind-gods' will which in every case happen to align with the distinct workings of mechanistic scientific explanations. Similarly, (H2)'s claim of overdetermination between physical and mental causes is a systematic one which makes it just as incredible, profligate, and desperate as above mentioned ways of reconciling the ancient stories in explaining the growth of plants and movement of clouds with their explicability in contemporary scientific terms.

Non-interactionist dualism is attractive because it does not face the foregoing dilemma regarding the causal closure of the physical. Parallelism is an early sort of non-interactionist dualism (often attributed to Leibniz 1686; see Braddon-Mitchell and Jackson 2007: 13-14). On this view, mental states have no causal interaction with the body and vice versa. This sort of parallelism held that God that sets up the sequence of mental states and the sequence of physical states in such a way that mentality and the physical world will be in a perfect harmony – meaning they seem to be interacting but they actually are not. For instance, although the physical world and mental states have no impact on each other, they are arranged in a way that beliefs and desires are followed by behaviours that would satisfy the desires if the beliefs were true. While parallelism avoids the dilemma faced by interactionist dualism in explaining causal closure, its appeal to a mysterious divine arrangement of mind and body seems just as unnecessary, complex, and implausible as the systematic overdetermination posited by (H2). This could be the reason why few people are proponents of this theory.

Epiphenomenalism is the latest version of non-interactionist dualism that has been proposed as a way of solving the forgoing problems. In this view, the impact of physical events on mental states is admitted, and only the influence of mental states on physical behaviours is denied (see e.g. Jackson 1982 for such a view in the case of qualia or subjectively identified features of experiences). As mentioned above, the *causal closure of the physical* would frustrate any interactionist dualism maintaining the influence of non-physical mental states on physical behaviours. However epiphenomenalism, due to its embracing the influence of physical events on mental states without accepting the influence of mental states on physical states, can smartly escape the problem, without the metaphysical troubles with parallelism.

Of course, epiphenomenalism is an unintuitive view. According to it, when one feels hungry and then pursues food, the pursuit is not caused by the feeling of hunger, but some neural activities in one's brain (the neural correlate of hunger) which at the same time cause both one's non-physical sensation of hunger and one's behaviour of searching for food. Epiphenomenalism asserts that the causal connection between the feeling of hunger and the behaviour of exploring for food is merely a false hypothesis that agents posit in trying to understand the causes of their own behaviour. This view, besides involving such a counter-intuitive denial of mental to behavioural causation, can be challenged by scientific discoveries in the field of evolutionary biology and how mental states could have evolved. According to considerable biological discoveries, we do know that initially, mentality was not present in the world, but gradually evolved in some species; we know, for instance, that the first single-celled organisms didn't have mentality while we human beings do.[3, 4]

Another problem for epiphenomenalism, owing to the rejection of mental to physical causation, is an epistemological one. If mental states have no causal role, they will leave no way to trace them, and consequently, there might be no way for us to get to know about them. The only way we infer other individuals' mental states is by witnessing their behaviours. Moreover,

[3] According to scientific discoveries, it seems that mentality (at least in a form that includes phenomenal consciousness) emerged by the very first vertebrates or perhaps in parallel it also emerged at an unknown time in in molluscs (given the likelihood of consciousness in cephalopod molluscs) and possibly arthropods (given the likelihood or possibility of consciousness in crustaceans, bees, and some spiders) – see e.g. Braithwaite 2010; Mather 2008; Tye 1997; and Elwood 2011). This view seems to be in contrast to what Descartes (1637; 1649) apparently believed about non-human animals. However, in my opinion, his view is due to a misapplication of the *intentional strategy.* I will discuss this strategy in more detail later, but briefly speaking, Descartes seems to assume that while explaining the behaviors of humans requires attributing beliefs and desires to them, the behaviors of non-human animals can be fully explained without requiring any intentional terms. However, as we discussed above, given the causal closure of the physical, there is just as much reason to think that the behaviours of humans have sufficient physical causes as that the behavior of non-human animals have sufficient physical causes. As we will see, our justification for attributing intentional states to humans is that doing so explains and predicts their behaviors by picking out a real pattern which is necessary to the most efficient way of explaining them. But this can be applied just as well to explaining and predicting the behaviors of many of vertebrates and probably also cephalopod molluscs and some arthropods (or at the very least to other mammals and birds– see e.g. Pappini 2008 Ch. 8 and 9 and Panksepp 2005).

[4] Someone might claim that it is still possible to consider the neural correlates of mental states as the main part which has been selected due to their important causal role and the mental states should be viewed merely as a byproduct of them (see e.g. Jackson 1982: 133-4). However, this view leaves the connection between mental states and their neural correlates as an open question. As I mentioned above, the best answer to this open question is presumably to explain the connection in terms of physio-psychic laws. But it can be challenged: why would there be such fundamental laws adding a fundamental aspect of reality (mentality) to neural correlates, the presence of which can be explained in virtue of their selective advantages? This makes the fundamental laws look more mysterious and dubious.

even in remembering and introspecting our own mental states, we need to rely on our own physical neural states and trace them back to some other mental states in a causal chain.

The final problem to mention for epiphenomenalism, which can be similarly applied to interactionist dualism, challenges the nature of physical to mental (not just mental to physical) causation. As mentioned before, in order to explain the causal relationship between physical and mental states, dualists can claim that there is a set of fundamental psycho-physical laws linking physical and mental states, similar to the way in which fundamental physical laws linking fundamental physical states to each other explain causal relations among physical states. However, as J.J.C. Smart (1959) argues, the reality of these fundamental psycho-physical laws is not convincing. In light of the causal closure of the physical, dualists are committed to holding that everything in the world can find a proper explanation and prediction in terms of physics, chemistry, biology and the like[5], except for mental states. Smart articulates how making such an exception for mental states would make them a mysterious sort of "nomological dangler," which require extra fundamental laws especially for them which can't be applied to anything else. As he puts the worry:

> It is not often realized how odd would be the laws whereby these nomological danglers would dangle….Certainly we are pretty sure in the future to come across new ultimate laws of a novel type, but I expect them to relate simple constituents: for example, whatever ultimate particles are then in vogue. I cannot believe that ultimate laws of nature could relate simple constituents to configurations consisting of perhaps billions of neurons (and goodness knows how many billion billions of ultimate particles) all put together for all the world as though their main purpose in life was to be a negative feedback mechanism of a complicated sort. Such ultimate laws would be like nothing so far known in science. They have a queer "smell" to them. I am just unable to believe in the nomological danglers themselves, or in the laws whereby they would dangle. If any philosophical arguments seemed to compel us to believe in such things, I would suspect a catch in the argument (Smart 1959: 142-3).

[5] Smarts takes biology and chemistry to be explicable in terms of and thus in a sense a branch of physics.

Smart's main concern thus seems to be that, although psycho-physical laws which are supposed to nomologically necessitate mental states based on neural/physical states can be held to be fundamental, they seem to be too narrow in scope to genuinely be fundamental. They are held merely to relate extremely complex physical states to simple mental states - states which are only possessed by individuals who constitute a teeny tiny fraction of the universe.

Although Smart's arguments against taking mental states, in general, to be just as fundamental as physical states are powerful, I think that they can be most effectively applied against the fundamentality of intentional psychological states. First of all, even if there were some special reasons to seriously hold dualism about sensations, subjective experience, or phenomenal states in the face of the costs Smart mentions, they would not support a similar dualism about psychological states such as belief and desires (see e.g. Chalmers 1996, 71-89). "The hard problem of [phenomenal] consciousness" challenges views which take subjective experiences to be simply built up out of physical states by questioning why any given physical state should have to be associated with any given subjective experience rather being associated with no experiences or different experiences. This problem is a legitimate one because the way in which we conceive and know about physical states is very different (e.g. third-personal and less direct) from the way we get to know about our subjective experiences (e.g. first-personally, and directly "from the inside"). However, psychological states such as beliefs, desires, and intentions, as we discussed before, do not seem to have an essential phenomenal feel; i.e. they can direct our behaviours without bringing any kind of phenomenal subjective experience to us. Therefore, we can conclude that the hard problem of phenomenal consciousness cannot make any problems for rejecting a dualistic account of psychological states, and seeking to explain them in naturalistic, non-intentional terms.

Besides, there are other reasons to hold that psychological states such as beliefs, desires, and intentions really are essentially causal states. Trying to conceive of the presence of these states, without their having any tendency to influence agents' behaviours by combining with other of the agents' psychological states doesn't seem coherent (see e.g. Braddon-Mitchell and Jackson 2007: 12). Our concept of these psychological states and their content - unlike subjective experiences - seems more closely tied to their tendencies to cause behaviours than what agents could feel about them or via them, by introspection.

1.2. Against Eliminative Materialism and the Non-Existence of Intentional Mental States

I have thus argued against the fundamentality of psychological mental states. If we are to believe that these states really exist, we need to find an account for them and also their content in the non-intentional terms of more fundamental physical entities and relations (e.g. neurons and the causal relations among neural states). The only remaining alternative to this project of naturalizing the intentionality of psychological states is to reject their existence. This is the position of *eliminative materialism,* which doesn't hold psychological mental states to be fundamental but doesn't count them as any sort of real entities either.

Eliminative materialism denies that there are any psychological states like beliefs, desires, or intentions, which (as we will explore more in the next chapter) represent the world in a special and underivative[6] way that is capable of genuine errors (or misrepresentations), which would seem to make such states distinct from more fundamental physical states. This view contends that in an ideal or (even partially ideal) scientific understanding of the world will not need to use folk psychological entities with underivative representational content such as beliefs and desires to explain and predict agents' behaviours. Hence, because these states will turn out to be explanatorily superfluous, considerations of parsimony dictate that we should conclude that they do not exist. This is similar to the way in which, because entities like spirits and deities are not needed in modern scientific explanations of natural phenomena such as the growth of plants, the motions of clouds, and the changes in seasons, we should conclude that such entities do not exist.

However, in opposition to this view, Jerry Fodor (1987) and Daniel Dennett (1981, 1991) have argued convincingly that, in essence, belief-desire explanations of our behaviours are

[6] By a special and underivative way, it is meant that the representational capacity of psychological states is not constituted by the representational capacity of something else. To illustrate, the resentational capacity of linguistic expressions can be considered to be built upon the representational capacity of related mental states. However, it is important to note that mental states' having intentional properties that are underivative in this sense doesn't mean that the intentional properties of mental states cannot be explained or analyzed as built up put of their causal roles. In fact, the whole purpose of this thesis in defending a naturalistic theory of mental content is to explore such a reductive explanation of the nature of such underivative representational capacities.

(unlike animistic accounts of natural phenomena) "too good to be false." Fodor (1987) begins his preface by mentioning some of his cat's regular behaviours such as her habit every morning of checking her bowl to search for food. These behaviours are different from the way that clouds, rocks, and trees interact with the environment. In explaining the movement of clouds we can appeal to how the wind, which is just a movement of air, pushes clouds, which are just blocks of steam, in a certain direction. He argues, in contrast to these phenomena, in order to explain the cat's behaviours we would say the cat every morning is hungry (and has a *desire* for food) and *believes* that food can usually be found in the bowl. Similar to this case, these belief and desire attributions can enable us to explain and predict an extensive range of flexible behaviors of hers; e.g. predicting all the various ways which she might come to check the bowl after a certain amount of time since her last meal, or explaining why she is used to approaching the bowl versus many other possible ways of finding food, or explaining why she may become particularly aggressive when other cats approach the bowl, and so on. This way of explaining and predicting the behaviours of agents by attributing beliefs and desires to them is called Commonsense Psychology (CSP).

Fodor also gives an example of setting an appointment to illustrate how CSP is indispensable in our everyday explanation and prediction of individuals' behaviours, and also in our communications. In the example, he sets an appointment with a stranger thousands of kilometres far away, by using only some sentences which do not even have to be expressed in precise terms. "Would you like to attend the conference?" the stranger might say and "I will be at your airport, on the conference day, early morning," he might answer. To explain how events such as a successful meeting at the airport can happen we have to attribute to the stranger a *desire* to arrange a pickup and a *belief* that the person will be there at a certain time. By using such simple belief and desire attributions, we can make a confident prediction about what will happen, which can guide our actions (such as purchasing the plane ticket and showing up for the meeting at the appointed time). Similarly, by attributing a *desire* of attending the conference and a belief about the time to the person, the stranger can make a confident prediction of the time that the person will arrive too.

Fodor argues that we use CSP in many aspects of our life. Even our ordinary languages have systematically evolved to include many structures which are associated with CSP in their deep structure. This is because we rely intimately upon the mechanism of belief/desire

attributions to agents. It has enabled us to make highly successful predictions of individuals' behaviours, despite knowing very little about most of the details of their internal states.

Fodor thus seems to be arguing that CSP is indispensable as part of the most effective way of enabling individuals to get out of themselves and view the world through the eyes of other individuals in order to predict and explain their behaviours. However, someone might claim that CSP might become dispensable if (or is even now in principle dispensable because) if someday we get to a technological point that enables us to communicate with other agents in a completely immediate way via some sort of highly sophisticated neural interconnection technique similar to what the movie Avatar (2009) portrayed. In order to access the workings of someone else's mind, the agents in this scenario could make a direct neural connection with someone else so that they could see what the other one is seeing, hear what the other one hearing and even observe what the other one is imagining. Could they then eliminate the role of CSP in enabling them to predict and explain others' behaviours? For example today, without this sort of technology, one might go on a date with a someone and feel that everything went completely fine, after which the other party might not answer one's calls and texts. To understand her behaviour one seems to have no choice but to start analysing her thoughts in terms of CSP. One may make many guesses about what happened and what she might think of one in light of the events. However, if one had the technology of sophisticated neural interconnection with the other party, it might seem that one would not need to use CSP to get her thoughts. It might thus seem that the indispensability of CSP is due to the lack of proper inter-communications, and thus does not give us reason to think that the beliefs and desires posited by CSP genuinely exist.

However, I believe that, although in this extreme scenario the usage of CSP in our public language would decrease, some aspect of CSP still would remain indispensable. First of all, even in explaining and predicting many of one's own behaviours, one needs to use a kind of third-personal application of CSP, although she has an immediate access to her experiences. It seems that in order to explain someone else's behaviours in light her sensations, one needs to form some hypotheses about how those behaviours are related to her background beliefs and desires. Besides, many of the beliefs and desires which motive us are subconscious, so that much of what we get by introspecting about our mental states can actually turn out to be merely confabulated third-personal belief-desire attributions that we falsely believe we have experienced (see e.g. Bargh and Chartrand 1999; Baumeister et al 1998; Norretranders 1998; Lakoff and Johnson

1999; Nisbett and Wilson 1977; and Wegner and Wheatley 1999). In conclusion, despite all of the possible inaccuracies and confusions of using CSP, it seems we have no choice but to try to explain our behaviours in terms what our most likely beliefs and desires in light of our experiences, and to try to minimize those which might be confabulated, mis-attributed, or mis-remembered.

Having thus argued that CSP is both non-fundamental but "too good to be false", it will help to examine a metaphysical account of the reality of CSP and other non-fundamental, but still genuinely existent, entities. I believe that the best such understanding of non-fundamental but real categories is that they pick out what Dennett (1991) calls as *real patterns* in the world, which makes it more efficient to explain phenomena in terms of such categories. The intentional belief-desire categories of CSP pick out such patterns, as well as all successful natural and social sciences such as physics, chemistry, biology and so on. The reason why we need to have several special sciences is precisely that there are several kinds of real patterns above the level of fundamental physics; otherwise, the most efficient explanations of everything would have to be couched in the terms of fundamental physics, which seems clearly to be false.

Dennett argues that, since CSP is a tool that successfully predicts agents' behaviours most of the time, and predictions cannot be made based on mere random information, there have to be some underlying discernible patterns in the world involving CSP. But what is a real pattern? To illustrate the concept of pattern, Dennett takes examples such as that of a centre of gravity, which is viewed in two very different ways by many philosophers. Some people think of it as a mere useful function; however, some people take is to be a real thing. The former corresponds to the stance of eliminative materialists, and the latter corresponds to the views of philosophers like Jerry Fodor who seem to treat the existence of beliefs to be as determinate as that of more fundamental entities like atoms. But, as Dennett argues, these two views are too extreme in their assumptions about what it takes for something to count as "real" or genuinely existent.

Dennett believes to make the best sense of the underlying patterns behind CSP, we should take a position between the two extremes. He calls his view *mild realism*; meaning that, it is neither true that facts about agents' beliefs and desires are as precise and determinate as very firm states of affairs such as "the earth is rotating around the sun," nor that beliefs and desires are

mere superstitious such as ghosts and gods, nor even that they are merely useful concepts for our current purposes due to our lack of proper knowledge and calculation power.

But what more precisely is the status of facts about agents' beliefs and desires according to the mild realism thesis? Isn't it true that everything is made of atoms and hence, any single behaviour of agents is basically a movement of atoms in space? In this regard Dennett (1981) considers an objection by Robert Nozick to the idea that beliefs and desires derive their reality from their indispensability to explanation, which suggests imagining a super-intelligent Martian civilization who are capable of predicting our behaviour via their ultra powerful calculation-simulation system in terms of atoms, just like Laplacean demons. Dennett argues that, by failing to use CSP in their explanations and predictions, those super-intelligent Martians are missing some very significant information about human beings and their social concepts and activities. Dennett takes an example of a stock market and a stockbroker who intends to buy 500 shares of General Motors Company. Here, Martians might be able to predict the behaviours (/motions) of the stockbroker's hands on the computer mouse and also her fingers when she clicks on the required buttons and so on. Dennett argues that, by viewing these behaviours in terms of merely physical categories, the Martians will miss certain patterns; such as those present in different ways of ordering and receiving 500 shares of GM (e.g. slightly different ways of using a computer or phone to order the shares, communicating with slightly different individuals, and so on). By attributing beliefs and desires to the stock broker, CSP provides a way to understand what is common to all of the different ways of ordering the shares by attributing a desire of buying them (e.g. by calling, using a smart phone app, text messaging etc.) and also all of the different way of receiving orders for shares by ascribing to her beliefs about such orders (due to seeing a computer screen, a reliable friend's voice, text in a message etc.).

These belief and desire attributions pick out related patterns by which our explanations and predictions are objectively more efficient than those of the super intelligent Martians. In the example above, by spending tiny amount of information in attributing a desire of buying 500 shares at a specific price and (true) beliefs about which actions will lead to a purchase at that price, we can make a successful general prediction about her behaviors; e.g. when the stock reaches a specific price, then she will buy them. Martians, on the other hand, in order to make this sort of prediction about her behaviours, would have to use all sort of extremely detailed

information and deal with elaborate calculations regarding every possible price scenario. Thus, although fundamental physical entities like the constituents of atoms and the laws governing them are the basis of events, there are some real patterns which are indispensable to efficient explanations that capture information about what unites similar cases. It is not just a matter of accident that CSP plays a deeply significant role in many aspects of our everyday life by enabling us to predict other individual behaviours.

To illustrate and explain the reality of such patterns, Dennett takes an example of compression/transmission algorithms to explain the significant role of CSP and the categories of special sciences. As we know, a completely randomly figured collection of data will take the largest compressed format because there is no way to re-express it in any shorter way. In fact, we need to transmit such a collection bit by bit. On the other hand, consider a file which contains only data which is fairly organized, e.g. one that contains 1000 bytes with a value of 0xF0 and only a few with a different value (e.g. Y, Z). Here it is not necessary to report all bits of data at all, but we can simply transmit "All bytes of 0xF0 – the exceptions being case 1 with value Y, case 2 with value Z, etc." Although it would be highly accurate to transmit the collection of data bit by bit, it is not efficient, even if complete accuracy is desired. Bit by bit transmission would be comparable to what Nozick's example of super-intelligent Martians do. It is quite inefficient to explain/ predict individuals' behaviours in terms of mere atom movements, while CSP enables us to compress the data (by recognizing the patterns). Although in some such efficient compressions the information which does not conform to the patterns could be transmitted (as in the above example the non-conforming ones were transmitted), for many explanatory purposes we do not need to care about such non-conforming information or "noise"– it may be sufficient to know only how much of the information is non-confirming, or even that it is not too much. Explanations and predictions by CSP (which seem to ignore the "noise" of behaviour that does not conform to the belief-desire pattern due to various neural and other physical causes) might thus not be as accurate in some cases as what Martians would get, but its massively greater efficiency renders it highly effective.

In conclusion, CSP (or a data compression algorithm) picks out the patterns which are perfectly real and necessary for efficient explanations that carry information about similarities among cases; i.e. the particular patterns which are most efficiently utilized independent of the

explanatory project at issue; the patterns which may not be as completely determinate as very fundamental physical facts. (In the section on interpretivism in the next chapter I will have more to say about Dennett's promising views about how intentional mental states pick out a real pattern in the course of explaining his *interpretationist* view about their content, which I will argue helps to motive Success Semantics as the best theory of their broad content).

1.3. The Practical Importance of Attempting to Naturalize Intentionality

So far, I have argued that: 1) the intentional mental states such as beliefs and desires are not fundamental, but 2) they are still real and explanatorily indispensable. With these two conclusions in hand, it seems there is no other option to make sense of the metaphysical status of intentional mental states except to naturalize them; i.e. to explain how such intentional states are built out of non-intentional ingredients. However, alongside this primary metaphysical motivation, there are considerable practical motivations for the project of naturalizing intentionality. Even if someone does not find the arguments in favour of the metaphysical motivations convincing (and thinks of intentional mental states as either fundamental or in principle eliminable), the practical motivations can make sense of the attempt to naturalize intentional states.

It seems important to achieve a consistent and reasonable understanding (or at least guess) of what it would take for an artificial intelligence [AI] to have states which should count as intentional mental states in a sense similar to or in the same way that many animals' mental states count as intentional. Achieving such an understanding not only enables us to evaluate the success of various AI projects (see e.g. Dretske 1994), but also helps us to understand to what extent AI systems can be considered as moral subjects, towards which we have moral obligations such as not to harm them or to benefit them for their own sakes (see e.g. Basl 2014). Understanding what it takes for an entity to have intentional mental states can also be practically helpful for us in determining the extent of intentionality in various non-human animals and also very young or intellectually disabled humans (see e.g. DeGrazia 1996; Knutsson 2015; Boly et al. 2013). Some people have argued that the capacity for various subjective experiences seems

crucial to the extent to which an agent or object can be harmed or benefited in either a literal or morally relevant sense (see e.g. Singer 1975; DeGrazia 1996), and many have argued that the intentional mental states are essential for such subjective experiences (see e.g. Tye 1997; Baars 1988; Merker 2005; and Rosenthal 2002). Some authors have also argued that intentional mental states such as beliefs and desires, apart from being essential to subjective experiences, are themselves morally relevant (see e.g. DeGrazia 1996; McMahan 2002; Feldman 2004; Basl 2014).

Even if the metaphysical ideas behind the project of naturalizing intentional mental states for some reason happen to be proven false, I believe the attempt to give a naturalistic account of what it takes for an entity to have these states can still be helpful for the above mentioned practical reasons. Clarifying the plausible physical correlates needed for the presence of mental states can help us in finding a more systematic method of determining the presence of intentional mental states via physical evidence. Thus, even if the project of naturalizing intentionality, as a metaphysical account of how intentional of mental states fit into our picture of what there is, fails, our best naturalistic theories in this regard should be re-interpreted as our best theories of which physical structures give rise to which distinct intentional states (at least in the world as it is).[7]

To illustrate, I would like to refer David Chalmers' parallel case of qualia and machine consciousness (Chalmers 1996, ch 7and 9). He argues that, although subjective experiences are indeed non-physical, our best, least arbitrary understanding of how physical states give rise to non-physical experiences should allow that systems which have the same sort of physical states as our brains will have the similar kinds of experiences.[8] Presumably, the relevant physical states must be specified in the broad, multiply-realizable way that physicalistic theories of subjective

[7] On the other hand, if it turns out that the correct metaphysical view towards intentional mental states is actually a version of eliminative materialism, then we should re-interpret our the naturalistic theories as theories of which physical structure should count as a mental state and which should not. This is significant not merely for scientific explanatory purposes, but also in understanding the extent to which various systems share similar abstract structures with us, which can help us in evaluating the ethical importance and determining the relevant ethical stance of them (even if they might not be very important for scientific explanations.)

[8] Chalmers argues against the physicalist thesis that mental states supervene on physical states, seeking to establish that it is metaphysically possible for entities to have the same physical states as us but different or no phenomenal states. However, he believes that this can not happen in the actual world or better to say it is not nomonologically possible (i.e. it can not be consistent with the fundamental laws of the actual world). Although Chalmers argues that philosophical zombies which are our materially identical doubles with no subjective experiences are possible, it would be intolerably arbitrary to think (and also our understanding of the mind must not allow us to think) of some actual individuals who are physically just like us as mere zombies with no subjective experiences.

experience have sought to specify the physical states that they take to be subjective experiences themselves (such as Rosenthal's (2002) higher-order thought theory; Tye's (1995) first-order representationalist theory; Baar's 1988 global workspace theory, and Tononi's (2008) integrated information theory). For example, if we are dualists about the relationship between physical states and subjective experiences, we should re-interpret the best physicalist theories of subjective experience as giving us the best, least arbitrary understanding of the fundamental psycho-physical laws that connect any physical states that play the relevant functional roles to the relevant subjective experiences. Similarly, if we take a neutral monist view (or even an idealist view) about the relationship between physical and phenomenal states, we should then re-interpret our best naturalistic theories of subjective experience as explaining how the "proto-experiences" (or "micro-experiences") of the most fundamental entities combine to give rise to genuine experiences (or "macro-experiences) in beings like us (cf. Chalmers 2015).

I believe that, if for some reason we had to admit that intentional states are fundamental (or a distinct, non-physical aspect of the right sort of combinations of fundamental states that are neither physical nor mental as we understand them), then the exact same should be said about the re-interpretation of naturalistic theories of intentionality. Our best, least arbitrary understanding of how physical states give rise to intentional states must allow that systems which the same relevant physical states as our brains will have the same sort of subjective experiences. In this case the relevant physical states must be specified by our best naturalistic theory of intentionality, re-interpreted either as our best, least arbitrary understanding of the fundamental psycho-physical laws linking physical states with intentional mental states, or our best understanding of how the "proto-intentional" states of fundamental entities give rise to "genuinely intentional" states in beings like us.

Chapter Two

Naturalistic Theories of the Content of Mental States

After motivating naturalistic theories of mental intentionality in the first chapter, the second chapter of this thesis will examine the main naturalistic theories that have been proposed and also some of their crucial problems and shortcomings. This will motivate success semantics, which, as I will argue in the final chapter, can provide a successful response to these problems.

Let's begin with the simplest kind of naturalistic theory of mental content, which claims that mental states *represent* the features of the world that have *caused* them, or the features about which they *carry information* (see Dretske 1994, 1981 for this notion of indication or information carrying). This view claims that what makes a mental state a representation of a dog is that its tokenings (or occurrences) are reliably caused by dogs or reliably correlated with dogs.

2.1. The Crude Causal and Information Carrying Theories

One of the simplest naturalistic theories of what it is for something to have a given representational content can be found in what Fodor (1987) calls the Crude Causal Theory (CCT) of content. According to this theory, entities with representational content denote their causes (with tokenings of these entities representing their specific causes and types of these entities representing the properties whose instantiations reliably cause their tokenings). This theory, regarding the content of mental states, would hold that a particular mental state (which might itself be a brain state or be realized by a brain state) represents a kind of thing in virtue of being reliably caused by that kind of thing. As an example, suppose one has a mental state, which we will call HORSE[9]. According to CCT, HORSE represents horses (or the property of being a horse) in virtue of being reliably caused by horses.[10]

[9] Mental entities like concepts are denoted by capital letters.

[10] Someone might object that in almost every instance of something causing a mental or neural state, there are many other features of the world that take part in the causation. For example, in causing HORSE, not only horses but also some events such as light, chemical processes, sound waves and etc. enable us to get the relevant information. The important question is that why should we take only horses to be the content of HORSE and not any of these other causes? To answer this objection we need to appeal Dretske's informational theory of content, which is a slightly

Fred Dretske (1994) argues that something like the CCT might be an acceptable naturalistic account for some sort of intentional or 'aboutness' relation; the relations which we might call 'indication' or 'meaning'. For example, we might say that smoke indicates or means fire, and the direction of the needle of a compass indicates or means the direction of the magnetic poles. However, CCT is actually unable to give a satisfactory account even of the meaning or indication relation. As a counter example for CCT, we might say dark clouds mean or indicate rain in the near future – while it is obvious that the future rain could not cause the appearance of dark clouds in their past. With these shortcomings of CCT in mind, Dretske, in his influential book, *Knowledge and Flow of Information (1981),* introduces his informational theory of content by providing the most careful analysis of this indication relation. His proposal is that what it is for a natural state S to indicate another state of the world P is for S to *carry the information* that P, where this is for S to be lawfully correlated with P. In his view, smoke is lawfully correlated with presence of fire, the direction of a compass needle lawfully correlates with the direction of the magnetic poles, and the presence of darks clouds lawfully correlates with rain in the near future. In applying Dretske's information carrying theory of representation to mental states, it would entail that my mental state HORSE represents horses (or the property of being a horse) in virtue of carrying the information about (or being reliably correlated with) horses or the instantiation of the property of being a horse.

2.2. Misrepresentation and The Disjunction Problem

Although the crude causal theory and especially the information carrying theory may do a good job in naturalizing a particular kind of intentional relation, Dretske (1981;1994) and Fodor (1987; 1990) explain how they seem to face a serious problem as theories of the intentionality of mental states. Indication or information carrying can be found anywhere in nature; all sorts of entities in nature can indicate or 'mean' things about many sorts of others entities – and these relations seem relatively simple to understand in naturalistic terms. These relations do not seem

---

more sophisticated theory. It maintains that, although a mental state may have various causes, it only represents ones which it carries information about; i.e. HORSE only carries information about horses not light, sound waves or the like, because it can reliably occur in us without any particular configuration of light or sound (see e.g. Dretske 1981, esp. Ch 6 for this sort of point about the contents of perceptual states).

to belong the particularly difficult and distinctive sort of intentional relations regarding mental states. The representational content of mental states has led some authors (e.g. dualists) to hold that mental states cannot be physical states, or even that this complicated sort of intentionality does not really exist (viz. eliminative materialists). Dretske (1981; 1994), inspired by Roderick Chisholm's idea about the essential ability of thoughts to misrepresent the world, says: "Without the capacity to misrepresent, we have no capacity for the kind of representation which is the stuff of intelligence and reason;" "what Chisholm describes as the first mark of [special, mental] intentionality [is] the power to say that something is so when it is not so, the power to misrepresent how things stand in the world." As Dretske and Fodor suggest, the underivative capacity of mental states to misrepresent the world seems to be the difficult and distinctive feature of these states. Misrepresentations in other things, such as spoken and written sentences, seem only to be derived by their conventionally serving to express the content of underlying mental states (see e.g. Grice 1957; 1969).

The problem with the crude causal and information carrying theories as theories of mental intentionality is essentially that they make genuine misrepresentation impossible (or at least fail to give an adequate account of it). We might see a cat on a dark night but believe that it is a dog. Indeed, cats on dark nights might be mistaken for dogs and thus reliably cause or correlate with tokens of our dog representation.[11] In this case, our dog representations are

[11] In the examples above, HORSE or CAT are sub-propositional representations, while the content of beliefs, desires and intentions are typically taken to be propositions; that's why they are called propositional attitudes. I illustrated the way that the crude causal and information carrying theories could explain how a sub-propositional representation, like HORSE or CAT, represents a kind or property. In suggesting an account of how such theories can get from an account of sub-propositional representations to the propositional representations involved in propositional attitudes such as beliefs and desires, Fodor's (1987) thesis of a 'language of thought' is a viable explanation. The basic idea of this thesis maintains that propositional representations are built up out of sub-propositional representations combined via a syntax, like in a language. Of course, 'syntax' in Fodor's thesis is simply a general physical property of how the sub-propositional representations occur together, which determines both their causal role and the content of the combined state (see e.g. Fodor 1987, especially Ch. 1 and Appendix). Although Fodor in explaining the content of mental states rejects both the crude causal and information carrying theories, like I mentioned, this is a viable explanation especially for the next two naturalistic theories of content (the teleological theory and Fodor's own) in this regard. The relation between sub-propositional representations and propositional attitudes is important because my discussion of these theories will focus (as is standard in the literature) on their accounts of how the contents of (non-logical) sub-propositional mental representations (e.g. concepts) are determined.

The crude causal and information carrying theories, like the upcoming teleological and asymmetric dependence theories, in the first instance give accounts of the content of 'mental representations' (such as mental concepts and thoughts or representations of propositions formed by combining these concepts) rather than propositional attitudes like beliefs and desires. Fodor in explaining how such theories can constitute accounts of beliefs and desires suggests this following general account. What makes a state (e.g. a neural state), a mental

reliably caused or correlated with not only dogs under good lighting conditions but also cats on dark nights. Therefore, if we try to explain the content of a mental representation as states of the world that reliably cause or are correlated with its tokenings, we must concede that what seems to be my representation of a dog actually represents both dogs and some cats on dark nights, thus making the content of my mental representation 'dog or some cats on dark nights'. In this situation, the content of my mental representation is not 'dog' or 'some cats on dark nights' but 'dog or some cats on dark nights'. The problem generalizes to make misrepresentation impossible—any time we might want to say a mental state misrepresents x as y, the crude causal / information carrying theory entails that it actually correctly represents x as x or y.[12]

2.3. Teleological Approaches to Solving the Disjunction Problem

Teleological approaches can be viewed as ideas inspired by causal or informational theories of content with some deviation and innovations. These theories seek to explain the content of mental representations, not as whatever reliably causes them or whatever they carry information about, but as what they have the *teleological function* of being caused by, indicating, or carrying information about.

---

representation of a proposition in general is that: it is the right sort of (e.g. syntactic) combination of states which bear the right causal, informational (or as we will see possibly historical) relations to things in the world. More specifically, what makes a mental representation of a proposition a *belief* is that it plays the causal roles characteristic of beliefs, such as tending to get caused by perceptions, tending to combine with other beliefs in inferences to produce new beliefs, and tending to combine with desires to produce behaviour that would fulfill the desires if the beliefs were true. Similarly, what makes a mental representation of a proposition a *desire* is that it plays the causal roles characteristic of desires, such as tending to combine with beliefs to produce derivative desires and behaviour that would fulfil (or bring about its content) if the beliefs were true. (These causal roles do tend to make reference to each other, but as we will see this need not give rise to a vicious form of circularity). Fodor and other authors often refer to a mental representation playing these causal roles distinctive of particular propositional attitudes (of belief, desire, etc.) as the propositional attitudes occurring 'in the belief box', 'in the desire box', etc.

[12] One might think that the crude causal and information carrying theories could at least explain states as misrepresenting things as having properties they do not have if the states are caused by things that do not have properties that tend reliably to cause or lawfully co-vary with them. For example, an old compass needle might not be able to show the directions of magnetic poles correctly, and someone might think that the old compass is misrepresenting the directions. But as Dretske (1994) argues, this does not seem to be a form of genuine misrepresentation. It seems incorrect to say that the state of the compass needle is itself in error, false, or directed towards a state of affairs that does not obtain. As Dretske (1994, 474) puts it, *we* could be misled into misrepresentation by these states, or "they could fool us," but they do not themselves seem to be incorrect, or part of a system that can itself be fooled.

A commonsense concept of something having a teleological function, as Dretske (1994) argues, can be understood in terms of that function's explaining how the thing came to have its features. Dretske introduces these functions by calling them 'natural functions', indicating that things can have them independent of what we might happen to use them for, which is also distinct from whatever it might tend to do with them. In general, Dretske claims that what it is for a system to have a natural function of doing F is for the doing of F in the *past* by the system or the ancestors of the system to explain the presence and operation of F in the system. He then introduces two kinds of natural functions: phylogenetic, or those related to the history of the *evolution of the system* from ancestral systems, and ontogenetic, or those pertaining to the system's own *learning history*. The most straightforward examples of phylogenetic teleological functions come from biology. For instance, it seems that a teleological function of the heart is to pump blood, which can be understood in terms of hearts having been biologically selected to pump blood. That is, the reason hearts came to proliferate and operate the way they do is because ancestral versions of hearts that pumped blood tended to proliferate because they did so (because the organisms with such hearts tended to pass on genes that tended to give rise to such hearts). Beside pumping blood, hearts do other things such as making certain noises, but they do not have a teleological function of making noises, because the presence of making noises in hearts through the history of evolution did not cause hearts to proliferate. (The presence of making noise in hearts was simply a byproduct of the physical mechanism of pumping blood; i.e. the blood-pumping features of this mechanism are what do the work in explaining how they came to proliferate and have the causal tendencies that they have).

2.3.1. Phylogenetic Teleological Accounts

Ruth Millikan (1989) seeks to explain the content of mental states in terms of their phylogenetic or biological functions, while Dretske pursues the explanation of the content of psychological states like beliefs and desires in terms of their ontogenetic functions or functions acquired through a history of learning. Phylogenetic functions in Millikan's theory might seem to give a naturalistic account of the content of perceptual states, for example, visual experiences of motion (e.g. V5 visual cortical activity – see Block 2005) and the experience of burn sensations. It seems plausible to hold that the content of experiences of motion is a motion, even though these experiences can be reliably caused by other sources such as illusion-of-motion images and

direct neural stimulation (Block 2005). A plausible explanation of why their content is motion (as opposed to such images and neural stimulations) seems to be that unlike the cases of illusions of motion and direct neural stimulation, the phylogenetic function of experiences of motion or states of V5 visual cortical activation is actually to track or correspond to the genuine motion. To illustrate, the ability of ancestral versions of the V5 visual cortex to detect or correspond to genuine motion (as opposed to illusions of motion, or direct neural stimulation) is what explains the proliferation and causal tendencies of the V5 visual cortex today – since tracking genuine motion is what explains why ancestral organisms reproduced and passed on genes that coded for V5 visual cortices with their current causal tendencies. Similarly, it seems plausible to say that the content of experience of beign burned is a kind of bodily damage at the location (see e.g. Tye 1995), although these can be reliably caused by things like direct neural stimulation or the like. Similarly, Millikan's view, by appealing the phylogenetic teleological account of representational content, seems to have a plausible explanation of why the content of experiences of burning sensations is one's tissue being damaged. She would say that this is because our ancestors' state of burning sensations tracked actual bodily damage (and also motivated them to do something in response) – rather than direct neural stimulation or the like – that our ancestors reproduced and passed on genes that coded for tendencies to have states of burning sensations with their current causal tendencies.

However, phylogenetic teleological accounts such as Millikan's face several problems in explaining the content of mental states. First, we can find instances of phylogenetic functions in some organisms which are uncontroversial non-mental cases. Dretske (1986) discusses this problem by taking the example of magnetosomes in anerobic bacteria that pull the bacteria toward geomagnetic poles[13] in order to keep the bacteria away from oxygenated water, in which they cannot survive. According to Millikan's account, the pull of bacterial magnetosomes represent the direction of oxygen-free water, because they have a phylogenetic function of tracking or carrying information about the direction of oxygen-free water. This is because their carrying information about the direction of oxygen-free water is what explains why ancestral bacteria to survived, reproduced, and passed on genes that coded for the magnetosomes with their current properties. It is also possible on Millikan's account for magnetosomes to genuinely

[13] Magnetosomes tend to pull the bacteria towards geomagnetic north (in the northern hemisphere) and towards geomagnetic south (in the southern hemisphere).

misrepresent the direction of oxygenated water – for instance, if one holds a magnet over the bacteria from the direction of the surface (which will cause them to move upwards, into an environment that kills them). It might, however, seem to be a mistake to attribute the sort of genuine misrepresentation which seems distinctive of genuine mental states to bacterial magnetosomes.

Although Millikan (1989) precisely claims that the pull of the bacteria's magnetosomes represents the direction of de-oxygenated water and that the bacteria misrepresent the direction of the surface as the direction of de-oxygenated water when one holds a magnet above them, (which entails a potentially implausible claim that the non-psychological system of bacteria can genuinely misrepresent the things, or in Dretske's (1994) words "itself be [literally] fooled."), someone might think that there is still some room for Millikan to escape the objection. In fact, it seems Millikan need not claim that the bacteria possess something like false *mental states* or false beliefs or perceptions; it is open for her to claim that the mark of the mental is representation capable of misrepresentation together with the right kind of flexible, domain-general functional roles. However, it still might seem counterintuitive to hold that there can be genuine error without a mental individual "to be fooled."

Perhaps, more importantly, a problem of indeterminacy threatens phylogenetic teleological theories like Millikan's. As Dretske (1986) argues, there seems to be no principled way to determine whether the biological function of magnetosomes is to indicate oxygen free water or to just geomagnetic poles. That is, it seems just as plausible to say that the magnetosomes were selected to correspond to magnetic poles. Although it might seem that tracking the magnetic poles was adaptive because magnetic poles were the direction of oxygen free water, still it was the direction of magnetic poles which have been selected to track, not the absence of oxygen in the water per se. After all, it is the direction of magnetic poles that is actually causing the magnetic response of the magnetosomes. Moreover, any plausible attribution of phylogenetic functions must allow that it can be something's function to do something, even though this was only adaptive due to certain background conditions, without its function being to track those background conditions as well. For instance, it was adaptive for sensations of burning to track tissue damage only because various environmental conditions obtained (e.g. the pull of gravity, the absence of universal spontaneous healing, etc.), but that

does not mean that the function of such sensations – and thus their content on Millikan's account – includes the presence of those background conditions. To generalize, we can put Dretske's indeterminacy objection as follows. For any alleged case of a state failing to correlate with state X, which it was allegedly biologically selected to track, it seems that we can instead say that it is successfully correlating with conditions Y which it was in fact selected to track. It's just that its correlating with Y was adaptive, or had gene-proliferating results, because in past environments with Y tended to correlate with X (for a similar objection see Fodor 1990).

Another substantial problem for phylogenetic teleological accounts like Millikan's is that even if we can consider them as convincing theories in explaining the content of relatively hard-wired mental representations such as perceptions and physical pains, it seems difficult to generalize them to give a plausible account of the content of many of psychological states such as beliefs and desires. For example, the belief that earth is rotating around the sun, which is an acquired mental state, does not seem to be selected by genetic evolution to correspond to what it represents; i.e. there are no genes that code for this belief (even in conjunction with standard environments of development.) There would have been no time for biological evolution to select for such genes even if they did exist. Millikan (1989) in attempting to provide a solution to this problem claims that, although there are no genes to code for these beliefs, there are some genes which *indirectly* select for these sorts of beliefs which correspond what they represent. In fact, evolution selected for a general cognitive system, which functions adaptively when its process of inquiry is successful. This is similar to the idea that genes indirectly selected for calluses to arise in particular areas of friction on the skin. That is, evolution selected for this general tendency of skin tissue, which functions adaptively when friction on the skin arises at particular locations (see Millikan 1989.)

However, Millikan's response faces the problem that it is unclear how exactly the content of particular beliefs and desires, which of course can be true or false (or fulfilled or unfulfilled) can be explained, even if they are formed under conditions that would have been maladaptive. To illustrate, suppose I might come to form a true but quite unjustified belief that earth is rotating around the sun (e.g. via a sophisticated neuroscientist manipulating on my brain in the future.) It is unclear how some genes indirectly select for this belief. Another problem with phylogenetic teleological theories is that phylogenetic evolutionary history seems irrelevant to whether our

beliefs are true. For example, consider a scenario in which there had been no biological evolutionary history in the world and all organisms came perfectly formed into existence by a complete fluke about 300 years ago (or if you like 30 years ago; if we opt for the shorter duration we need only imagine many of coming into existence with quasi-memories of things that did not happen). Although this scenario is very unlikely to be true, it provides an intuition that we could still have intentional mental states, even though (in virtue of lacking an evolutionary history – or anyway an evolutionary history long enough to give rise to a biological function for our cognitive systems) our cognitive systems would lack a biological function. As another example, let's consider a less extreme version of above scenario (which I will tweak and elaborate on later as the swampman scenario). Suppose that 30 years ago I came into existence as an infant, not as result of biological reproduction, but as a result of lightning hitting a swamp and, by complete fluke, a being who was molecule-for-molecule exactly like a human infant popping into existence. I then grew and came to have my own personal history. While it seems my cognitive system (in virtue of my completely lacking a phylogenetic history) would have no phylogenetic function, I could still have the intentional mental states that I currently have.

2.3.2. Ontogenetic Teleological Accounts

Dretske's (1994) teleological theory of mental content is based upon his original (1981) informational theory. He discusses how besides phylogenetic functions we need to appeal ontogenetic functions to explain the content of mental states. While phylogenetic functions are realized by a history of selection, ontogenetic functions can be realized by learning processes within individuals' lifetimes. As mentioned before, the basic idea of teleological theories is that what it is for a state N (e.g. a mental or neural state) to represent T (e.g. a kind of thing or property) is for it to be that N's natural function is to be caused by or carry the information that T is present. In other words, there is a tendency in one to token N in virtue of previous instances of N being caused by or covarying with instances of T. When a system learns, in general, an internal state N will be associated with an external condition T. If N's occurs each time in the presence of T, and N has a tendency to cause a response that, if performed in the presence of T, makes the agent more likely to respond in such a way in similar circumstances in the future, a tendency in the agent to token N in the presence of T will be *reinforced*. This can constitute a

kind of ontogenetic version of the selection of the state to correspond to or carry information about something.

What Dretske (1988) calls discriminative instrumental conditioning is perhaps the simplest and clearest case of the above mentioned emergence of ontogenetic functions of being caused by or carrying information about certain states of the world. In this process, a behavioural response is rewarded only in the presence of a certain environmental condition. To illustrate, suppose that Mickey the mouse learns that pressing a lever gives food when a certain human is present. If a certain human is present Mickey's pressing the lever will cause him to obtain food. Mickey has some state N which occurs after hearing the human's footsteps, or hearing her voice, or smelling her scent. Each time Mickey presses the lever when he is in state N and the human is in fact present, the food reward reinforces state N's tendency to cause him to press the lever when he, in fact, wants food (i.e. when he is hungry). Thus, state N comes to occur in presence of the human because it's doing so in the past reinforced it. So Mickey's learning history has selected the state N to correspond to the state of human's presence: its doing so in the past explains Mickey's having the state and its tending to work as it does. It is possible that state N in Mickey could be tokened by things other than the human as well; e.g. playing a recording of human footsteps or voice or scent instead of a real human presence. Dretske argues that these are in fact instances of genuine misrepresentation because N has (via a process of learning) acquired the function of responding to the presence of the human rather than merely these pieces of evidence of the human's presence. It was the presence of the human, rather than the mere presence of these pieces of evidence, that caused him to receive the food reward and thus reinforced his tendency to token the state in the presence of the human.

In the previous subsection, we saw how Millikan's account faces the problem of indeterminacy. The parallel problem regarding Dretske's account would be why Dretske gets to say that state N was selected to track the human's presence, rather than selected to track things like the sound of her footsteps (which were only rewarded because they tended to occur in the presence of the human). In response, Dretske (1986; 1988) argues that unlike the case of the phylogenetic functions of magnetosomes in anerobic bacteria, Mickey's state N tends to be caused not only by the sound of the human's footsteps, but the sound of her voice, her scent, and so on, which causes Mickey to have many ways of obtaining information about the human's

presence. Besides, Mickey could (and given the chance would) associate other forms of evidence of human's presence with N. All these various forms of evidence of the human's presence tend to reinforce Mickey's state N because of tending to track the human presence. Dretske concludes that, with states capable of learning like Mickey's state N, his ontogenetic teleological theory can give a determinate account of the content of mental states.

Although Dretske's account might seem to escape the problem of indeterminacy, there are still some potential problems with it. As we already discussed, Dretske's account maintains that for state N to represent thing T (in a sense that allows for genuine misrepresentation) is for N to have the natural function of carrying the information that T is present, and for N to have this function is for it to be the case that (i) N's presence and causal propensities are explained by N's being caused by T in the past, and (ii) to the extent that N's presence and causal propensities are or would be explained by various actual and hypothetical things other than T, this would only be because those intermediaries are or would be correlated with T. Fodor (1990) discusses how the idea in clause (ii) seems to make the content of N depend upon the actual and hypothetical pattern of possible rewards (and in particular whether they would be tied to the presence or absence of T). However, this idea seems to make it too difficult for an agent to be mistaken about the possible situations in which she will be rewarded, or to over-generalize a concept that is less helpful in terms of being rewarded. To illustrate, through the process of learning Mickey might come to believe that pressing the bar in presence of every human (instead of a specific human) will bring rewards. This idea is made by an over-generalizing a concept that is less helpful for getting rewards. However, it seems Dretske's account would implausibly entail that Mickey's belief already represents the presence of the specific person who feeds him, because he would only be rewarded for pressing the button in the presence of that human.

Someone might think that there is a room for Dretske to revise clause (ii) so that it does not refer to what would explain the reinforcement of the state under various hypothetical conditions. However, this would make Dretske's account face the problem of indeterminacy in the same way that Millikan's account does. In other words, there seems to be no principled reason to hold the content of Mickey's mental state is the presence of the human, rather than the presence of the pieces of evidence that correlate with the human's presence. Although it is true that those correlating evidences did reinforce Mickey's state only because the human was

present, it is still equally true that they reinforced his states only because of some things unrelated to the presence of the human – such as the absence of overly loud noises, obscuring smells, and so on. Besides, Dretske's account seems to presume that one in order to acquire a representation of a property one needs to be exposed to all or many of the instances of the property, while in practice agents normally acquire representations of general properties by being exposed to few or even one instance of it. In the above mentioned example, Mickey seems to be able to acquire a representation of any human's presence in response to the presence of only the one who feeds him, while by restricting the determination of content to Mickey's actual pattern of reinforcement, it seems impossible for him to acquire such representations.

2.3.3. Swampman and the Irrelevance of History

An additional problem that seriously challenges all historical-teleological theories of mental content (whether phylogenetic or ontogenetic) is posed by Donald Davidson's (1987) scenario of swampman As mentioned above, historical-teleological theories hold that the content of a mental state is determined by a function which is constituted by a history of selection; either by biological evolution or learning. However, Davidson proposes a scenario in which an object is created in an amazing sheer accident of the combining of materials which happens to be completely molecularly similar to us. For instance, lighting, hits and mixes elements in a swamp in such a way that it results in a swampman, who is molecule-by-molecule exactly like Donald Davidson. Consequently, because it/he has a structure molecule for molecule similar to Davidson, it/he behaves exactly like Davidson and seems to demonstrate rational and emotional behaviors just like him. Clearly, although the swampman has not earned its/his mental states through a history of learning or evolution, it/he behaves in such a way that it seems we need to ascribe beliefs and desires to it/him.

Dretske (1994) argues that entities like swampman cannot possess mental states, or at least mental states with content; i.e. mental states which are capable of representing the world in a sense that allows them to genuinely misrepresent it. He believes that scenarios like those of swampman are similar to cases in which, for instance, we might try to make money in our basement. The fabricated money might be molecularly just like genuine money, but it will not be genuine, because it lacks the right history behind it. Dretske and Millikan reject the view that entities like swampman have mental representations of the kind that can genuinely misrepresent

the world, because according to their historical-teleological views, the contents of a person's mental states are what they are in virtue of his or her history. Is it, however, acceptable for them to bite the bullet in this way and simply insist that, intuitions to the contrary, swampman lacks mental representations of the kind that can be genuinely mistaken?

There seem to be good reasons to insist that swampman does in fact have mental states that can represent the world in a way that allows for genuine error. In Davidson's scenario, swampman seems to be more complicated than a dollar bill. A swampman, by definition, would react and behave in the exact same sorts of ways, as a result of the exact same sorts of internal mechanisms, as us. If we hurt him, he would certainly act in a way that we would sympathize with. Since he has the exact same responses as us as a result of the exact same internal mechanisms, it is very difficult to deny that he feels pain, suffers, and desires not to have such experiences in the same ways we do. Moreover, it would seem that we can fluently communicate with him – he will seem to have the exact same conversational intelligence, as a result of the exact same sorts of internal states with the exact same causal propensities as us.[14] This also makes it very hard to think of him as a merely physical/non-mental entity.

In conclusion, it seems we are as justified in attributing psychological states like beliefs and desires to swampman as to ourselves, because its/his behaviors are to be just as well explained and predicted by such psychological states as ours; i.e. we can predict and explain what swampman does in terms of what he believes and desires just as successfully as we can predict and explain each others' behaviors. For example, admittedly, swampman will do some things, such as walk in certain path, stop at some stores, buy some things from them and so on. It seems we can best explain these behaviors by attributing beliefs and desires to it/him. This is exactly the same as the way that we would want to explain Davidson's behavior if were to do

[14] Some authors, like Block (1981) and Searle (1980) argue that certain entities might have the exact same behavioural dispositions as us but lack mental states. Block (1981) proposes a scenario of a blockhead which has the exact same dispositions to say what we would say in a conversation in response to any given history of conversation and input, but does so on the basis of an enormous pre-programmed lookup table (see also Braddon-Mitchel and Jackson's 2005 discussion of generalizing such a blockhead to all behavioural responses).
Perhaps our doubtfulness in recognizing these entities as mental is due to what Block (1978) earlier calls "internal architecture chauvinism" (see e.g. Braddon Mitchel and Jackson 2007 for an argument that this is so in the case of an enhanced version of Searle's (1980) water-pipe room connected to a robot). Although there are some controversies about whether these entities should count as having mental states, it should be vastly less controversial that swampman has mental states like ours, since his behaviour is produced by causal mechanisms (viz. neural states) exactly like our own.

these things; e.g. we would assign a desire to Davidson that he wants some bread, and a belief that there is a bakery at certain place, a belief that in order to get bread he should walk to there and so on.

2.4. Fodor's Asymmetric Dependence Theory

In order to provide a solution to the disjunction problem that avoids the problems with teleological theories, Fodor (1987) proposes his own causal theory of content which is known as Asymmetric Dependency Theory (ADT). The general idea of Fodor's account is that misrepresentation is dependent upon correct representation in a way that correct representation is not dependent upon misrepresentation. Consider a mental (or a neural) state in me, N, which represents cats. Consider cases where something other than cats cause N – e.g. my seeing a dog on a dark night might cause N because the dog resembles a cat. It seems that the dog's ability to be misidentified as a cat and cause N depends upon the propensity of cats to cause N. But cats' ability to cause N doesn't depend upon dogs on dark nights being able to cause N – e.g. cats would have the same tendency to cause N even if there were no dogs around, even if dogs were always easily distinguished from cats, and so on.

Fodor thus explains the distinction between misrepresentations and representations is terms of an asymmetric dependence of certain tendencies of things to cause a mental state; things which are not what a mental state represents versus things which *are* what the state represents,. Fodor proposes the following fine-grained version of his ADT: what it is for a state (e.g. a mental or neural state), N, to represent a thing, T (e.g. a kind or property), is for it to be the case that:

1. Ts tends to cause Ns,
2. For all Xs that are not Ts, if Xs tend to cause Ns, then Xs tendency to cause Ns depends upon Ts tendency to cause Ns. (That is, Xs would not tend to cause Ns if Ts did not tend to cause Ns), but
3. Ts tendency to cause Ns does not depend upon the tendency of Xs other than Ts to cause Ns (that is, Ts would still tend to cause Ns even if Xs did not tend to cause Ns).

ADT, as a rival to historical-teleological theories of mental content rejects the relevance of teleology in its core. However, one other significant idea in Fodor's theory which rejects the relevance of history in explaining the content of mental states is that the above mentioned asymmetric dependence is, in fact, a synchronic one and not diachronic. To illustrate, consider how we are supposed to understand the dependence of the X to N causal pathway upon the T to N causal pathway (and lack of dependence of the T to N pathway upon the X to N pathway). One might initially think, for instance, that the dependence of the X to N causal pathway is supposed to depend historically, or upon a previous (selected) T to N causal pathway, in which case this looks rather similar to a historical-teleological account, and then might face many of the same problems. For instance, in an online discussion, Fred Adams suggests that Fodor's theory has a problem with *semantic promiscuity* much like Ruth Millikan's theory in attributing genuine error to bacteria.[15] He considers a case in which antelope biting trees causes them to emit a distasteful chemical, as a result of biological evolution. Because of this, humans disturbing the trees can cause them to emit this chemical. Here, the human disturbance would not cause the emission of the chemical unless antelope biting did, but antelope biting would have caused the emission of the chemical even if human disturbances never came to do so. It might thus seem that Fodor's theory must declare this a case of genuine asymmetric dependence of the human to chemical pathway upon the antelope to chemical pathway, in which case the emission of the chemical represents antelope bites, and misrepresents human disturbances as antelope bites if the emission is caused by human disturbances.

However, as I already mentioned, Fodor (1987) clarifies that the dependence of the pathway from things the representation does not represent (Xs) to the representation (Ns) upon the pathway from things the representation does represent (Ts) to the representation (Ns) is supposed to be read *synchronically*. It is supposed to be a claim about the first path's dependence upon the second at the very time of the tokening of the representation, not an historical claim about how the first path came into existence due to the initial existence of the second path. That is, we are to read conditions 2 and 3 as claiming that:

[15] See Adams, "Fodor's Asymmetrical Causal Dependency Theory of Meaning" (http://host.uniroma3.it/progetti/kant/field/asd.htm).

2*. For all Xs that are not Ts, if Xs tend to cause Ns at time t, then Xs tendency to cause Ns at time t depends upon Ts tendency to cause Ns at time t. (That is, Xs would not tend to cause Ns at time t if Ts did not tend to cause Ns at time t), but

3* Ts tendency to cause Ns at time t does not depend upon the tendency of Xs other than Ts to cause Ns at time t (that is, Ts would still tend to cause Ns at time t even if Xs did not tend to cause Ns at time t).

Understanding Fodor's ADT in synchronic terms seems to enable it account to avoid semantic promiscuity. It might seem that while historical or diachronic asymmetric dependence is plentiful in nature, synchronic asymmetric dependence is rare, and is perhaps restricted to mental individuals. Besides, the synchronic asymmetric dependence in causal pathways seems to make it possible to attribute mental states to swampman. In other words, the newly made neural states in swampman's brain would have the same causal dispositions as Davidson's, so it seems plausible to consider them having the same synchronic asymmetric dependencies. However, Fodor's clarification that his ADT concerns synchronic asymmetric dependence, in some people's view, seems to be even more problematic. Adam and Aizawa (1994, 2017) observe that, while the diachronic dependence (or independence) of one causal pathway upon another seems at least easy to understand, synchronic asymmetric dependence actually seems at least as mysterious as representation (capable of misrepresentation) itself, and Fodor's claims about it seem to be unsupported by anything more than his intuitions about the actual content of mental representations. Thus, appealing to synchronic asymmetric of dependence (or independence) in explaining representational content seems to lead Fodor's ADT to be viciously circular.

As Adams and Aizawa (2017) explain, when we actually think about what is going on in the brain, there do not seem to be any genuinely synchronic asymmetric dependencies of the kind Fodor claims there are. The ability of all of the relevant entities to cause our neural states that are (or realize) our mental representations *at a given time* seem to be equally independent (or equally dependent) upon the ability of the others to cause the representation *at that very time*. Our tokening the representation of a cat, for instance, is "some set of neurochemical events. There should be natural causes capable of producing such events in one's brain under a variety of circumstances. Why on earth would [dogs on dark nights] be able to cause the neurochemical [representation of a cat at t] only because [cats] can [cause this representation at t, while the

ability of cats to cause this representation at t does not depend upon the ability of dogs on dark nights to cause it at t]?" The only explanation for an appearance of asymmetry seems to be the circular one, which appeals to intuitions about the content of the representation: "One might be tempted to observe that [N represents cats],…we associate [dogs on dark nights] with [cats], and that is why [dogs on dark nights] cause [N at t] only because [cats cause N at t]… This answer, however, involves deriving the [synchronic] asymmetric causal dependencies from [representational content], which violates the background assumption of the naturalization project [to explain representational content in non-representational terms].[16]

2.5. Interpretationism

The shortcomings of historical-teleological naturalistic theories of intentionality and problems with Fodor's ADT can naturally motivate a new theory about mental content called interpretationism (see especially Dennett 1981). According to this view, mental states and their contents are accounted for by their explanatory and predictive role instead of their atomistic causal, information, and historical relations to facts and objects in the world. A core idea of

[16] To the extent that we *can* make sense of synchronic asymmetric dependencies without circularly invoking representational content, Adams and Aizawa (1994) have another compelling objection, which is that the theory is either (i) semantically promiscuous, or (ii) reliant upon a more restrictive notion of synchronic asymmetric dependence, which *does* depend in a circular way upon the idea of representational content. Adams and Aizawa consider a hypothetical case of pigeon dropping synthesis, in which mad scientists synthesize matter chemically identical to pigeon droppings, and could not do this if pigeons didn't produce the droppings. Adams and Aizawa claim that this is a case where (i) pigeons tend to cause droppings, (ii) scientists tend to cause droppings, and this tendency depends upon that of pigeons to cause droppings, and (iii) pigeons' tendency to cause droppings does not depend upon that of the scientists. The threatening, semantically promiscuous upshot might seem to be that pigeon droppings represent pigeons (in a sense capable of misrepresentation) and misrepresent scientists as pigeons when scientists cause the droppings. But for reasons we have seen, this only works as an objection if the asymmetric dependence in (ii) and (iii) is synchronic, and although Adams and Aizawa flatly claim that it is, it is not very easy to imagine that this is so. It certainly sounds, from everything Adams and Aizawa say about the case, that the dependence is merely diachronic; that the scientist to droppings pathway came into existence because the pigeon to dropping pathway was already there for them to learn from.

To try to imagine (ii) and (iii) as cases of synchronic asymmetric dependence, we might try to imagine the following. Suppose that there is a pigeon god, Coosalcoatl, who only lets the synthesis go on because pigeons produce droppings and Coosalcoatl feels that the synthesis honours pigeons. If at any point pigeons were to stop producing droppings, Coosalcoatl would feel that the synthesis no longer honoured pigeons, and would step in and make the synthesis stop working. (We can also imagine Coosalcoatl occupying a standpoint outside of space and time, and thus having the ability to make dropping synthesis at t fail instantaneously if pigeon dropping production at t is to fail). To the extent that we can make sense of all of this, we might seem to have a genuine instance of synchronic asymmetric dependence, but it still seems mistaken to say that the pigeon droppings represent pigeons in a sense capable of misrepresentation, and that the pigeon droppings are themselves mistaken if they are caused by scientists. To avoid this conclusion, it seems that Fodor would have to invoke a special kind of synchronic asymmetric dependence, distinct from that in (ii) and (iii) given the presence of Coosalcoatl. It is difficult to see, however, how Fodor could explain this more restrictive notion of synchronic asymmetric dependence without invoking the notion of representational content, thus making his account viciously circular.

functionalism about mental states is that their role in explaining behavior is essential to them – e.g. what makes a state count as a belief is its role in combining with desires to produce behavior (and in combining with other beliefs to produce further beliefs, which may also combine with desires to produce behavior; see e.g. Braddon-Mitchell and Jackson 2006). Interpretationism goes further, and claims that mental states' role in explaining behaviour determines not only what kinds of mental states they are (e.g. that a certain state of my brain is a belief), but also what their content is or what they represent (e.g. that this state is a belief that there is a desk in my room).

2.5.1. The Intentional Stance

Dennett (1981) introduces the idea of an *intentional stance* in his argument that the attribution of mental states is part of the best way to predict and explain agents' behaviors. To explain Dennett's intentional stance, it seems helpful to start by explaining some other stances that he describes. Perhaps the most fundamental stance one can take in explaining and predicting phenomena is what Dennett calls the *physical stance*. Using physical stance, we can predict and explain physical systems' behavior merely by using the laws of physics; e.g. in explaining the time that a falling object would take to reach the Earth, we only need these laws and the distance of the object from surface. The physical stance is commonly used in physics, chemistry, biology, astronomy and other natural sciences[17]. Scientists use the physical stance to predict that the Andromeda galaxy and the Milky-way galaxy will collide each other about three billion years later, that the sun will become a red giant in about five billion years, and the like. In addition to these brutal physical facts, even some sorts of human behavior can be explained and predicted by the physical stance (or at least from a bio-chemistry point of view); e.g. in explaining why, when we get drunk, some irrational and unusual behaviors might happen, or predicting how quickly a person can move a stone with a certain weight and shape, etc.

Besides the physical stance, Dennett introduces the *design stance* as a way to understand the explanation and prediction of some of a system's behaviors by attributing certain functions to

[17] As Dennett describes it, the stance which is used in chemistry and mechanistic biology is same as that used in physics. However, in my opinion, there can be some real patterns in chemistry and certainly in (even mechanistic parts of) biology which will be missed by a stance which is used in physics. In my opinion, it would not be wrong to consider biological and chemical stances in addition to the physical stance.

the system or its parts.[18] For example, in explaining my phone's behavior when its lock key is pressed, we would use the design stance by holding that the design of my phone implemented by producer and programmers is to operate defined functions that dictate what it does. Although the design stance cannot straightforwardly be applied to explaining and predicting most human behaviors, it would not be right to think that the use of this stance is limited to merely explaining the behaviour of artifacts. In fact, we can consider the role of genes and what they code for in explaining and predicting many physiological features of individuals. Because they often abstract away from the fundamental physical details (and are often compatible with multiple fundamental physical pathways that achieve the relevant results, see e.g. Kitcher 1984) these explanations and predictions can hardly be viewed as the result of the physical stance, and they are actually coming from a design stance.

After introducing the physical and design stances, Dennett argues that there is also an *intentional stance*, which works by attributing mental states with intentional contents to agents. Dennett argues that in explaining the behaviour of certain complex systems, the intentional stance can work when design and physical stances are not practically accessible (or, as we will see, most enlightening). In order to explain a system's (or agent's) behaviors with this stance, we assume that the individuals will be in a minimal sense rational, and we can explain the system's (or agent's) behaviors by attributing beliefs and desires to the system (or agent). Dennett suggests that this works in accordance with what he calls the *intentional strategy*, in four steps: 1) consider the object as a **rational** agent which has some **purposes** (which are ultimate and underivative desires) and a place in the world (which naturally involves interacting with external factors and receiving certain kinds of sensory inputs), 2) figure out what **beliefs** this agent ought to have with regards to her purposes and her place in the world, 3) determine what **desires** this agent ought to have in a way similar to step #2, and finally 4) **predict and explain** what this rational agent would or will do to achieve her purposes in light of her beliefs and desires. As we can see, in Dennett's view, the content of an agent's beliefs is defined based on the content of her sensory inputs and desires. The intuition behind this idea is inspired by the ways we typically form beliefs. Suppose you (a rational person who has a place, purposes and desires) were walking down a street and your friend asks about the number of computer shops on the street.

[18] Of course, the proper functioning of the functions is assumed in explaining/predicting the systems' behaviors.

Having any idea about the number of computer shops depends on your interests in computer stuff (as well, of course, as whether you were able to get sensory information about these shops). If you are not into any of those things (or for some reason could not, e.g. see the shops), it is natural for you not to form any beliefs about them.

However, Dennett's interpretationist thesis as he explains it faces some questions. First, how can an agent possess a false belief? This question is similar to the problem of explaining how mental states can be capable of misrepresentation. For instance, if we see a dog on a dark night, how can Dennett's theory explain how we come to form the belief that it is a cat? More specifically, why would the intentional strategy, in order to gain the best explanation and prediction of an agent's behavior, attribute a false belief to an agent instead of claiming that the agent believes that the object is dog and that she just has a desire to behave in certain ways consistent with truly believing it to be a dog but behaving as she does (e.g. desiring to report falsely that it is a cat if we ask her)? Similarly, there is a question about how Dennett's view can explain the possibility of agents' forgetting some very important beliefs. Suppose I have a very important job interview appointment, however I simply forget and miss such an important opportunity. It certainly seems that this can happen, but how could the intentional strategy attribute to me the belief that I have a very important appointment, since it holds that the attribution of mental states should be based on what would help fulfill my desires given my information?

In response Dennett argues that beliefs with errors and mistakes are in fact based on true beliefs manipulated by hallucination, illusion, misperception, memory loss, or even fraud. As he puts it, "… the false beliefs that are reaped grow in a culture medium of true beliefs" (Dennett 1981). In fact, Dennett is appealing to a holistic view about how the intentional strategy should go about attributing mental states. To illustrate, in the case of forgetting an important interview, Dennett would posit that the best explanation of my overall pattern of behaviour was that I did have a strong tendency to form the belief that was counteracted by a memory lapse – since my tending to have the belief might help to explain such things as my scheduling the appointment, my taking some initial steps to keep it, and so on. In other words, on a holistic view of the intentional strategy, attributing such false beliefs gives a better explanation than supposing that

all of the agent's beliefs are true and interpreting her as having only those desires that would explain her behaviour on the assumption that all of her beliefs are true.

2.5.2. Real Patterns and True Believers

After explaining the intentional strategy, Dennett argues that using this strategy cannot be helpful in every context. Taking an example of a lectern, he observes that all behaviors of a lectern can be predicted and explained using the physical or at best the design stance, and attributing mental states to it via the intentional stance would be explanatorily redundant.

Here, as I discussed in the first chapter, Dennett himself considers the objection that the selection of a stance to gain the best explanation and prediction of a system's behavior depends upon our epistemic or cognitive limitations, and hence interpretationism might be viewed as some sort of implausible relativism or subjectivism about mental states. If what makes it true that a subject has mental states with certain contents is that these states play a role in an observer's explanation and prediction of her behaviour, then her mental states seem to exist only in the observers' interpretations, and there seem to be no mental states independent of what observers happen to find useful in explaining things. For instance, if super-intelligent aliens could explain our behaviour using only the laws of physics governing the constituents that compose us, interpretationism would seem to say that, relative to them, we have no contentful mental states.

Dennett (1981), however, argues that there is a version of interpretationism that avoids these worries by making what determines a state's status as a mental state with a particular content is something real in the world – in Dennett's (1991) words, there are some **real patterns** of behaviour, and explanations that fail to capture those patterns (like those of the super-intelligent aliens) are missing something perfectly objective.

As already discussed in the first chapter, these patterns consist in the similarities between different ways that the same outcomes could be realized, and enable us to predict and explain behaviors in a way that is objectively more efficient and also successful (with a pretty high degree of accuracy on the basis of radically less information.) To take another of Dennett's examples, consider how the intentional strategy as opposed to the Martians would explain and predict the behaviour of someone driving home from work. All the user of the intentional

strategy might need would be a conversation of this individual with her partner on the phone, like “Oh, hello dear. You're coming home early? Within the hour? And bringing the boss to dinner? Pick up a bottle of wine on the way home, then, and drive carefully." On the basis of only such information the user of the intentional strategy can predict with a high degree of reliability “that a large metallic vehicle with rubber tires will come to a stop in the drive within one hour, disgorging two human beings, one of whom will be holding a paper bag containing a bottle containing an alcoholic fluid” (Dennett 1981). As we discussed in the first chapter, to reach the same prediction using the physical stance, the Martians would have to use an extremely large amount of information and also make laborious calculations about specific neural events in humans, and also many other things involving the behaviors of non-human objects such as specific pathways of vehicles, and so on. Moreover, even with overcoming all these inefficiencies, the Martian explanation would still miss information about how the same results could have been achieved slightly differently – e.g. if the very specific neural events of the driver had been slightly different, if she had taken a slightly different route, and so on.

On Dennett’s (1981) view, what makes a state of a system count as a contentful mental state is its being disposed to play a causal role in a general pattern of behaviour that is captured by mental (e.g. belief-desire) explanations. So understood, the general interpretivist idea is that what makes it true that an agent has a mental state with a certain content is that she is disposed to behave in ways that would be best explained (in the sense of capturing a real, objectively-efficient-to-use pattern) by attributing to her mental states with this content. This view seems to capture very nicely our sense of why swampman should be deemed to have mental states with contents in the same “literal” sense that we have them: because his behaviour follows the same sort of extremely complex patterns as us (unlike, say, a lectern), the best (i.e. real-pattern-capturing, objectively-efficient-to-use) explanation of his behaviour unavoidably makes reference to mental states with the kind of content that ours can have.

To demonstrate why a lectern and a swampman should be discriminated, let us consider Dennett’s explanation of how genuine mental states can arise by adding the right sort of complexity to systems that lack them. Take his example of an ordinary thermostat, all of the behaviours of which can be adequately explained by the design (if not something like the

physical[19]) stance. If we choose, we can still attribute some beliefs to the thermostat such as that the room is colder or hotter than the set-point, and that the heater is on or off. We can also attribute desires, such as that if the temperature is colder (hotter) than the set-point that the heater be turned on (off). Notwithstanding, these belief/desire attributions don't seem quite the same as attributing them to us. Dennett's view is that these belief/desire attributions to the ordinary thermostat do not capture any real patterns beyond what is already captured by the design (or physical) stance. To see why this is so, we can argue that this thermostat in contrast to us is not plausibly viewed as having a concept of heat or cold or the like at all. To illustrate, if we replace the temperature sensor of the thermostat by a photo-cell, it can control the projection of light just as well as the temperature of a room. Similarly, if we replace it by a water-level meter, it can control the level of water tank. In fact, Dennett argues that, due to the indeterminacy of what this thermostat is responding to or controlling, we can conclude that there is no state of affairs for this system to determine what its beliefs and desires are about. In other words, such a system doesn't have the kind of rich, sufficiently complex causal dispositions that are best explained by attributions of beliefs and desires. Those beliefs which we attribute to systems like ordinary thermostats are thus useful only as a heuristic (which captures no real patterns other than the design or physical stances), or as Dennett calls them, are mere belief-like states. These states are very simple, so that we can easily change their metaphorical content by changing their inputs; i.e. their content is too indeterminate to be literal. Hence, Dennett concludes we can fully predict and explain the behaviors of this system using the physical or design stances, so that any attempts to explain its behaviors in terms of the intentional strategy would be redundant.

On the other hand, Dennett considers complicating the above discussed thermostat to explain how a true believer could emerge, so that its behavior could be best predicted and explained (in the real-pattern-capturing, objectively-efficient-to-use sense) only by the intentional stance. He suggests enhancing the thermostat by proving some other inputs such as an eye (such as a motion detector to sense the presence of people in the room), an infra-red eye (to scan for other possible sources of heat and cold in the room), an ear (such as a sound or voice detector to sense the occupants' bodies shivering due to coldness), and also the ability to receive

[19] As I mentioned in note 17, there are likely many real patterns above the level of fundamental physics other than those best captured by the intentional stance, some of which might help to explain the behaviour of such thermostats. The point here is simply that the intentional stance does not seem required to best explain (i.e. in a real-pattern-capturing, objectively-efficient-to-use) the ordinary thermostat's behaviour.

other regional information such as the outside temperature and humidity. Dennett also suggests adding some other outputs that the device can manipulate, such as giving it the ability to relight the furnace or the ability to purchase fuel when it runs out, etc. He also suggests equipping the device with some levels of reasoning such as inference and induction in order to make some new belief-like states or desire-like states out of its existing belief-like states and desire-like states. Consequently, the device will use its various belief-like states and also its desire-like states to guide a wide variety of different behaviors. For instance, suppose the thermostat has run out of fuel for the first time. Regarding its desire-like state which tends to cause it to fill the empty tank, it orders the fuel. However, in practice it takes some hours for fuel to arrive and consequently the temperature of the room significantly goes down. The enhanced thermostat spots occupants' bodies shivering. This information will combine with the desire of preventing the occupants from feeling cold. As a result, the thermostat will form a new desire-like state which causes it to seek to avoid running out of fuel in the future, which could cause it to perform some specific behaviors. The thermostat might order the fuel not right at the time that the tank is empty, but when it has depleted half of the fuel stock. Such a process of forming new desire-like and belief-like states can continue iterating so that the complexity of the system becomes increasingly greater. Dennett argues that eventually, as the complexity of internal mental-like states of the system gets greater and greater, some significant behavioral patterns will emerge so that the physical and design stances cannot give the best explanation or prediction for the system's behaviors anymore. To illustrate, as discussed above, there was no designed function of early ordering fuel in the enhanced thermostat. This function emerged out of the system's belief-like and desire-like states. As the thermostat continues working these kinds of new functions will emerge; the role of belief-like and desire-like states in explaining and predicting its behavior gets more significant. Eventually, the complexity in the enhanced thermostat system gets to a point where its behaviors are best explainable by the intentional stance.

To further see how this will be so, consider the complicated belief-like and desire-like states of the enhanced thermostat (which resulted from the system's transactions over the time) will, in contrast to those of the simple thermostat, be interpretable as having sufficiently determinate content to be understood as genuine beliefs and desires. E.g. the belief-like state of 'it's colder than set-point' in a simple thermostat would function in a way that it could be just as well interpreted as representing many other things (such as the position of the sensing

mechanism, or in slightly different environments the level of a water tank, etc.). But the belief-like state in an enhanced thermostat would, in light of its highly complex dispositions to tend to cause different behaviours in the presence of different sensory states, belief-like states, and desire-like states, be very hard to interpret as representing anything other than coldness. For suppose we try to adapt this system for instance to control the light of the room, similar to what we could do with a simple thermostat. It is obvious that in an enhanced thermostat there are a lot of desire-like and belief-like states which cannot do anything relevant to light; they are only about coldness, which they differentially tend to influence very differently in the presence of different other coldness-relevant sensory information, belief-like states, and desire-like states. Thus, as Dennett argues, by attributing the beliefs and desires to the enhanced thermostat, the intentional strategy is not giving a mere shorthand for what the design and physical stances explain or predict, but it is capturing genuine patterns in the behaviors of the system which cannot be discovered by those stances.

2.5.3. Interpretivism, Conceptual Role Semantics, and Success Semantics

As we discussed above, Dennett‘s version of interpretationism avoids the threat of subjectivism by holding that what it is for a system to have intentional mental states is for there to be a real pattern in its behavioural tendencies which is best explained by certain causal dispositions among its internal states. This does, however, diminish the difference between interpretationism and other naturalistic accounts of intentionality according to which the content of mental states can be explained in terms of their holistic, interacting causal propensities. Dennett (1981) stresses that the key to arriving at a genuine intentional system (such as a sufficiently enhanced thermostat) from a merely metaphorically intentional system (such as an ordinary thermostat) is “giving its belief-like [and desire-like] states more to do.” One part of this is giving it enough internal states with sufficiently flexible propensities so that they can interact with each other to produce new internal states in the ways characteristic of our beliefs and desires; as Dennett describes this it is necessary to “provide more and different occasions for [belief and desire-like states’] derivation or deduction from other states, and by providing more and different occasions for them to serve as premises for further reasoning.” A second part of this is giving it sufficiently flexible desire-like states so that they can combine with a wide

variety of belief-like states, and giving it sufficiently flexible belief-like states so that they can combine with a wide variety of desire-like states, to produce a wide variety of behaviour that would fulfill the desires if the beliefs were true. This kind of flexibility in the behavioural dispositions of belief- and desire-like states seems to be central to the real pattern captured by the intentional stance in the dispositions of the enhanced thermostat, which is absent in the ordinary thermostat, and in particular seems central to the determinacy of the interpretation of the content of the enhanced thermostat's states.

The first sort of content-determining causal relations among internal states that concerns their influencing other internal states is that held by naturalistic versions of *Conceptual Role Semantics* (CRS) to determine the content of mental states. According to naturalistic CRS, what it is for a mental state to have a given content is for it to play a certain causal role in an agent's cognition (Field 1977; Harman 1982; Block 1986). The second sort of content-determining causal relations among internal states that concerns their combining with each other to produce behaviour is that held by success semantics to determine the content of mental states. As I mentioned in the first chapter, success semantics holds that what it is for a belief to have a given content is for it to combine with the agents' desires to produce behaviour that would fulfill those desires if that content were true (and what it is for some of an agent's desires to have a certain content is for them to combine with the agent's beliefs to produce behaviour that would bring about that content if those beliefs were true).

I thus do not believe that we should view Dennett's interpretationism as a rival to naturalistic versions of conceptual role semantics and success semantics. Rather, we should view Dennett's ideas about the intentional stance capturing a real pattern that is best explained by the right kinds of causal relations among internal states as giving us an account of how something like conceptual role semantics or success semantics captures what we should want a naturalistic theory of content to capture, and fits mental states into a broader metaphysical account of what there is on which mental states are not fundamental but still genuinely exist. As I will argue below, we likely need both conceptual role semantics and success semantics to fill out Dennett's general account of the nature of the content of mental states, as each is needed to capture a kind of content.

2.6. Conceptual Role Semantics and the Need for a Distinct Theory of Broad Content

Naturalistic versions of Conceptual Role Semantics (CRS) thus hold that the content of a mental state is determined by the mental state's causal roles in an agent's cognition (Field 1977; Harman 1982; Block 1986). The most common and prominent causal roles considered by proponents of CRS are mental states' inferential roles. The idea that inferential roles determine the meaning or content of a mental state is very attractive in the case of explaining what it is for a mental state to have contents involving logical connectives. What, for instance, would it be for a mental (e.g. a neural) state N to have the content of the logical connective 'A or B'? Plausibly, this is for the agent to be disposed to infer a mental state with the content 'A' from N together with a mental state with the content '~B', to infer N from a mental state with the content 'A', to infer N from a mental state with the content 'B', and so on. CRS is also an attractive way to explain the content of mental states that involve complex concepts, such as BACHELOR. It seems that a contemporary concept of a bachelor is something like a male who is not in a relationship but is in a position to enter one (which explains, e.g. why priests and seven-year-olds are not bachelors, but technically married men who are separated, living alone, and waiting for their divorce papers to clear are). What, then, is it for a mental (e.g. neural) state N to have the content that 'X is a bachelor'? Plausibly, it is for one to be prepared to infer N from thoughts with the content 'X is male, not in a relationship, and in a position to get into a relationship,' to infer the thoughts 'X is male', 'X is not in a relationship' and 'X is in a position to get into a relationship' from N, and so on.[20] CRS can also give a convincing explanation of the content of mental states involving necessarily co-extensive concepts such as three-sided and three-angled. For instance, the difference between thinking that something is three-sided and thinking that it is three-angled seems to consist precisely in the difference between the inferences one must be disposed to make

[20] It is certainly true that these accounts of the content of one mental state makes reference to the contents of others. This might appear to be viciously circular, but as I will discuss in the next chapter one can use the Ramsey-Carnap-Lewis method of defining theoretical terms to show that it is not (cf. Lewis 1970).

to count as thinking about sides and the inferences one must be disposed to make to count as thinking about angles.

However, the inferential tendencies should not be understood in a way that allows no room for phenomena such as occasional forgetfulness, "failing to put two and two together," and more generally logical non-omniscience (so, for instance, if our mathematical concepts logically entail which interesting mathematical claims are theorems, one need not know more than a professional mathematician to count as possessing these concepts! See e.g. Block 1986 and Fodor 1987). The basic, plausible idea is simply that one could not completely lack dispositions to make the "close-in" logical entailments associated with a content like 'A or B' or 'X is a bachelor' and count as having thoughts with those contents. There are also different ways of approaching which exact inferential tendencies should count as determining the contents of mental states. A thought or claim is said to be *analytic* if it is true simply in virtue of its content or meaning, standard examples of which would include the thoughts that BACHELORS ARE MALE and IT IS RAINING OR IT IS NOT RAINING. A thought is said to be *synthetic* if its meaning alone does not settle whether it is true or false, standard examples of which would include the thoughts that BACHELORS ARE LESS LIKELY TO BE STRESSED THAN NON-BACHELORS and IT IS RAINING. Philosophers who are not afraid of drawing such an analytic-synthetic distinction and holding that there are analytic truths can hold that what it is for a mental state to have a given content is for the agent to be disposed to make inferences that follow *analytically* from that content.

Of course, there are some philosophers such as Quine (1951) who for some reason are skeptical about the possibility of drawing the analytic-synthetic distinction or the existence of genuinely analytic truths. In their view, CRS will entail that *all* of the inferences associated with a mental state will determine its content (see e.g. Fodor 1987). To illustrate, since the *total* set of inferences we are prepared to make with any one mental state does seem somehow to be influence by the inferences we are prepared to make with all of the others, it would seem to follow from this that a very radical form of *content holism* is true, according to which the content of any mental state is determined by the content of all of one's other mental states. From this, it would seem to follow that no two thinkers (and no one thinker at different times who changed anything about her beliefs) could ever have mental states with the same content. This might

appear to be a difficulty, but as Block (1986) and Fodor (1987) discuss, we could switch to talking about gradational similarities in content (understood as similarities in the inferences one tends to draw from a given mental state) in the place of strict identities in content.

As successful as CRS may be at explaining one kind of mental content, Ned Block (1986) argues convincingly that it must be supplemented by a theory of another kind of mental content. He contends that, although CRS can explain the *narrow content* of mental states, it is unable on its own to give an explanation of the *broad content* of mental states. Narrow mental content is the kind of content that depends only upon the thinker's internal states, while broad content depends as well upon her environment. To illustrate the distinction between narrow and broad content using the concept of WATER or the concept of I, the broad content is the profile of what that content picks out in each possible world. For instance, when I token the concepts I and WATER, these concepts pick out Meysam and H_2O in every possible world that I might think about, and these profiles of what they pick out in all possible worlds are their broad contents.[21] Similarly, the broad content of a thought such as THE LIQUID IN THE OCEAN IS WATER is its truth conditions, or its patterns of being true or false in each possible world. Thus, as thought by me this thought would be true in the actual world (and any possible world alike to it in this respect) and false in any world in which the ocean was filled with something else (or empty, or non-existent) – and this profile of truth or falsity in all possible worlds is its broad content.

The feature that distinguishes broad content is that it is not determined solely by internal facts about the possessor of that state. In fact, it often depends upon objective, external facts. Putnam's twin earth scenario seems helpful to understand this feature, and to motivate the existence of a distinct kind of content (namely narrow content). Assume that there is a twin earth which is identical to ours with the exception that H_2O is replaced with XYZ (which is a different chemical structure that has the same superficial properties as water and plays the same practically important roles). On twin earth, there is also a person who is my twin and is as similar to me as possible — call him Schmeysam — who possesses exactly the same internal states as me. When I think I, WATER, or that THE LIQUID IN THE OCEAN IS WATER, the broad content of these thoughts is respectively Meysam, H_2O, and truth in all worlds in which the

[21] The broad content of referring expressions is not always invariant across possible worlds; e.g. non-rigidified descriptions, such as 'whatever mountain on earth happens to be the tallest' refers to Mt. Everest in the actual world, but would refer to K2 in a world that was as much like the actual world as possible but had no Mt. Everest.

oceans contain H_2O (and falsity in all others). However, although my twin's internal states are identical with mine, their broad contents will not be the same as mine; they will be, respectively, Schmeysam, XYZ, and truth in all worlds in which the oceans contain XYZ (and falsity in all others).

As discussed above, there is a kind of content of mental states which makes the mental states of my twin's and mine have different contents. However, besides this sort of content, there also seems to be a kind of mental content that is the same between my twin and me. First, it seems there is a kind of mental content responsible for a mental state's causal powers, where two mental states with the same contents will have the same causal powers (see e.g. Fodor 1987). In fact, the states of Schmeysam and I seem to have the same causal powers: they both direct us to interact in the same way with the local watery stuff. Second, there seems to be a kind of content that is (under the right conditions) introspectable, in that one can determine what one believes and desires simply by introspection; and unlike the broad content of the mental states of my twin and mine, what my twin and I introspect seems to be the same thing. We cannot tell simply by introspection to which specific chemical structure our concept WATER refers. As Putnam discusses, the concept of water (when thought by people on earth) always referred to H_2O, even prior to the time at which we had a chemical theory in which to describe H_2O, and also prior to many agents' learning such a theory (e.g. the concept of water as thought by people who have never learn chemistry can refer to H_2O as well.) Similarly, there seems to be an important kind of mental content that is bound up with determining which thoughts are trivially knowable (or at least knowable by analyzing one's concepts and thinking about what theoretical role one should want them to play) and which thoughts take substantive empirical work to evaluate. For instance, on the broad way of individuating contents, my thought that WATER IS H_2O has the same content as my thought that WATER IS WATER (and as thought by Schmeysam WATER IS XYZ has the same content as WATER IS WATER). But WATER IS H_2O took empirical work to verify, and can be reasonably doubted in the absence or ignorance of such work, while WATER IS WATER is trivially true. It thus seems that, in addition to broad content, there is this second kind of content, referred to as *narrow content*, which depends only upon one's internal

states and plays a central role in explaining one's thoughts' causal powers, introspectable properties, and roles in reasonable inference.[22]

As I mentioned above, Block argues that although narrow mental content could be explained by CRS, broad mental content cannot be convincingly explained in this way. Authors such as Harman (1982) do attempt to use CRS to explain mental states' broad content, or truth and reference conditions. Such views employ what Block (1986) calls "long arm" causal roles of interacting with the environment (in addition to the "short arm" causal roles of interacting with other internal states to explain narrow content). But as Block argues, by making content depend upon such external factors, CRS loses its ability to explain narrow content's central causal and epistemic features, and faces the problems faced by a theory of reference and truth conditions without offering any distinctive solutions. Most relevant for our purposes, by invoking 'long arm' causal roles of being caused by and causing certain things in the external environment, Harman's version of CRS immediately faces Jerry Fodor's disjunction problem. A mere causal connection to the environment cannot simply determine a mental state's broad content, i.e. we cannot simply hold that mental states refer to whatever reliably causes them, as this would make error impossible (or at the very least offer a radically inadequate account of error, since mental states can be reliably caused by things they misrepresent, such as dogs on dark nights causing us to token a representation of cats).

Harman can, of course, replace the crude causal theory with another theory of the determination of the broad content of the kind we have discussed. He could replace it with a teleological theory (although we saw problems with that), Fodor's asymmetric dependence theory (although we saw problems with that), or perhaps best the success semantical theory that for a belief to have the content that P is for it to tend to combine with desires to cause behaviour that would fulfill those desires if P were true (and that, for at least some desires, for a desire to have the content that S is for it to tend to combine with beliefs that cause behaviour that would bring about S if those beliefs were true). However, even if he succeeds in providing a broad content theory to add to CRS, he has actually invoked something very different from the causal roles between internal states that in a theory like Block's can so convincingly explain what

[22] One might object that narrow mental content is not actually a kind of content due to its lack of reference and truth conditions. Block (1986) suggests that if one wishes to use the term 'content' in this way, then he simply can replace references in the text to 'narrow content' with 'narrow determinant of content'.

makes it the case that a mental state has a given narrow content. Thus, Block concludes that we should understand such attempts to give a successful naturalistic account of what makes it the case that a mental state has the reference and truth conditions it does as supplementing CRS's explanation of what makes it the case that a mental state has the narrow content it does with a distinct theory of broad content. Block calls such an account a "two-factor" theory, with CRS explaining the narrow factor of content, and something else explaining the broad factor of content. In the following chapter, I will argue that the success semantics is the best naturalistic theory of broad content, and thus the best candidate to serve alongside conceptual role semantics in a two-factor theory.

Chapter Three

Success Semantics: Problems and Solutions

After delineating the problems and shortcomings with some attempts to naturalize the intentional content of mental states in chapter 2, in this third chapter I will try to demonstrate (a) what success semantics offers as a significant resource to those interested in naturalizing intentionality, (b) what the main problems for success semantics are, (c) what responses and solutions the proponents of success semantics have given to these problems, and finally (d) how I think that problems for success semantics can successfully be solved.

3.1. Success Semantics and its Problems

As I suggested in the last chapter, it seems that to provide a persuasive account of the content of mental states, we should follow Dennett's ideas in looking to the explanatory role of these states. We discussed that, in order to avoid the threat of subjectivism, we need to understand this explanatory role as a real pattern of the causal dispositions of intentional states to interact with other states to generate behaviours. By investigating the nature of these causal dispositions, we came to articulate two kinds of contents for these states; 1) narrow content, and 2) broad content. Narrow content is content which depends only on the agent's internal states, and might be adequately captured by conceptual role semantics [CRS], which holds that a mental state's intentional content is determined by its tendencies to be caused by and cause other mental states. However, as Block (1986) argues CRS cannot adequately capture the whole picture of the content of mental states, and we need a distinct theory to explain the broad content of mental states; i.e. the kind of content which depends upon the agent's environment as well as her internal states. As theories which are best construed as naturalistic theories of broad content, we already discussed Millikan and Dretske's historical-teleological accounts and also Fodor's ADT. However, we saw how the historical-teleological accounts face serious problems such as the indeterminacy of content, the irrelevance of history, and an inability to account for swampman's mental states. We also saw how Fodor's ADT invokes a notion of synchronic asymmetric

dependence which is even more mysterious than intentionality itself. My suggestion, following authors such as Bermudez (2003), is that success semantics yields the best account of broad content in a "two factor" theory.

It is commonly believed that basic ideas of success semantics can be traced back to Frank Ramsey's works (1927). However, as a fine-grained theory of content, success semantics was introduced by J.T. Whyte (1990, 1991). Whyte (1990) begins to introduce his idea of success semantics by considering why the truth of our beliefs matters. In general, we want our beliefs to be true because the truth of our beliefs causes the **success** of our actions. Going further, Whyte suggests that, if we understand the success of an action as the fulfillment of the desires motivating it, we can use tendency of beliefs to combine with desires to make our actions successful to explain the intentional content of beliefs. We can put the general idea of success semantics as: what makes a state of an agent a belief that P is its tendency to combine with desires to produce behaviour that would fulfill those desires if P were true. For example, what makes a state of me a belief that there is a desk in my room is that the state is disposed to combine with my desires (e.g. to have exactly one desk in my room) to cause actions (e.g. declining an offer of what would be a second desk) that would fulfill those desires if there were a desk in my room.

In this general understanding of success semantics, it is left open exactly how strong is the relationship between the obtaining of a belief's content and the success of the actions that the belief tends to cause. Whyte takes this relation to be one of *guaranteeing* success. Whyte claims that "truth just is the property of a belief that **suffices** for your getting what you want when you act on it." This understanding of a belief's truth brings him to put forward following account of what it is for a state to be a belief with a given broad content, or truth condition (which Whyte calls (R) after Ramsey)

> "(R): A belief's truth condition is that which **guarantees** the fulfillment of any desire by the action which that belief and desire would combine to cause."

It might immediately be noticed that (R) explains the intentional content or truth conditions of beliefs in terms of the intentional content or fulfillment conditions of desires. This would seem to be a problem with (R) as a guideline to a fully naturalistic theory of the

intentional content of psychological mental states. This issue has been addressed by Whyte (1991), who argues that for every agent there is a set of simple or immediate desires the fulfillment conditions (or content) of which do not depend on the content of any of her beliefs. To put his proposal roughly: what makes it the case that a state of an agent is a simple or immediate desire that S is that S would "satisfy" the state, in the sense that S would make the state "go away" or stop influencing the agent's conduct, in a way that would reinforce a disposition to engage in similar behaviour if the state was manifest in the agent in similar circumstances. Whyte suggests that (R), accompanied by this account of the content of simple or immediate desires, can give a fully naturalistic account of the content of mental states. In what follows, I will try to explain these ideas in more detail.

As another immediate question about (R), someone might ask how it is supposed to work when we notice that a belief's contribution to causing an action depends not only upon the agent's desires, but upon on the content of her other beliefs as well. To illustrate, consider my desire D1 that: I have some bread, and my belief B1 that: store close to my home is a bakery. In practice, what leads me to do things like walk to the actual location of the store, enter it, and buy bread, is not only B1 together with D1, but in fact other beliefs of mine such as B2: the store is open 8AM to 6PM except Sundays, B3: it is 9AM Friday now, B4: bread costs around $3, B5: I have at least $3, and so on. Whyte (1990: 151) in response claims that (R) must be understood as explaining the truth condition for the conjunction of all of an agent's beliefs that combine with her desires to guarantee their fulfillment. According to this, what makes a state of me, namely B1&B2&B3&B4&B5&…, the conjunction of beliefs such as 'the store close to my home is a bakery', 'the store is open 8AM to 6PM except than Sundays' and so on, is its tendency to combine with my desires (e.g. my desire for bread) to cause behavior (e.g. walking to the store, entering, choosing a bread and paying, etc.) that would guarantee the fulfillment of those desires if those beliefs were true. Next, in order to identify an individual belief's content, Whyte (1990: 151) suggests that, we can take what is common to the success conditions of every possible set of beliefs with which it could combine to cause behaviour that would guarantee the fulfillment of the agent's desires. Thus, in order to specify the content of B1, we need to consider not only the success conditions of actions caused by B1 together with B2, B3, and so on, but also B1 together with all sorts of other beliefs. To illustrate, we may have another desire D2 that can combine with B1 to cause behaviour, such as my desire to learn the recipe of bread. B1 will play a role

here, but instead of combining with B4 and B5 it combines with B6: my friend works at the store, B7: people who work at bakeries know better than many people about bread recipes, and so on, which cause me to call him and talk to him. We continue on, for every possible set of beliefs with which B1 could combine to cause actions that would guarantee the fulfillment of my desires. We then look to see what is common to (i) what guarantees the fulfillment of my desire for bread when combined with B1, B2, B3, ….; (ii) what guarantees the fulfillment of my desire to learn the bread recipe when combined with B1, B6, B7, …. and so on. This will be its being the case that the store close to my home is a bakery close, which is thus the content of B1.[23]

Whyte's success semantics, in characterizing the content or truth condition for each belief, makes reference to other beliefs (and desires). Someone might, owing to its holistic approach, accuse Whyte's theory of involving an ineliminable or even vicious form of circularity. In response, I believe that we can reject this worry, by understanding Whyte's theory as an application of what many call the "Ramsey-Carnap-Lewis" method of what it is for something to have a particular feature in terms of its role in a general theory (see e.g. Lewis 1970; Braddon-Mitchell and Jackson 2007). To illustrate this method, suppose we want to explain what it is for something to be a state described in terms of a theory; here it would be a state's being (without loss of generality) a belief B1: that the store close to my home is a bakery, described in the intentional terms of commonsense psychology, or its essence as analyzed and characterized by Dennett 1981 and Whyte 1990; 1991. First, we take everything that the theory says about something that has the state; here, this would include 'my possible beliefs B1, B2, B3, … to tend to combine with my possible desire D1(desire for bread) to cause behaviour that would guarantee d1(=my having bread) if b1(=the store close to my home is a bakery)&b2&b3…'; 'my possible beliefs B1, B6, B7, … combine with my possible desire D2 (desire to get the recipe) to cause behaviour that would guarantee the d2 if b1&b6&b7…' and so on (a complete version of which would include desires having the features that Whyte

[23]Interestingly, Whyte (1990: 155) suggests that this method can also be used to identify the contents or reference conditions of individual concepts: "If an entity E occurs in the truth conditions of all beliefs that include a concept C, then C refers to E. For example, if one of [B1's] sub-propositional components, C1, is also a part of other beliefs, all of whose truth conditions involve [the bakery] but have nothing else in common, then C1 refers to [the bakery]." Thus, while theories like causal / informational accounts, teleological accounts, and Fodor's asymmetric dependence account seek to identify the contents of sub-propositional representations and then work its way "upward" to the contents of propositional representations like beliefs and sets of beliefs, success semantics due its holistic nature begins by identifying the contents of sets of beliefs, and then works its way "downward" to the contents of individual beliefs and sub-propositional representations.

characterizes as essential to their having the fulfillment conditions they do, which we will explore below). Then, we form a sentence that says that there are things that play above mentioned roles described by the theory (which is known as the "Ramsey sentence" of the theory); here this would be:

> Whyte's Ramsey Sentence: (∃x1) (∃x2) (∃x3)… (∃x6) (∃x7)…(∃y1) (∃y2)…[x1, x2, x3, … tend to combine with y1 to cause behaviour that would guarantee d1 if b1&b2&b3&…; x1, x6, x7,… combine with y2 to cause behaviour that would guarantee d2 if b1&b6&b7&…]

Finally, we identify something's being a state describe by the theory as (i) the holding of the Ramsey sentence, which says that there are things that play the relevant theoretical roles, and (ii) the thing plays the role associated with the relevant state. Here, we would explain what it is for a state to be a belief that the store close to my home is a bakery with (i) Whyte's Ramsey Sentence, and (ii) the state is identical to x1.

Success semantics has many advantages as a theory of broad content. First it has the general advantages of the causal-tendency approach suggested by a "real-pattern" understanding of Dennett's interpretationism. Second, as a dedicated theory of mental states' reference or truth conditions, it allows their narrow content to be explained by a different theory (as Block suggested, by Conceptual Role Semantics). Success semantics offers a straightforward way to explain misrepresentation; misrepresentation or error is what happens when a state tends to cause (or is part of a state that tends to cause) actions which would have been successful (i.e. desire fulfilling) if something were the case but in reality that thing is not the case. To illustrate, when I see a cat on a dark night and I form a false belief that it is a dog, the belief tends to combine with my desires (e.g. to offer her a food that she will like) to cause behaviours (e.g. providing Evolution dog food, instead of Ami Cat cat food) that would fulfill those desires if she were a dog, but in reality she is a cat. Because success semantics attributes content to the mental states on the basis of their causal tendencies, it does not require the relevance of biological or learning history. This lets success semantics attribute to swampman's states the same sort of content as it would attribute to our mental states due to his having the same sort of causal tendencies as us. This account does not (like Fodor's) invoke a notion of synchronic asymmetric dependence which, as we discussed in the previous chapter, is at least as mysterious as mental content and is

arbitrarily attributed to fit our intuitions about content. The theory relies simply upon the notions of causal tendencies and which behaviours would bring about which outcomes if certain conditions were to obtain.

Finally, success semantics seems to give a plausible account of what it is for a state to have a certain content. It illuminates Dennett's idea about how mental states with genuine content emerge with sufficiently complex, flexible dispositions. It seems plausible that what is it for a state to be a belief that P is for it to guide us to act in a way that would succeed (i.e. fulfill our desires) if P were true. This offers a credible account of what is it to have states that can themselves be in error, so that not only can they fool others (like tree-rings) but can themselves be fooled. It explains how, for us to say determinately that one believes some particular thing, we must be able to talk about what is common to what it would lead one to do in combination with many other possible beliefs and desires. Because rocks, tree rings, and bacteria are incapable of having such different, flexible beliefs and desires, this seems to be a convincing explanation of they do not have states that represent specific features of the world in a sense that allows for genuine misrepresentation.

Despite these considerable advantages, Success Semantics does face several problems. In the remainder of this section I will explain these problems, and the main responses to them that proponents of Success Semantics have offered.

3.1.1. The Problem of Insufficiency

The first problem with Success Semantics as characterized by Whyte has to do with (R)'s idea that beliefs that P combine with desires to cause behaviour that would *guarantee* the fulfillment of the desires if P were true. The problem, which we can call *the problem of insufficiency*, is that the truth of beliefs which combine with a desire may not be sufficient for the fulfillment of the desire. Blackburn (2005) attempts to present a version of this problem for success semantics when he argues that an agent simply might have some other false beliefs which could cause him to fail in fulfilling the targeted desire. To illustrate, suppose I truly believe that I have a laptop, which combines with my desire to work with it to cause me to go home. However, while my belief that I have a laptop is true, it turns out that I also had a false

belief that my laptop is at home, when in actuality my laptop is at my friend's home and I have forgotten about this, so my behaviour of going home does not in fact fulfill my desire to work with my computer. Therefore, my true belief that I have a laptop combines with my desire to work with it to cause the behaviour of going home, but this does not guarantee the fulfillment of my desire.

However, this version of the problem of insufficiency does not seem to succeed against Whyte's holistic account of the content of individual beliefs, because a belief's content in Whyte's view is derived from the content of conjunctions of all of the beliefs that motivate an action. As already discussed, in order to explain the content of individual beliefs, Whyte begins with an explanation of the content of *all* of the beliefs on which an agent acts as the guaranteed success condition of the actions that they all, together with the agent's desires, motivate. For instance, in the above mentioned example, in order to capture the content of B1: my belief that I have a laptop, we need to first consider it together with all other relevant beliefs such as B2: my belief that this laptop is at home, which combine with my desire to work with my computer and cause my behaviour of going home. It seems that it *would* be sufficient to fulfill my desire to work with my computer if B1 (that I have a laptop) *as well as* B2 (that my laptop is at home) (together with the rest of my beliefs) were true. On Whyte's theory, the content of B1 is not actually identified directly as what combines with my desires to cause behaviour that would guarantee their fulfillment if B1 were true. It is identified as what is *common* to the various sets of beliefs in which it participates which combine with my desires to cause behaviour that would guarantee their fulfillment if they were *all* true.

A more serious version of the problem of insufficiency is presented by Robert Brandom (1994), who argues that even if all of an agent's beliefs are true, the actions they motivate are still not guaranteed to fulfill her desires. This is because it is still possible that the agent is ignorant of some fact that can cause her to fail in fulfilling her desires. For example, in the case of my desire of working with my laptop, my beliefs about the laptop may be true—e.g. that it is sitting in my room—but I may be unaware that it has already been broken, e.g. by an electrical shock. As we can see, while I have no false beliefs, my beliefs may combine with my desire (to work with my laptop) to produce the behaviour of going home and trying to turn it on, which does not fulfill my desire.

In response to this problem, Whyte (1990,1997) argues that the full set of beliefs on which an agent acts must include *no-impediments belief*. Whyte appeals to the idea that it is impossible to act unless one believes that one's action will be successful (e.g. in the sense that, if one desires S, one believes that b1&b2&b3…, and this causes one to do A, then one must believe that 'if b1&b2&b3…, then A will bring about S'). So in addition to whatever else they believe, agents must always have an additional belief that they face no impediments. In fact, Whyte's idea will convert all sorts of cases in which we might initially think that agents fail only due to ignorance to cases in which agents fail due to some false beliefs. In the case of the broken laptop, Whyte would say that I fail to fulfill the desire of working with my laptop, because in acting on my ignorance about the viability of laptop, I am actually acting on a false belief that by, say, turning it on and typing I face no impediments to these things fulfilling my desires. So, if the *full* set of beliefs on which I acted were true, including my no impediments beliefs, my actions would be guaranteed to succeed in fulfilling my desires.

Nevertheless, important objections to Whyte's insistence that agents must always act on such no-impediments beliefs have been raised by Brandom (1994), Hattiangadi (2007), and Nanay (2013). First, these authors deny the idea that it is impossible to act unless one believes that one's action will be successful or that, if one's other beliefs are true, one's act will bring about the content of one's desires. It is plausible that it is impossible or incoherent to do A if one believes that doing A will *not* bring about the fulfillment of the motivating desire. But this is not inconsistent with one *not* positively believing that doing A *will* bring about the fulfillment of the desire motivating it. For instance, one might not be completely sure or might even suspend judgment about the success of an action, but still quite coherently perform it. Thus, Whyte seems to lack a plausible motivation for thinking that agents must always be acting on no impediments beliefs.

Second, as Brandom (1994) and Hattiangadi (2007) show, no-impediments beliefs can seem to have a trivial content. They seem to assume that the exact same no-impediments belief is operative in each case of an agent's acting. However, according to Whyte's holistic account in understanding the content of an individual belief, we need to take what is common to the success conditions of every action it motivates. Since the unique no-impediments belief would motivate all actions, Whyte's account makes its content a trivial one which obviously cannot entail, as

Whyte was hoping, the success of any individual action in conjunction with the truth of all of the agent's other beliefs. However, as Whyte (1997) and Dokic and Engel (2002; 2005) argue, instead of a unique no-impediments belief, for every context we can consider a different relevant no-impediments belief. For instance, Whyte (1997) suggests that, if the other beliefs motivating me to do A on the basis of a desire to bring about outcome O are B1, B2, …, Bn with contents b1, b2, …, bn, then what I must believe in doing A is the conditional belief that 'if b1&b2&…&bn, and I do A, then O' (e.g. if the store close to my home is a bakery, and the store is open 8AM to 6PM except than Sundays, and …, then if I walk to the location of store, enter there, choose a bread, pay for it, etc., then I will obtain bread). Dokic and Engel (2002; 2005) similarly argue that the relevant beliefs are instrumental beliefs, of the form 'if I do A, then O' (e.g. if I walk to the location of store, enter there, choose a bread, pay for it, etc., then I will obtain bread).

Nanay (2013) worries, however, about the redundancy of attributing such conditional or instrumental beliefs to agents, especially since, as we have seen, it seems sufficient for agents to act that they simply not believe the negation of the contents of such beliefs. However, Dokic and Engel (2002; 2005) argue that there are highly principled reasons to attribute such instrumental beliefs. First, they claim that it is plausible that such instrumental relations figure into the contents of what agents perceive in the form of 'perceptual affordances' – e.g. a door appears as if it can be opened, and one's surprise when it does not open indicates that one has been under an illusion or misrepresentation. Second, they claim that considerations analogous to those which support a principle of epistemic closure (according to which, if one knows that p, and q entails that one does not know that p, then one at least implicitly knows that ~q) support an analogous principle of pragmatic closure, according to which, if one intentionally does p, and q implies that p will not succeed, then one at least implicitly knows that q is not the case.

A full evaluation of Dokic and Engel's (2002; 2005) grounds for attributing instrumental beliefs to agents (the truth of which would, together with that of their other beliefs, be sufficient to guarantee the success of the actions they motivate) is beyond the scope of this thesis. It is worth noting that Dokic and Engel spend more of their efforts sketching possible defenses of the idea that true beliefs guarantee or suffice for success than actually arguing that these defenses are sound. In my opinion, there is room to doubt that any - or at least all - such instrumental relations

figure into the contents of one's perceptions. For instance, even if doors appear to be such that they can be opened, the location of the store does not seem to appear to be such that it will take me to bread. Similarly, there is at least as much room to doubt Dokic and Engel's principle of pragmatic closure as there is to doubt the analogous principle of epistemic closure. For a stock example of a reason to doubt the latter, it seems that I know that I have hands, but it also seems that I do not know that I am not a brain in a vat who might or might not have hands (for this and much more, see e.g. Luper 2016). Perhaps Dokic and Engel's grounds for attributing the relevant instrumental beliefs can ultimately be defended, and if they cannot, perhaps alternative grounds for such attributions can be defended. But it is open to doubt whether any such grounds will succeed, and the condition that true beliefs genuinely guarantee success seems extremely strong. As such, it seems worth exploring versions of success semantics that do not require beliefs to combine with desires to cause actions that strictly guarantee the fulfillment of the desires.

3.1.2. The Problem of Non-Necessity

Another problem that challenges the characterization of the relationship between actions motivated by true beliefs and success is that true beliefs do not seem to be necessary to the success of our actions. We can call this the *problem of non-necessity.* Godfrey-Smith (1994) and Hattiangadi (2007) argue that true beliefs are not necessary for the fulfillment of a desire that relies on them. As Hattiangadi says: "Just as we can fail through ignorance, we can also succeed through good fortune." To illustrate, suppose I am away from my home and I falsely believe that I have not already brought a pot home to place on my desk. This combines with my desire to have exactly one pot on my desk and no pots anywhere else at home to cause me to buy a new pot, when, in fact, I had already purchased a pot and placed it on my desk. Now, imagine that, before I arrive home, a thief comes and steals the pot that I had previously bought. In this scenario, I might happily arrive home and place the new pot on my desk—thereby successfully fulfilling my desire to have exactly one pot on my desk and no pots anywhere else at home—despite the fact that the belief that prompted this successful action was, in fact, false.

In response to this objection, Whyte (1997) argues that, even if false beliefs occasionally lead to success, for any false belief, there will be some desires and also some true beliefs with which it would combine to cause behaviour that would not fulfill the agent's desires. In other

words, although false beliefs may sometimes cause successful actions, only true beliefs can 'guarantee' their success. However, as Hattiangadi (2007) mentions, this distinction between true and false beliefs' relation to successful action seems hostage to the success of Whyte's idea that true beliefs combine with desires to cause behaviour that genuinely guarantees the fulfillment of the desires (i.e. his proposed solution to the problem of insufficiency). As we have seen, there are reasons to doubt the truth of this view, and to seek to find a version of success semantics that does not hold that desires combine with true beliefs to cause behaviour that is sufficient to fulfill the desires in every case. But as Hattiangadi observes, if the word "guarantee" is supposed to mean something weaker than sufficiency for the success of every action, then this notion of a guarantee is in need of further clarification.

3.1.3. The Problem of Circularity

As I mentioned earlier, Whyte himself addresses another problem with success semantics having to do with the contents or fulfillment conditions of desires. We already discussed how the basic idea of success semantics (e.g. as articulated by R) explains the content of beliefs in terms of the fulfillment of desires, while a desire itself is an intentional state in the first-place. If success semantics is to be a naturalistic theory of broad content, it has to explain the contents of beliefs (and desires) in terms that are ultimately non-intentional. That is, as a naturalistic theory of content, success semantics must ultimately avoid using any intentional ingredients in its account of intentionality.

All would be well if we could explain the contents of desires in terms that do not reference those of beliefs. Then the contents of beliefs could be explained in terms of the contents of desires, which would ultimately get a non-intentional explanation. However, what Whyte (1991) proposes as the most natural way in explaining the content of desires does seem to make reference to the contents of beliefs, namely:

> (F): a *desire's* fulfillment condition is that condition which is guaranteed to result from any action caused by that desire, if the *beliefs* with which it combines to cause the action are true.

Just as the rough, basic idea behind R, or the success semantical account of the contents of beliefs, is:

> (SS_B) what it is for a state of a system to be a *belief* that P is for the state to be disposed to combine with some of system's *desires* to cause behaviors that would satisfy those desires if P were the case (i.e. if that belief were true),

The rough, basic idea of F, or a natural success semantical account of the contents of desires, is:

> (SS_D) what it is for a state of a system to be a desire that O is for the state to be disposed to combine with some of the system's beliefs to cause behaviours that would bring about O if those beliefs were true.

Here, we can clearly see how R / SS_B is using desires' fulfillment conditions to explain the contents of beliefs, and F / SS_D is also presupposing beliefs' truth conditions to explain the fulfillment conditions of desires; which makes the account circular.

Regarding this problem, some people might argue that the case of circularity here is not actually a vicious one. For instance, Dokic and Engel (2002: 62-4) argue that the above mentioned circularity is not vicious as long as one has good independent evidence upon which to attribute either desires or beliefs. However, in my opinion this solution would degrade success semantic to a mere epistemic principle for attributing beliefs and desires, and frustrates its ambitions to be a metaphysical theory of what it is for mental states to have the contents they do (which, one might have thought, can ultimately help us determine what should count as good evidence about the presence of mental states with certain content in the first place). Similarly, someone might refer to the above discussed Ramsey-Carnap-Lewis method of explaining theoretically identified entities to argue that SS_B and SS_D explain the content of beliefs and desires holistically without actually involving a vicious form of circularity. The rough idea here would be that the Ramsey sentence of SS_B and SS_D together says something like: there are two sorts of states; the first combine with the second to produce behaviour that makes the second have property F when the first have property T. We would then try to explain what it is for a state to be a particular belief or desire in terms of its being a particular one of these states, and the Ramsey sentence holding. However, the main problem with this attempt to use the Ramsey-Carnap-Lewis approach is that it does not provide us with a sufficiently substantial view about

the content of beliefs and desires. All we can say, on this view, is that there are two kinds of states which are disposed to interact with each other and which produce behavior that fulfills the content of one only if the content of the other is true. But SS_B and SS_D alone do not seem together to give us enough content to understand what property of fulfillment or property of truth they are, and thus they do not give us enough content to understand what it is for one of the states to be a belief or a desire with a particular content. For instance, why should we understand the first sort of states as beliefs, the second as desires, the first property as truth, and the second property as fulfillment – as opposed to the other way around (i.e. the first sorts of states are desires, the second beliefs, T is the property of being fulfilled, and F is the property of being true)?

Whyte (1991) thinks that success semantics should avoid the circularity of an account that includes only R (SS_B) and F (SS_D) by supplementing the theory with an independent account of the fulfillment conditions of certain basic desires. He suggests that we can find a way to explain the content of these desires independent from being their disposed to combine with beliefs to produce behaviors that would bring about certain outcomes if the beliefs were true (as mentioned by F). Whyte is inspired by Bertrand Russell's idea that fulfilling (at least some) desires is psychologically satisfying in the sense that when you get what you want, your desire 'goes away' or ceases exerting causal influence on your behaviour. Whyte tries to provide an explanation for the content of these desires in terms of states of affairs which would in this way *satisfy* them. The notion of satisfaction and also a satisfaction-condition here can thus be explained in a way that has nothing to do with the content of the agent's beliefs. To explain further, the idea is that a neural state in my brain is my desire to make myself comfortable by sitting on my chair in virtue of being satisfied by this state of affairs. That is, my comfortably sitting on my chair would make this desire "go away" or to cease exerting causal influence on my behavior in the right kind of way. Whyte's proposal is that for at least some desires, D, what makes it the case that D is a desire for (or that D has a fulfillment condition of) O is that O satisfies D, in the sense that D "goes away" (or ceases to exert influence on action) when O obtains.

Whyte thus suggests what we might call (W) (for Whyte's proposal) as an alternative for (F) in the case of these relevant desires:

(W): What makes it the case that some state of a system "D is the [basic] desire that [O is that] D *would* be satisfied iff [O]."

It is worth noting that Whyte does not claim that the content of all desires can be explained in this way. The idea is that W can be used to explain the content of a set of "basic" desires, which can then be used in conjunction with R and F to explain the content of the rest of the agent's beliefs and desires. Although Whyte does not go into further depth here, in what follows I will try to explore in more detail how by using W, R, and F we can build up to give an account of the rest of an agent's mental states.

As Whyte also notes, this attempt to understanding the fulfillment of certain desires in terms of their satisfaction or "going away" faces the objection famously made by Ludwig Wittgenstein that some desires might go away when they are not actually satisfied. For instance, when I am hungry, and I want to eat food, my desire might go away as a result of watching disgusting videos, but this has not caused the desire to be satisfied or fulfilled. Whyte calls these [apparent] cases of *unfulfilling satisfactions*. He also mentions another sort of case as another problem for the attempt to understand fulfillment as satisfaction, which are [alleged] cases of *unsatisfying fulfillment*,. These are cases in which a desire is fulfilled but it doesn't "go away" and still exerts its causal influence on the agent's behaviors. For example, we might want to have a cup of tea, but after drinking tea we then come to want some more.

Whyte (1991) believes that these alleged cases of unsatisfying fulfillment are not a serious problem for success semantics. Whyte argues that what happens in these cases is not in fact an unsatisfying fulfillment of desires, but the emergence (or persistence) of some new (or other) desires with new (distinct) satisfaction conditions. Whyte does not believe that, when someone wants a cup of tea and still wants more after drinking it, her initial desire for a cup of tea was unsatisfyingly fulfilled. In actuality her initial desire for drinking a cup of tea has been fulfilled and a new desire of wanting second cup has emerged (or persisted – e.g. if one wants of each of two cups of tea to drink them).

Nevertheless, apparently the case of unfulfilling satisfaction is trickier. In order to distinguish the sort of satisfaction in terms of which we should understand the fulfillment of basic desires from what is going on in the apparent cases of unfulfilling satisfaction, Whyte

argues that we need a more sophisticated conception of the satisfaction of desires than their merely 'going away'. Inspired (like Dretske) by the idea of reinforcement learning, Whyte suggests that, while unfulfilling "satisfactions" cause the relevant desires merely to go away, proper (i.e. basic desire fulfillment-explaining) satisfactions *reinforce* the actions that have led the related desires to go away. That is, an agent will be more likely to repeat those actions the next time in similar situations of having this desire. To illustrate, suppose you are hungry late at night and you find a twenty four hour restaurant. By ordering food from that restaurant, it will not only make your desire to eat food 'go away', but it will also make you more likely to order from the restaurant the next time you get hungry late at night. In contrast, if you become hungry at midnight and happen to watch a disgusting video that causes your desire for food to 'go away', this won't make you more likely to watch such videos the next time that you become hungry late at night. [24] Hence, ordering the food from the restaurant would in the relevant sense satisfy the desire to eat food while watching disgusting videos would not.

Whyte, of course, notes that the idea of reinforcement in this solution must be understood in *counterfactual* terms. Whyte's idea of using counterfactual reinforcement to explain the content of basic desires can thus be seen as a modification to Dretske's (1988, ch 4) attempt to use reinforcement learning to explain the content of certain mental states. Due to his historical-teleological view, Dretske essentially appeals to the actual states of affairs that cause a desire go away in explaining its fulfillment conditions. For instance, a state of Mickey counts as a desire to press a button in a human's presence when he is hungry (as opposed e.g. to a desire to press the button in the presence of mere human-like sounds when he is hungry) because pressing it in the human's presence when he was hungry in the past is what has reinforced his tendency to press it today. Dretske requires a particular actual history behind the desire to explain what it represents (although as we saw, in trying to explain the determinacy of what one represents, Dretske does seem to want to appeal to counterfactual reinforcement as well). However, by relying upon a purely counterfactual understanding of reinforcement, Whyte can explain how a state in me is e.g. a desire for food because food **would** make it go away in a way that reinforces a **tendency** of

[24] Watching the disgusting videos at midnight, which causes the desire to eat food to 'go away', might reinforce such behaviour in the service of some *other* desires, e.g. the desire to avoid desiring food late at night, which is a second-order desire. However, this kind of reinforcement will not take place unless we have a second order desire (or something like that) such as the desire to stop one's feeling of hunger.

mine to perform the acts that caused it to go away, so that it **would**[25] cause me to perform similar actions if I **were** in a similar situation. This reinforcement never has to happen, nor if it happens, need it ever cause me to act on the reinforced tendency.[26] Whyte thus clarifies W as

> (W'): What makes it the case that some state of a system D is the basic desire that O is that O would (i) cause D to "go away" or cease exerting causal influence on the system's behaviour, (ii) in such a way that D's going away reinforces the system's **disposition** to act in the way that caused D to go away in similar circumstances.

However, as Whyte concedes, there are some cases which show that his solution will not work as it stands, even as an account of the content of desires that he proposes as examples of basic desires. For instance, there are some cases in which outcome O might cause an agent's desire to (i) go away in a way that (ii) it reinforces the behaviour that led to O, but in actuality O is not the desire's fulfillment condition (i.e. the desire is not a desire for O). As an example, suppose you want to increase your wealth and you perform a certain action (e.g. send money to what you believe is an investment service), which brings about outcome O (e.g. the numbers on your computer screen indicate that the value of your investment has increased), which you believe is increasing your money. In reality, your belief is false and you are actually losing money (e.g. because the alleged investment service is fraudulent, and they are simply pocketing the money you send to them). However, outcome O (i.e. your seeing increased numbers on your computer screen) will (i) at least temporarily cause your desire to 'go away' or stop influencing your behavior, and also (ii) reinforce some tendency in you to perform those actions again (i.e.

---

[25] The word 'would' here is supposed to refer to possibly counterfactual situations.

[26] The idea of counterfactual reinforcement preserves Whyte's ability to escape Swampman problem faced by Dretske's account in terms of actual learning history. As we discussed, swampman comes into existence without having any history of reinforcement (i.e. actual reinforcement). But because it seems that swampman can have beliefs and desires (especially basic desires), the scenario motivates the denial of the necessity of any actual learning history/selection for a system to have such mental states, causing serious problems for any historical-teleological theory. Because it appeals to counterfactual states of affairs in explaining the content of mental states, Whyte's account provides a way to consider swampman's physical states as real mental states. For instance, although swampman does not have any former experience of making himself comfortable by sitting on a chair, Whyte can argue that because sitting on the chair would reinforce his tendency to do so in the future if he did so, we should take his current physical state as a desire which has satisfaction conditions (of making himself comfortable by sitting) and consequently has a certain content. The fact that counterfactual states of affairs can be considered in explaining mental content enables Whyte to confirm our sense that the lack of any past (or even future) experience in the swampman scenario is an impediment to his having mental states with the sort of content that can misrepresent the world. For instance, Whyte's counterfactual reinforcement account of the contents of desires can explain how a swampman can just come into existence for some seconds and then die, but still at those moments he existed have mental states with certain contents.

send more money to this apparent investment service that seems to be securing for you a very nice rate of return) — at least, until you are aware of your false belief.

Someone might object that the desire for additional money is not appropriately a "basic" desire, and hence it cannot be a counterexample to W'. However, Whyte (1991) takes the example of desires for things like cherries (understood as the desire to eat cherries)to be uncontroversially "basic", and argues this problem can arise even in these cases. Suppose there is something which tastes just like real cherries (let's call it "imitation cherries"), so that one cannot distinguish it from real cherries. Eating such imitation cherries will cause one's desire for cherries to (i) go away (ii) in such a way that will reinforce a disposition to act (i.e. eat imitation cherries) in the way that caused the desire to go away in similar circumstances. Because both cherries and imitation cherries thus reinforcingly satisfy one's desire, W' seems to say that it is a desire for cherries *or* imitation cherries. But it seems that one could have a desire for cherries per se (and not imitation cherries) that could be in this way reinforcingly satisfied by imitation cherries.

As another problem for W', Whyte observes that it seems possible to fulfill a "basic desire" like the desire for cherries without necessarily (i) causing the desire to go away (ii) in such a way that it reinforces the behaviour that led it to go away. Suppose one desires cherries, but an external force brutally puts cherries in one's mouth, forces one's jaws to chew them, and also forces one's throat muscles to swallow them all while one is paralyzed and incapable of doing anything. It seems one's desire for cherries is fulfilled. While in such cases one's desire for cherries will presumably (i) go away, it was not (ii) in such a way that it reinforces one's doing the same behaviour in similar situations. This is simply because there was no actual behaviour on one's part that led the desire to go away that can be reinforced.

In response to these problems, Whyte (1991) proposes a further modification of W to restrict what it says to "normal conditions." The basic idea is that a desire's content is what would reinforcingly satisfy it under *normal* conditions. Thus, desires of earning money and eating cherries are not fulfilled in the above scenarios because, although the numbers on the screen and "imitation cherries" reinforcingly satisfy them, or similarly the desire for cherries is technically fulfilled by the brutal external force, the conditions in these scenarios are not relevantly "normal." Whyte notes, however, that it would be circular to understand "normal"

conditions as any conditions where a desire is reinforcingly satisfied if and only if it is fulfilled, and that it would be a mistake to identify normal conditions with whatever conditions are statistically typical, as it is conceptually possible for error to be statistically typical. He thus proposes to understand normal conditions as those conditions that would remain reinforcingly satisfying no matter how much the agent's perceptual capacities were improved. Whyte thus arrives at:

> (W''): What makes it the case that some state of a system D is the basic desire that O is that O would (i) cause D to "go away" or cease exerting causal influence on the system's behaviour, (ii) in such a way that D's going away reinforces the system's disposition to act in the way that caused D to go away in similar circumstances, and (iii) (i) and (ii) would remain true no matter how much the system's perceptual capacities were improved.

However, Hattiangadi (2007) claims that there is a serious circularity in Whyte's characterization of normal conditions. As he puts it, "In order to decide what counts as an 'improvement' of my perceptual abilities, assumptions have to be made about what I want, which is ultimately circular." However, the circularity might not be as obvious as Hattiangadi seems to suppose, as some someone might think that perceptional improvements could be characterized merely as increased abilities to discriminate among different states of the world, which might not seem to presuppose the content of the agent's desires. But, regarding the above mentioned example, what would be true if I were to gain the ability to discriminate among various microscopic events is not relevant to the content of my desire for cherries. It seems that for W'' to explain the content of my desire for cherries, it needs to consider what would be true if I were able to discriminate cherries from imitation cherries. Presumably Hattiangadi's point is that to pick out that perceptual improvement as relevant seems to presuppose the content of my desire. In addition to this, an arbitrary improvement of my perceptual capacities might alter the content of my desires, causing (i) and (ii) to cease to hold of someone outcome O not because O was not what I desired, but because the changes mentioned in (iii) changed what I desired. For example, I might now have a desire to eat cherries, but because of having improved perceptual abilities I might come to only desire to eat more particular kinds of cherries (for an analogous point in the case of full information analyses of ethical facts, see e.g. Gibbard 1990, Ch 1).

Finally, Whyte does not seem to notice that his understanding of "normality" in W'' in terms of perceptual improvements does not solve the problem of the possibility of fulfillment without reinforcement. As we saw if an external force places a cherry in my mouth and causes me to chew and swallow, there would be no behaviour of mine to be reinforced, so (i) and (ii) would not hold of my eating cherries even though I desired to eat cherries - and this would remain true even if my perceptual abilities were arbitrarily improved.

3.1.4. The Problem of Complex Attitudes

Some authors believe that success semantics cannot explain the content of certain more complex beliefs, and—in the case of Whyte's view (in which the content of desires is also explained)—more complex desires. Let us call this the *problem of complex attitudes*. For example, Bermudez (2003) argues that although success semantics can give the best account of the content of the sort of beliefs and desires that are shared by users of public languages and non-linguistic thinkers (such as preverbal human children and at least many non-human animals), this account is unable to give a persuasive account of the content of many of the beliefs and desires of language-users. First, Bermudez (2003: 66) claims that "There are no prospects for giving an account of the complex belief systems of language-using creatures in terms of success semantics, for the simple reason that so many of our beliefs have little direct contact with actions or desires." Second, Bermudez (2003: 68) observes that Whyte's (1991) account of the fulfillment conditions of certain basic desires (i.e. W'') cannot be extended to more complex desires.

Third, Bermudez (2003: 68, and Ch. 8 & 9) argues in particular that success semantics cannot adequately explain the content of certain of what Frankfurt (1971) calls "second order" mental states, which are mental states the content of which concerns other of the agent's mental states (in contrast to 'first-order' mental states, the contents of which are non-mental). Bermudez argues that public language is necessary for an important subset of second-order mental states. By a public language Bermudez (2003: 155-6) means a symbol system for intra-personal communication that exhibits compositionality, or the feature that the meaning of complex symbols is determined by the meaning of their parts, in a way that involves sub-propositional symbols that signify predicates and referring expressions being combined into complex symbols that signify propositions. Bermudez (2003: 158-64) argues that language is necessary for second-

order mental states involved in what Clark (1996) calls "second-order cognitive dynamics," or "capacities involving self-evaluation, self-criticism and finely honed remedial responses" such as "recognizing a flaw in our own plan or argument, and dedicating further cognitive efforts to fixing it; reflecting on the unreliability of our own initial judgements in certain types of situation and proceeding with special caution as a result" and so on. Moreover, Bermudez (2003: 170-188) argues that public language abilities (which allow for the right kind of logical operations) are necessary to have beliefs and desires about several important kinds of things, including: other agents' desires for states of affairs (as opposed to simply their desires for objects or properties), other agents' perceptions and beliefs that one can think of as mistaken (as opposed to simply their having representations of the world that one takes to be accurate), necessity and possibility (in the general sense in which we conceive of it), the abstract notion of past and future (as opposed, e.g., simply to sequences of events), universal and existential quantification (as opposed, e.g., to simply have tendencies to attribute properties to each of a set of things), and the ability to think about things in ways that are maximally abstracted from their contextual features.

A detailed discussion and evaluation of Bermudez's arguments about the dependence of various kinds of thinking on public language is beyond the scope of this thesis. They are certainly controversial, not least because many of them hinge upon Bermudez's unsupported and dubious assertions that the relevant second-order cognition must always be conscious, that public language is required for conscious second-order cognition, and that logical operations require second-order cognition (see e.g. Bargh and Ferguson 2000; Fodor 2003; Butterfill 2004). The relevant question for my purposes is why Bermudez suggests that a potential link between public language abilities and various kinds of thought poses any kind of problem for success semantics as a theory of the broad content of our mental states.

I believe that Bermudez's arguments that success semantics cannot explain the content of complex attitudes suffer from three crucial oversights. First, in suggesting that success semantics faces a problem because "many of our beliefs have little direct contact with actions or desires," Bermudez seems to have overlooked the ability of Whyte's (1990) success semantics to explain the content of such beliefs that motivate actions only in conjunction with other beliefs. It may be, for instance, that beliefs about general relativity motivate action only in the presence of many other beliefs, such as those connecting fundamental physics to the engineering of satellites. But

as we saw Whyte's success semantics does not attempt to identify the content of beliefs about general relativity in terms of their combining with desires all on their own. It begins by identifying the content of the whole package of beliefs (e.g. about general relativity, satellite engineering, and so on) that combine with desires (e.g. to have accurate readings of a GPS receiver's position), and identifies the content of beliefs about general relativity as the constant contribution that they make to the success conditions of all possible packages of beliefs (and desires) in which they could participate to cause action. Second, in suggesting that Whyte's success semantics cannot explain the content of desires (and beliefs which interact with them) beyond the basic ones identified by W'', Bermudez is overlooking the fact that W'' and R were never supposed to work on their own to explain the content of all of an agent's beliefs and desires. As Whyte suggests, an explanation of the content of all of an agent's mental states is to be built up out of W'' and R together with F (or SS_D), or the idea that what it is for (many) states to be desires for O is for them to tend to combine with one's beliefs to cause behaviour that would bring about O if those beliefs were true. Finally, third, Bermudez seems to assume that if a thought depends upon public-language abilities, its content cannot be explained by success semantics. The relevant assumption seems to be that we already have a clear, correct account of the content of mental states that reflect public language, which is nothing like success semantics (or any of the other theories that I discussed in chapter 2; cf. Fodor 2004). But the content of thoughts that reflect public language are just as much in need of explanation as any other – as are the meanings of public language expressions themselves, which, as Grice (1957; 1969) convincingly argued, must be explained in terms of the content of the mental states that they are conventionally used to express. There is thus no reason why success semantics cannot offer itself as an account of the broad content of mental states that reflect public language, or indeed (as Mellor 2012 argues) the meaning of public language expressions themselves.

In fairness to Bermudez, Whyte did not sufficiently develop his proposal about how to construct an explanation of all of an agent's beliefs and desires out of W'', R, and F. We might thus read Bermudez charitably as challenging the success semanticist to develop this proposal. I will seek to meet this challenge in section 3.2.3. below.

Nanay (2013) also argues that although success semantics can adequately explain the content of certain relatively simple mental representations, it cannot explain the content of all

mental representations. In particular, Nanay argues that, although success semantics can explain the content of domain specific intermediaries between perception and action which he calls "pragmatic representations", it cannot adequately give an account of the content of the highly interdependent action-tendencies of domain-general beliefs (and desires).

Nanay appears to suggest this modification of success semantics to provide a solution for the problems of non-necessity and insufficiency by limiting the scope of account. After reviewing the problem of insufficiency and the apparent problems with Whyte's introduction of "no impediments" beliefs, Nanay relates how the original idea of success semantics as proposed by Ramsey, unlike what Whyte tries to do, was not intended to explain the content of all sorts of intentional states. Nanay proposes to follow Ramsey by in particular narrowing the scope of account to the contents of the immediate perceptual antecedents of action which he terms "pragmatic representations."

Nanay argues that these domain-specific perceptual mental states are the immediate antecedents of action, and that they should not be considered beliefs. He argues that the content of these very basic representations, unlike beliefs, do not depend on any other intentional states. To illustrate, suppose, you that you know you are wearing a pair of goggles which distorts images of objects, so that you see the objects with a certain shift from their actual position. Now, suppose you want to throw a ball into a basket. Because of the distortion caused by the goggles you won't be able to throw correctly, and hence your desire of putting the ball in the basket won't be fulfilled. But your beliefs about the location of the basket and general layout of your surroundings may not have changed (since again, you know that the goggles are distorting your vision, and that no change occurs your environment when you put them on). Thus, Nanay concludes, your actions of throwing must be guided by a kind of perceptual representation of the location of the basket and your relation to it, which misrepresents this information, in contrast to your beliefs about the basket and your relation to it, which continue to accurately represent the basket and your relation to it. While your beliefs about the basket and your relation to it can remain constant across different circumstances due to their interconnection to your other background beliefs, our actions can are guided by pragmatic representations which, because of their insensitivity to our background beliefs, can vary with our immediate perceptual information.

Although the content of an agent's 'pragmatic representations' certainly has some influence on her other representations, Nanay argues that, regarding the content of these representations, we can provide a straightforward way to explain the success (or goal-satisfaction) of an agent's actions based upon them in a way that is relatively independent of the influence of her other representations. To illustrate, there is a laptop on the desk in my room. The pragmatic representations related to this laptop represent the features which are immediately derived from perception: e.g. that it is located directly in front of me. Nanay's idea is that these basic representations combine with my desires to cause behaviour that would fulfill those desires if they were correct, and that this is relatively independent of the content of my other representations. For instance, if I want to lift the laptop up, my pragmatic representation of its location directly in front of me will tend to combine with this desire to cause me to reach for it in such a way that would lift it up if it were located directly in front of me. It will have a tendency to exert this influence even if I believe many things to the contrary - even if, e.g. I believe that I am wearing vision-distorting goggles and the laptop is actually located a bit to my left.

Nanay seems to think that this relative independence of the influence of pragmatic representations from the content of one's other representations helps to solve the problems of insufficiency and non-necessity in giving an account of their content. However, as Nanay also discusses, the obtaining of the content of an agent's pragmatic representations is not sufficient for the success of an action, because they might not include representations of something which might lead the agent to fail. For instance, my pragmatic representations might correctly portray my laptop as directly in front of me and as having such a weight that it can be lifted by a certain way of reaching for it, but this way of reaching might fail to lift it because is glued to my desk and I have no pragmatic representation of this. Similarly, the accuracy of an agent's pragmatic representations is also not necessary to the success of her actions because sometimes mere good luck might lead the agent to act successfully on the basis of inaccurate pragmatic representations. For instance, my pragmatic representations of my laptop might portray it as too heavy to be lifted by one hand, and thus I will lift it by two hands, while in fact it was light enough to be lifted by only one but, because it was (unbeknownst to me and my pragmatic representations) glued to my desk, it required two hands to lift it after all.

To solve these problems of insufficiency and non-necessity in the case of applying success semantics to explain the content of pragmatic representations, Nanay argues that we should appeal to the idea that the obtaining of the content of an agent's pragmatic representations *increases the probability* of the success of actions based upon them. Thus, the idea is that what it is for a state to be a pragmatic representation that P is for it to combine with an agent's desires to cause behaviour that would, instead of guaranteeing the fulfillment of her desires, *increase the probability* of fulfilling her desires. For example, as we saw, the obtaining of the content of my pragmatic representations of the position and weight of my laptop will not guarantee the success of my actions relating to it – such as lifting or moving it, since it might, unbeknownst to me and unrepresented by my pragmatic representations, be glued to my desk. But the obtaining of the content of these representations will raise the probability of the success of these sorts of actions. If my representations of its position and weight are accurate, my reaching to lift my laptop in a corresponding way is more likely to fulfill my desires than the other things I might do. By contrast, acting on the basis of inaccurate pragmatic representations is unlikely to be successful, as one requires the right kind of lucky fluke for them to lead to success.

In order to clarify the relevant idea of raising the probability of success, Nanay explains that pragmatic representations increase the probability of the success of an action based on them in what he calls a *strong* sense. In fact, the truth of any sort of belief would raise the probability of the success of an action based upon them in the sense that, *so long as one has no other false beliefs*, the truth of the belief would raise the probability of the success of one's actions. Nanay terms this the raising of the probability of success in a weak sense – i.e. in the sense that, if the rest of one's representations were accurate, the obtaining of the content of the representation would increase the probability of success. Nanay claims that increasing the probability of success "only in a weak sense of the term…would be of not much use for us." Presumably Nanay's concern is that this explanation of the content of a mental state presupposes the content of the agent's other mental states, and would thus render a success semantical account viciously circular. Nanay distinguishes this weak sense of raising the probability of the success of an action from raising it in what he calls a strong sense. For the obtaining of the content of a representation to increase the probability of an action's success in this strong sense is for the obtaining of its content to increases the probability of the action's success regardless of whatever else is going on in the agent's mind.

Finally, Nanay considers that the accuracy of a pragmatic representation may increase the probability of the success of actions that do not depend upon all of its content. For example, the pragmatic representation that helps me to grasp and lift my computer also helps me to extend my hand in the general direction of my computer, but my merely extending my hand in the general direction of my computer does not need to utilize the entire content of the pragmatic representation that helps me to lift it (e.g. if I were wearing vision-distorting goggles that make objects look farther away than they are, my accurate pragmatic representations of the general direction of the computer could lead me to successfully extend my hand in its general direction, while my inaccurate pragmatic representations of its exact distance from me might lead me to fail to successfully grasp and lift it). Nanay's solution to this difficulty is to focus on the success conditions for actions which are not proper parts of other actions which a pragmatic representation might help to succeed. He thus identifies the content of a pragmatic representation with the conditions which (in the strong sense) raise the probability of the success of the actions it causes, which are not proper parts of any other actions the probability of the success of which it raises (in the strong sense). As Nanay puts this final proposal, "the correctness conditions of a pragmatic representation, R, is C if and only if C raises the probability (in the strong sense) of the success of the action R is the immediate mental antecedent of and this action is not the proper part of any other action the success of which R raises the probability of."

While I think that Nanay's probability raising solution to the problems of insufficiency and non-necessity is promising, I will explain in the next section how I think that it can be extended to solving these problems for a version of success semantics that, like Whyte's, seeks to explain the broad content of all mental states, including domain-general and interdependent beliefs and desires. I will thus argue that Nanay is mistaken in thinking that employing the idea of probability raising to solve the problems of insufficiency and non-necessity requires the restriction of the account to explaining only the content of pragmatic representations.

3.2. Solving the Problems of Success Semantics

In this section I will attempt to show how the problems for success semantics which I have discussed in previous section can be solved by drawing upon the literature which I have

reviewed and evaluated in that section. I will begin by showing how Nanay's probability-raising solution can be extended to solve the problems of non-necessity and insufficiency for a version of success semantics that, like Whyte's, seeks to explain the broad content of all psychological mental states. Next, I will explain how Whyte's solution to the problem of circularity can be made to work if we consider the right kinds of "basic" desires. I will show how the account of the content of certain immediate desires in terms of tendencies towards reinforcing satisfaction can be made to work in a way that does not require a problematic appeal to "normal" conditions. I will also develop Whyte's suggestion about how this account of the content of certain desires can be used in conjunction with a success semantical explanation of the content of beliefs in terms of the fulfillment conditions of desires, and an account of the fulfillment conditions of other desires in terms of the content of beliefs, to give a recursive account of the broad content of all of an agent's beliefs and desires. Finally, I will show how this recursive account constitutes a solution to the problem of complex attitudes, explaining not only the content of relatively simple beliefs and desires that immediately motivate action, but complex ones that depend for their influence upon many other beliefs and desires.

3.2.1. Solving the Problems of Insufficiency and Non-Necessity

As I discussed in section 3.1.1., Dokic and Engel (2002; 2005) outline a potentially viable way of defending Whyte's (1990) general solution to the problems of insufficiency and non-necessity. Whyte's general solution begins by noting that, although the truth of any individual belief is insufficient to guarantee the success of any actions it motivates, we can begin by explaining the broad content of complete sets of beliefs that motivate actions as the conditions that are sufficient for the success of those actions. Next, we can explain what it is for a state to be a belief with a specific broad content as what is common to the content of all of the possible sets of beliefs that include it (i.e. what is common to the conditions that guarantee the success of all of the possible actions it could motivate, in conjunction with all possible sets of other beliefs and desires). Finally, to deal with the apparent possibility that an agent's actions could fail due to ignorance rather than false belief, we argue that whenever an agent acts, she must have beliefs that entail that she faces no impediments in acting. As we saw, these cannot always be the same belief, and there are problems with motivating the view that agents always have beliefs that

entail that they face no impediments. But Dokic and Engel argue that there are principled grounds always to attribute to an agent performing action A to fulfill her desires that O an instrumental belief to the effect that 'if I do A, then O'. These are supposed to have to do with the content of such beliefs appearing in perception, and considerations motivating a principle of pragmatic closure analogous to the principle of epistemic closure. If this solution to the problem of insufficiency succeeds, we can solve the problem of non-necessity by noting that, although the obtaining of the content of any set of beliefs motivating an action is not necessary for its success (since an agent could get lucky by acting on a false beliefs), only the obtaining of the content of the beliefs motivating an action is sufficient for its success.

Although we cannot eliminate the possibility of the success of this Whyte-Dokic-Engel solution to the problems of insufficiency and non-necessity, as I noted in section 3.1.1., there are problems with Dokic and Engel's grounds for attributing instrumental beliefs to agents. First, there are reasons to doubt that all such instrumental beliefs are part of the contents of perception. Second, Dokic and Engel appear to say very little to motivate the principle of pragmatic closure besides claiming that it is analogous to the principle of epistemic closure. But in my opinion the principle of pragmatic closure seems dubious, because the reasons to doubt the principle of epistemic closure presumably correspond to the reasons to doubt the principle of pragmatic closure. Perhaps there are other principled grounds for always attributing to agents beliefs that essentially entail that they face no impediments in acting. However, as I noted in section 3.1.1., it is unclear if such grounds exist, and being strictly sufficient for the success of one's actions is a very demanding standard. So I believe that it is worth exploring an alternative solution.

As an alternative to the Whyte-Dokic-Engel solution, I want to propose applying Nanay's idea of probability raising in the strong sense (instead of guaranteeing success) to Whyte's general solution to the problems of insufficiency and non-necessity. Although, Nanay seems to clearly think that we must restrict the scope of success semantics to giving an account of the content of pragmatic representations in order to solve the problems of insufficiency and non-necessity, this seems puzzling because, (i) Nanay's main idea is to replace the idea of guaranteeing success with that of raising the probability of success in a strong sense (i.e. independent of the accuracy of the agent's other mental states), and (ii) he has adopted Whyte's holistic proposal for explaining the content of individual beliefs in terms of their common

contribution to the content of *sets* of beliefs. In my opinion, there is a plausible way to combine Nanay's idea of raising the probability of success in a strong sense with Whyte's holistic account of the content of sets of beliefs to explain the content of individual beliefs in a way that solves the problems of insufficiency and non-necessity.

More specifically, my proposal of combining Whyte's and Nanay's ideas will result in the following success semantical account of the content of beliefs in terms of their tendency to combine with desires to cause actions that fulfil those desires:

> (PR1[27]) What it is for a set of states of a system, S, to be a set of beliefs that together have the content C, is for it to be the case that S tends to combine with the system's desires to cause behaviour that would increase in the strong sense the probability of the fulfillment of those desires if C were the case.
>
> (PR2) What it is for a state of a system, s, to be a belief that P is for it to be the case that P is what is common to the content of all possible sets of beliefs of which s is a member [i.e. if the possible sets of beliefs that include s are S1, S2, …, Sn, … and the possible desires with which they can combine to produce actions A1, A2,…An,… are Δ1, Δ2, …, Δn, ….; then P is what is common to (i) the conditions under which A1 (caused by S1 and Δ1) increases in the strong sense the probability of the fulfillment of Δ1, (ii) the conditions under which A2 (caused by S2 and Δ2) increases in the strong sense the probability of the fulfillment of Δ2, …, (n) the conditions under which An (caused by Sn and Δn) increases in the strong sense the probability of the fulfillment of Δn, …[28]].

For instance, much as I discussed in section 3.1, (PR1) entails that what makes it the case that a state of me is the conjunction of beliefs that 'the store close to my home is a bakery, the store is open 8AM to 6PM except than Sundays and so on, …' is its tendency to combine with my desires (e.g. my desire for bread) to cause behaviour (e.g. walking to the location of store, walking up to the door, entering the store, looking for breads on the shelf, etc.) that would (not

[27] 'PR' standing for 'Probabilified Ramsey's' Principle.

[28] Of course, it seems possible for a set of beliefs to be such that all and only its members can play a role in combining with more than one set of desires to cause more than one action. If so, then this full list of possible actions caused by possible sets of beliefs and desires will have to include such combinations – e.g. the conditions under which some action (call it A1,2) caused by S1 and Δ2 increases in the strong sense the probability of the fulfillment of Δ2, and so on.

guarantee but) increase in the strong sense the probability of the fulfillment of those desires (e.g. my obtaining bread) if that content were to obtain (i.e. if it were the case that the store close to my home is a bakery, and the store is open 8AM to 6PM except than Sundays, and it is Wednesday 9AM, and …). Similarly, (PR2) entails that what makes it the case that some state of me is the belief that the store close to my home is a bakery is for the store's being a bakery to be what is common to the content of all possible sets of beliefs of which the state is a member. That is, it is for the store's being a bakery to be the only thing common to: (i) what (does not guarantee but) increases in the strong sense the probability of the fulfillment of my desire for bread when combined with this state together with the belief that the store is open 8AM to 6PM except than Sundays, the belief that it is Wednesday 9AM, etc.; (ii) what (does not guarantee but) increases in the strong sense the probability of the fulfillment of my desire to to learn the recipe for bread when combined with this state, the belief that my friend works at the store, the belief that people who work at bakeries know better than many people about bread recipes, etc; …. and so on.

This account can also give a solution to the problem of non-necessity in the exact same way that Nanay's account has proposed. In fact, while it is true that the obtaining of the content of the (entire) set of beliefs motivating an action is not necessary for its success, it is necessary to raise its probability of success in the strong sense. To illustrate, although some actions motivated by sets of beliefs which contain some false beliefs might still succeed, in reality the probability of the success of actions motivated by all true beliefs is higher that the probability of actions motivated by sets of beliefs some of which are false. It seems, then that by a wise combination of Whyte's original proposal with Nanay's idea of raising the probability of actions' success in the strong sense, we can successfully solve success semantics' problems of insufficiency and non-necessity.

3.2.2. Solving the First Part of the Problem of Circularity

In subsection 3.1.3, I discussed how the basic success semantical account of the contents of beliefs explains the content of an agent's beliefs in terms of the content or fulfillment conditions of her desires. This is also the case for our newly developed success semantical account of

beliefs constituted by PR1 and PR2. However, success semantics as a naturalistic theory of the content of mental states must explain the content of an agent's desires, and do so in such a way that does not reference the content of her beliefs in a way that makes the account viciously circular. As we saw, Whyte's general solution to this problem of giving a non-circular account of the fulfillment conditions of an agent's desires is to first give an account of the fulfillment conditions of a subset of the agent's desires (her 'basic desires') that does not reference the content of her beliefs. Next, the solution seeks to use this account of the fulfillment conditions of an agent's *basic* desires, together with the basic success semantical explanation of the content of her beliefs in terms of the fulfillment conditions of desires, and an account of the fulfillment conditions of other desires in terms of the content of beliefs, to give an account of content of the rest of her beliefs and desires (although Whyte himself never developed this second part of the solution).

As we saw, Whyte's basic idea for explaining the fulfillment conditions of an agent's basic desires appealed to the idea of reinforcing satisfaction or "going away":

(W'): What makes it the case that some state of a system, D, is the basic desire that O is that O would (i) cause D to "go away" or cease exerting causal influence on the system's behaviour, (ii) in such a way that D's going away reinforces the system's disposition to act in the way that caused D to go away in similar circumstances.

But Whyte raised two problems for this. First, even what he took to be basic desires, such as the desire to eat cherries, can be reinforcingly satisfied without being fulfilled – e.g. if the agent eats a cherry-like substitute that she cannot distinguish from cherries. Second, Whyte worries that it is possible for even basic desires to be fulfilled without being reinforcingly satisfied – e.g. if an agent were caused to eat cherries by wholly external forces, there would be no behaviour of hers to be reinforced. Whyte's solution to these problems was to invoke the idea of what would reinforcingly satisfy basic desires under "normal" conditions, which he understands as conditions of arbitrary perceptual improvement, i.e.:

(W''): What makes it the case that some state of a system D is the basic desire that O is that O would (i) cause D to "go away" or cease exerting causal influence on the system's behaviour, (ii) in such a way that D's going away reinforces the system's disposition to act in the way that

caused D to go away in similar circumstances, and (iii) (i) and (ii) would remain true no matter how much the system's perceptual capacities were improved.

But as we saw, Hattiangadi (2007) objected that what counts as a perceptual "improvement" presupposes what the agent actually desires, which makes the appeal to normal conditions circular. Moreover, it seems that some improvements of perceptual capacities might change what the agent desires, and ruling such changes out seems to presuppose the content of the desires in question. Finally, it does not seem that W'' solves the problem of fulfilment without reinforcing satisfaction, since no matter how much the agent's perceptual capacities are improved it still seems possible for wholly external forces to fulfil her desires, in which case there is no behaviour of hers to be reinforced.

I believe, however, in providing a solution for this problem we can still defend the essence of Whyte's reinforcing satisfaction account of the fulfillment conditions of basic desires, W', without an appeal to normal conditions. We simply need to identify the right set of desires as "basic," or restrict the application of W' to the right subset of the agent's desires. In this regard, I suggest that, we should restrict the domain of desires in our account into a subset of desires the content of which can be explained in terms of reinforcing satisfaction; I will call this subset *immediate desires*. The most familiar immediate desires may be an agent's desires for her own *phenomenal states*. As an example, the desire for cherry-like taste is an immediate desire while desire for cherries per se is not. The content of these desires is for the agent to be in a certain phenomenal state, and they certainly are reinforcingly satisfied by the agent's being in that phenomenal state. Moreover, unlike other desires, there does not seem to be any room for a gap between what reinforcingly satisfies these desires and what they are actually desires *for*. Although it is possible for one not to be able to discriminate the taste of cherries from the taste of imitation cherries, because the taste of both reinforcingly satisfies a desire one has (i.e. according to W', because the taste (i) causes the desire to "go away" or cease exercising causal influence on one's behaviour, (ii) in such a way that its going away reinforces the one's disposition to act in the way that caused it to go away in similar circumstances cause), it seems that one really does have a desire for the phenomenal experience of cherry-like taste that is common to experiences of eating cherries and imitation cherries.

What reinforcingly satisfies these immediate, phenomenal desires is the obtaining of their phenomenal content per se, which constitutes their being genuinely fulfilled. This reinforcing satisfaction of phenomenal desires by the phenomenal states that are their content takes place quite independently of whether the agent believes that these desires have been fulfilled. Consequently, there is no conceivable way for agents to be deceived about their fulfillment in a way that leads them to be reinforcingly satisfied without their actually being fulfilled. To illustrate how the reinforcing satisfaction of immediate desires is independent of her beliefs, suppose that an agent lacks beliefs about what is fulfilling her desire for cherry-like taste. She might be an infant who lacks beliefs about what she desires, or even a regular adult who isn't that much into navel-gazing and might not form any reflective beliefs about what, precisely, she wants. Or we can even consider a regular adult who falsely believes that she only wants to eat cherries per se, while in fact, what she has at least some desire for mere cherry-taste.[29] We can observe that, even in these cases, when she wants a cherry-like taste, her desire can be reinforcingly satisfied simply by her drinking or eating something with a cherry-like taste – which constitutes the fulfillment of her desire for cherry-like taste.

Therefore, for all agents, we can find a set of immediate desires (which we can call D_1), for which their fulfillment conditions can be explained as the conditions that reinforcingly satisfy them. Since this reinforcing satisfaction does not depend upon the agent's beliefs about whether the object of her desires obtains, we do not have to worry about the possibility that these desires are reinforcingly satisfied but not fulfilled. Hence, if we replace 'basic desires' with 'immediate desires', we can stick to the basic idea of W'. We do not need Whyte's idea of normal conditions or his move to W'' to try to deal with the possibility of the relevant desires being reinforcingly satisfied but not fulfilled under any conditions. We can thus avoid the problems with Whyte's appeal to normalcy such as those Hattiangadi (2007) mentions.

We might still need to worry, however, about the possibility Whyte raises of whether the relevant desires can be fulfilled without being reinforcingly satisfied. Again, it seems that an agent's desire for cherry-like taste could be fulfilled by external conditions over which she has

[29] It is possible that someone has both an immediate desire for cherry-like taste *and* a more reflective desire particularly for "cherries" at the same time. In these cases, due to having the immediate desire for cherry-like taste, any form of experiencing this taste, whether caused by cherries or imitation cherries, will reinforce the actions that have brought about the experience. However, the fulfillment condition of the more reflective desire will be explained in my response to the problem of *complex attitudes* below.

no control (such as randomly occurring direct stimulation of the neural correlate of the experience of cherry-like taste). Perhaps, however, we can say that in such cases, the agent's desires still have an essential causal propensity to reinforce behaviour leading to the fulfillment of the desire. It is simply that, because of the circumstances, there is nothing to reinforce and that causal propensity is unable to do its characteristic work. After all, if there *were* anything that the agent *could* do to increase the likelihood that the external forces would operate to bring about her experience of cherry-like taste (such as pressing a button, or asking whomever is directly simulating her brain), her doing it would cause her desire for cherry-like taste to "go away" or cease exerting causal pressure on her behaviour in a way that *would* reinforce her doing it in similar circumstances.[30] Thus, while the fulfillment of an immediate desire may not always reinforce the disposition of the agent to act in a way that caused the desire to go away, it does always have a tendency to reinforce any such acts.

I thus believe that we can solve the problems faced by Whyte's basic approach to explaining the content of a subset of desires without reference to the content of the agent's beliefs, and without recourse to W'' and an appeal to "normal conditions," by revising W' to:

(W*): What makes it the case that some state of a system, D, is the immediate desire that O is that O would (i) cause D to "go away" or cease exercising causal influence on the system's behaviour, (ii) in such a way that D's going away tends to reinforce the system's disposition to act in any way that caused D to go away in similar circumstances.

Having thus explained the content or fulfillment conditions of agents' immediate desires, we can use the basic formula of success semantics (SS_B, or more precisely PR1 and PR2) to go further and formulate an account of the content of a basic set of the agents' beliefs (which, to avoid confusion with epistemic basicality, let's call 'level-1 beliefs'). To illustrate, some state of an agent is a level-1 belief that P because of how it interacts with one or multiple possible immediate desires (like the desire for cherry-taste) to produce behavior that would tend to reinforcingly satisfy them if P were true. We can call this set of level-1 beliefs B_1. An instance of

[30] For the fulfillment of an immediate desire for experience E to reinforce act A which caused it be satisfied or "go away," must the agent happen to hold an instrumental belief to the effect that, 'if I do A, then E will occur'? This does not seem to be the case. Suppose that an agent lacks this belief; suppose, for instance, that she holds the opposite belief – e.g. if asked to bet on whether E will occur if she does A, she is motivated by her beliefs and her desire for money to bet against this. Still, if her doing A does cause E, it seems that she will find herself with a reinforced tendency to do A that goes against the content of her beliefs.

a belief in B_1 might be my belief that the stuff in my fridge has a cherry-like taste. This belief is a level-1belief because it directly combines with the immediate desire for cherry-like taste to produce behaviors (such as opening the fridge and searching for the stuff) that would reinforcingly satisfy the desire, if the belief were true. As another example, suppose I believe that there is some cherry-like tasting juice in the mug on my desk. This belief is a level-1 one because it directly combines with an immediate desire for cherry-like taste to produce a behavior (e.g. drinking it) which would reinforcingly satisfy my immediate desire if it were true. Of course, multiple level-1beliefs can work together to produce behaviour that would satisfy the agent's immediate desires if they were true, and the behaviours caused by level-1beliefs and immediate desires are not guaranteed to reinforcingly satisfy the desires. So to adequately characterize the content of an agent's level-1beliefs we would want to employ PR1 and PR2, where the fulfilled desires in question are restricted to immediate desires:

($PR1_{B1}$) What it is for a set of states of a system, S, to be a set of *level-1* beliefs that together have the content C, is for it to be the case that S tends to combine with the system's *immediate* desires to cause behaviour that would increase in the strong sense the probability of the fulfillment of those desires (i.e. their being reinforcingly satisfied) if C.

($PR2_{B1}$) What it is for a state of a system, s, to be a *level-1* belief that P is for it to be the case that P is what is common to the content of all possible sets of level-1 beliefs of which s is a member [i.e. if the possible sets of level-1 beliefs that include s are S1, S2, …, Sn, … and the possible *immediate* desires with which they can combine to produce actions A1, A2,…An,… are Δ1, Δ2, …, Δn, ….; then P is what is common to (i) the conditions under which A1 (caused by S1 and Δ1) increases in the strong sense the probability of the fulfillment of Δ1 (i.e. Δ1's being reinforcingly satisfied), (ii) the conditions under which A2 (caused by S2 and Δ2) increases in the strong sense the probability of the fulfillment of Δ2 (i.e. Δ2's being reinforcingly satisfied), …, (n) the conditions under which An (caused by Sn and Δn) increases in the strong sense the probability of the fulfillment of Δn (i.e. Δn's being reinforcingly satisfied), …].

This might seem to be a nice, non-circular way to explain the content of agents' immediate desires and level-1 beliefs. But one might worry that, if all immediate desires are desires for phenomenal states, and phenomenal states have intentional content (e.g. the experience of cherry-like taste represents certain features of what one is eating, or of one's bodily conditions –

see e.g. Tye 1995), then W*, $PR1_{B1}$, and $PR2_{B1}$ do not ultimately succeed in naturalizing intentionality. One response to this concern might seek to give a naturalistic account of the content of phenomenal states which, when conjoined with W*, $PR1_{B1}$, and $PR2_{B1}$ does not reference the contents of an agent's beliefs and desires in a way that is viciously circular. This, however, would be beyond the scope of this thesis.

A more modest response is that immediate desires are not *essentially* desires for phenomenal states; there may well be non-phenomenal immediate desires, and if there are not it seems that there could be. What makes a state of an agent an immediate desire is (i) its being such that its tendency to be reinforcingly satisfied depends upon what the obtaining of what actually fulfills it, rather than what the agent thinks fulfills it, and (ii) its tendency to combine with the agent's level-1 beliefs in the way described by $PR1_{B1}$ and $PR2_{B1}$.[31] Because the agent's internal mechanisms of reinforcing satisfaction must be able to detect the content of such a state directly (i.e. independent of something in the external world that typically correlates with it, but might fail on occasion, to do so), it does seem that immediate desires must be desires for states of the agent herself. But there is no need for them to be phenomenal states that have their own intentionality, or intentional states of any kind.

In order to see how W*, $PR1_{B1}$, and $PR2_{B2}$ describe an intentional system in naturalistic terms, and appreciate why immediate desires should be common to intentional systems, it may be helpful to return to Dennett's ideas about how a system without genuine intentionality (of the kind that admits of genuine misrepresentation) can be enhanced into a system with genuinely intentional states. As I discussed in the second chapter, Dennett takes an example of a thermostat and suggests that we can attribute some belief-like and desire-like states to it as a heuristic, such as the thermostat believing that the room is colder or hotter than the set-point or the thermostat wanting the heater to turn on/off when the temperature is colder/hotter than the set-point. Nevertheless, the intentional stance of attributing all of these belief-like and desire-like states is actually redundant because the behavior of the thermostat can be equally explained/ predicted from design (or physical) stance. Such attributions merely describe a pattern of causal processes

[31] Note: this is not viciously circular, as we can use the Ramsey-Carnap-Lewis method of explaining theoretically identified entities to give a non-circular account of immediate desires and basic beliefs that replaces mentions of them in W*, $PR1_{B1}$, and $PR2_{B1}$ with a Ramsey sentence that says that there exist things with the properties attributed to them in W*, $PR1_{B1}$, and $PR2_{B1}$.

that is just as well captured by the design or physical stances. One way to see why this is so is that what we assign to the thermostat as belief-like or desire-like states which are supposed to be about the temperature of a room can, depending upon what it is hooked up to, equally be *about* many other issues such as the brightness of lamp-light or the level of a water-tank.

However, considerer enhancing the thermostat with some sophisticated outputs (such as actions that purchase fuel), which can be caused by various desire-like states (that incline it to order fuel when the tank is 25% full instead of when it is empty) in conjunction with various sophisticated belief-like states (such one that combines with the foregoing desire to cause the ordering of fuel – i.e. which plays the role of a belief that the tank is 25% full). As Dennett argues, after a while we cannot (most efficiently, in a way that captures real-patterns) explain the behaviors of thermostat from the design or physical stances as well as with the intentional stance, and we have to attribute intentional states to the thermostat.

I believe that the simplest enhanced thermostat that requires the intentional stance to explain its workings will do so because its outputs (i.e. actions), goal-like states, and belief-like states conform to W*, PR1, and PR2. These are what help us determine why the belief-like states we attribute to it *cannot*, unlike those of a simple thermostat, be interpreted as *about* anything other than the temperature of a room. This belief-like state must stand ready to combine with many of the system's desire-like states and other belief-like states to produce behaviours that would be more likely to fulfill them if the room were a certain temperature as opposed to if some other condition were to obtain (such as a lamp's having a certain brightness, the level of a water in a tank being a certain height, and so on.) A fairly sophisticated enhanced thermostat might have goals that can be interpreted as about the temperature of the room per se, or the amount of fuel on hand, and so on. But such desires require the ability to represent, or have beliefs about these external world conditions. The simplest desire-like states that could get the belief-like states with which they combine off the ground to require an interpretation as genuine beliefs about such things in the first place would not presuppose the presence of other beliefs. But they would still have to be states that we could interpret as fulfilled (e.g. by actions motivated by belief-like states that get things right) and unfulfilled (e.g. by actions motivated by belief-like states that get things wrong), in a way that fits the attribution of minimal rationality, or trying to do things that fulfill one's desires in light of one's beliefs. W* seems to explain what is necessary and

sufficient for a non-belief-presupposing desire-like state to capture the real patterns of behaviour of a genuine desire. It captures how such a desire-like state tends to continue to exert influence on our conduct until it is fulfilled, and thus how agents who do not fulfill their desires by doing one thing will tend to try something else. Moreover, it captures how when such a desire-like state is fulfilled, the agent will tend to learn from this experience, and try the things that successfully fulfilled it again in similar circumstances in the future.

3.2.3. Solving the Second Part of the Problem of Circularity and the Problem of Complex Attitudes

I have thus shown how W*, $PR1_{B1}$, and $PR2_{B1}$ constitute a success semantical account of the content of immediate desires and level-1 beliefs that successfully solves the problems of insufficiency, non-necessity, and circularity. This seems sufficient to constitute a successful naturalistic account of the broad content of the beliefs and desires of a minimal intentional system, or the simplest sort of belief-desires system that represents the world in a way capable of genuine misrepresentation (or that can, in Dretske's words, not only fool us but itself be fooled). I believe, moreover, that this account can be combined with the following idea (and further uses of the general idea of PRP1 and PRP2) to give a non-circular account of the broad content of the rest of the beliefs and desires of more sophisticated intentional systems like us in a recursive way:

> (PSS_{NID}[32]) What it is for a state of a system to be a *non*-immediate desire that O is for the state to be disposed to combine with the system's beliefs to cause behaviours that increase in the strong sense the probability of O if those beliefs were true.

As a first step, we can use PSS_{NID} to extend our account from (1) the account of the content of immediate desires (D_1) that we get from W*, and (2) the account of the content of level-1 beliefs (B_1) that we get from $PR1_{B1}$ and $PR2_{B1}$, to (3) an account of a set of non-immediate desires which are in a position to interact with the agent's level-1 beliefs to produce behaviors (which we can call D_2). The idea is that what it is for a state of a system to be a D_2 desire that O is for the state to be disposed to combine with the system's level-1 beliefs to cause behaviours that

[32] For the 'probabilified success semantical account of non-immediate desires'

tend to increase in the strong sense the probability of O if those beliefs were true. For example, we discussed above my level-1 belief that there is some juice in the fridge which has a cherry-like taste, the content of which is explained in terms of W*, $PR1_{B1}$, and $PR2_{B1}$: there being juice in my fridge that tastes this way is what is common to the sets of level-1 beliefs of which it is a member which combine with my immediate desires (such as that for cherry-like taste) to produce behaviours (such as drinking the juice) that increase in the strong sense the probability of their fulfillment (such as my experiencing cherry-like taste: i.e. this desire being reinforcingly satisfied by this experience). Another level-1 belief might be one that a certain brand of juice (e.g. one in a bottle that looks a particular way) has a cherry-like taste. What PSS_{NID} can do is use level-1 beliefs like these ones to explain the content of non-immediate, D_2 desires such as my desire for there to be some juice with a cherry-like taste in my refrigerator. What it is for a state to be a non-immediate, D_2 desire for there to be a single bottle of juice with a cherry-like taste in my refrigerator is for it to combine with my level-1 beliefs (such as that there is no such juice in my refrigerator, and that a particular brand of juice has a cherry-like taste) to produce behaviour (such as my retrieving a bottle of that brand and placing it into my refrigerator) that will in the strong sense increase the probability of there being a single bottle of juice with a cherry-like taste in my refrigerator if those beliefs were true (i.e. if there is no such juice in my refrigerator, and that the particular brand does indeed have a cherry-like taste).

D_2 is the first set of non-immediate desires the contents of which are explained by the success semantical account. Agents who have these desires can indeed be mistaken about whether they are fulfilled (and again, without any problematic appeal to "normal conditions"). Notice that the account does *not* explain their content in terms of what would reinforcingly satisfy them – it explains their content instead in terms of their motivating, in conjunction with level-1 beliefs, actions that would tend to bring about those contents if the level-1 beliefs were true. What allows us to explain their content in terms of the content of beliefs in a non-circular way is that we have already explained the content of level-1 beliefs in terms of what they motivate in conjunction with immediate desires (and it is the contents of these desires alone which we explain in terms of reinforcing satisfaction, which avoids circularity in virtue of its not relying on the content of any beliefs). Thus, our account is consistent with the fact that it is possible for various desires that Whyte seemed to treat as "basic," (which in fact on our account are non-immediate) to be reinforcingly satisfied but not fulfilled (e.g. if I think that I've done

such a good job obtaining cherry-tasting juice, but unbeknownst to me the stuff tastes nothing like cherries) or fulfilled but not reinforcingly satisfied (if I put juice into my fridge that I don't think tastes like cherries but in fact does). According to the account, what reinforcingly satisfies a *non*-immediate desire is no part of the explanation of its content.

Now, once we have thus used W*, $PR1_{B1}$, $PR2_{B1}$, and PSS_{NID} to explain the content of (D_1) the agent's immediate desires, (B_1) the agent's level-1 beliefs, and (D_2) a first set of an agent's non-immediate desires, we can go further. We can now explain the content of a new set of beliefs, the content of which cannot be explained simply in terms of their interaction with D_1, but needs to be understood in part in terms of their interaction with D_2 as well. For example, the content of my belief that Safeway carries a particular brand of juice probably cannot be explained simply in terms of what it would motivate in conjunction with my immediate desires (e.g. for my own phenomenal states). But we might be able to explain the content of this belief in terms of combining with my D_2 desires, for such things as that brand of juice. What we can do is apply the basic success semantical account of the contents of beliefs in terms of the fulfillment of desires, PR1 and PR2, to both D_1 desires and D_2 desires, to explain the content of a new set of level-2 beliefs. This would be:

> ($PR1_{B2}$) What it is for a set of states of a system, S, to be a set of *level-2* beliefs that together have the content C, is for it to be the case that S tends to combine with the system's *immediate and D_2* desires to cause behaviour that would increase in the strong sense the probability of the fulfillment of those desires if C.
>
> ($PR2_{B2}$) What it is for a state of a system, s, to be a *level-2* belief that P is for it to be the case that P is what is common to the content of all possible sets of level-1 and level-2 beliefs of which s is a member [i.e. if the possible sets of level-1 and level-2 beliefs that include s are S1, S2, …, Sn, … and the possible *immediate and D_2* desires with which they can combine to produce actions A1, A2,…An,… are Δ1, Δ2, …, Δn, ….; then P is what is common to (i) the conditions under which A1 (caused by S1 and Δ1) increases in the strong sense the probability of the fulfillment of Δ1, (ii) the conditions under which A2 (caused by S2 and Δ2) increases in the strong sense the probability of the fulfillment of Δ2, …, (n) the conditions under which An (caused by Sn and Δn) increases in the strong sense the probability of the fulfillment of Δn, …].

For example, what makes a state of mine a belief that Safeway carries a brand of juice would be that Safeway's carrying this brand of juice is what is common to the sets of level-1 and level-2 beliefs of which it is a member which combine with my immediate and D_2 desires (such as my desire for that brand of juice) to produce behaviours (such as my going to Safeway, walking in, and searching on the shelves) that increase in the strong sense the probability of their fulfillment (such as my obtaining that brand of juice).

Once we have thus used W*, PR1, PR2, and PSS_{NID} to explain D_1 (immediate desires), B_1 (level-1 beliefs), D_2 (non-immediate desires), and B_2 (level-2 beliefs), we can keep going as long as we like. Having started with a base clause of W*, we can continue to recursively define different orders of beliefs and desires in terms of each other using PR1, PR2, and PSS_{NID}. This gives us an explanation of their contents and the relations of dependence among them in such a way that the content of no desire in D_{n+1} is explained in terms of the content of the beliefs in B_{n+1}. Instead, their contents will be explained in terms of the contents of beliefs in B_1 through B_n, which are explained in terms of the desires in D_1 through D_n, on down to B_1 beliefs which are explained in terms of D_1 immediate desires, the content of which does not depend on the content of any other intentional states. Thus, by starting with an account of immediate desires' content that does not depend on any other states' content, and building up level by level from there, the account avoids circularly.

More precisely, the recursive explanation is the following:

Let D_1 be the set of immediate desires.

> (W*): What makes it the case that some state of a system $d \in D_1$ is the immediate desire that O is that O would (i) cause d to "go away" or cease exerting causal influence on the system's behaviour, (ii) in such a way that d's going away tends to reinforce the system's disposition to act in any way that caused d to go away in similar circumstances.

Let B_n ($n \in \mathbf{N}$, $n > 0$; e.g. B_3) be the set of beliefs in a position to directly interact/combine with the desires in D_1 through D_n (e.g. D_1 through D_3).

> ($PR1_{Bn}$) What it is for a set of states of a system, S, to be a set of beliefs of levels-1 through n that together have the content C, is for it to be the case that S tends to combine

with the system's desires of orders 1 through n to cause behaviour that would increase in the strong sense the probability of the fulfillment of those desires if C.

($PR2_{Bn}$) What it is for a state of a system, $b \in S$, to be a level-n belief that P is for it to be the case that P is what is common to the content of all possible sets of beliefs of levels 1 through n of which b is a member [i.e. if the possible sets of beliefs of levels 1 through n that include b are S1, S2, …, Sn, … and the possible desires of orders 1 through n with which they can combine to produce actions A1, A2,…An,… are Δ1, Δ2, …, Δn, ….; then P is what is common to (i) the conditions under which A1 (caused by S1 and Δ1) increases in the strong sense the probability of the fulfillment of Δ1, (ii) the conditions under which A2 (caused by S2 and Δ2) increases in the strong sense the probability of the fulfillment of Δ2, …, (n) the conditions under which An (caused by Sn and Δn) increases in the strong sense the probability of the fulfillment of Δn, …].

Let D_{n+1} ($n \in \mathbf{N}$, $n > 0$; e.g. D_4) be the set of desires in a position to directly interact with beliefs in B_1 through B_n (e.g. B_1 through B_3).

(PSS_{Dn+1}) what it is for a state of a system $d \in D_{n+1}$ to be a *non*-immediate desire that O is for the state to be disposed to combine with the system's beliefs of levels 1 through n to cause behaivours that increase in the strong sense the probability of O if those beliefs were true.

Thus, after using D_1 to define an initial level of B1 beliefs, each level of beliefs opens up a new level of things that an agent can desire, and each of those desires in turn opens up a new level of goals that she can figure out how to fulfill, thus leading to the possibility for yet a new level of things she can believe. I believe that this recursive explanation of increasingly complex beliefs of an agent not only completes the solution of the problem of circularity (explaining beliefs and desires beyond those that are immediate or level-1 in non-circular way) but also solves the problem of complex attitudes. We have actually already seen in subsection 3.2.1. how PR1 and PR2's basic approach of using Nanay's idea of probability raising in conjunction with Whyte's method of explaining the content of individual beliefs in terms of sets of beliefs responds to Nanay's argument that success semantics must be restricted to the content of pragmatic representations. Moreover, I believe that this recursive account makes it clear why

Bermudez was mistaken to think that success semantics cannot explain the sorts of complex desires and beliefs of public-language using creatures like us, many of which go far beyond immediate desires and level-1 beliefs.

3.3. Conclusion

I have thus argued that success semantics offers us the best naturalistic account of the broad content of psychological mental states such as beliefs and desires, and that its problems can be successfully solved. We should naturalize the content of mental states because the alternatives are (like dualists) taking mental states as fundamental or (like eliminative materialists) rejecting the existence of mental states with their distinctive intentionality, neither of which is acceptable. Dualism is unacceptable because it either violates the causal closure of the physical, posits overdetermination by distinct causes, or rejects the causal efficacy of psychological states, and in any event the best versions seem to involve fundamental laws of nature that implausibly relate a tiny part of the universe to non-simple physical states. Eliminative materialism is unacceptable because predictions and explanations in terms of psychological mental states are "too good to be false"; such predictions and explanations of our behaviours are objectively more efficient due to their picking out real patterns among our internal states and what they lead us to do. Even if this were wrong, it would still be worth trying to naturalize intentionality so as to improve our evaluation of what should count as evidence that a physical system has mental states, which is important in the context of the evaluation of mental states in artificial intelligence systems, various non-human animals, and very young and intellectually disabled humans.

The main challenge to any project of naturalizing intentionality is to explain the possibility of genuine error or misrepresentation; an essential part of the sort of intentionality possessed by our mental states, which is philosophically difficult to explain. Phylogenetic teleological theories face a problem of indeterminacy in what our mental states were selected to track and face trouble explaining the content of acquired mental states, including the apparent irrelevance of the history of biological selection of our cognitive systems to the content of our mental states. Ontogenetic teleological theories seem to do better but still may face problems of indeterminacy, and they as well as phylogenetic teleological theories face the problem that the

absence of any learning history, as in the case of swampman, does not seem to entail that an entity completely lacks mental states. Fodor by his synchronic asymmetric dependence theory tries to escape all of these problems; however, unfortunately synchronic asymmetric dependence turns out to be at least as mysterious as intentionality itself. Fodor seems to lack a principled basis for his claims about what asymmetrically depends on what, relying on his intuitions about intentionality, and thus making his account viciously circular.

A more promising approach to solving the problem of naturalizing intentionality seems to come from Dennett's interpretationism, which seeks to explain the content of mental states in terms of what makes it the case that taking the intentional stance on a system captures a real pattern that makes explaining certain features of its behaivour and inner workings objectively more efficient. However, to avoid becoming an implausible form of subjectivism, Dennett's account must identify what it is for a system to have intentional states with its having states that play the right internal causal roles and / or its relating in the right holistic way to the world. Some of what Dennett says about the relevant features of an intentional system seem to motivate conceptual role semantics, which identifies a mental state's content with its internal causal roles, while other parts seem to motivate success semantics, which explains the content of our beliefs as the conditions that would lead to the success of the actions they motivate. As Block argues, although conceptual role semantics seems to do a very nice job explaining the narrow or internally determined content of our mental states, which corresponds to how we conceive of our representations of the world, it needs an independent theory of the broad content of our mental states, which corresponds to what makes them true and what they refer to (independent of how we conceive of those truths and referents). This is what success semantics importantly offers: a naturalistic theory of broad content that can work alongside conceptual role semantics' account of narrow content in a naturalistic "two factor" account of the content of our mental states. In addition, success semantics offers us a naturalistic theory of broad content that avoids the problems of teleological theories and Fodor's view, and retains the virtues of Dennett's general non-subjectivist interpretationist approach.

It is broadly believed that the original ideas of success semantics were proposed by Frank Ramsey; however, as a sophisticated theory of mental content it was developed and defended by J.T. Whyte (1990, 1991). The basic idea of success semantics is that what makes it the case that a

state is a belief that P is that it tends to combine with the agent's desires to cause behaviour that would fulfill those desires if P. Whyte observed, however, that to be a successful naturalistic account of intentionality success semantics requires a non-circular account of the content of desires, and that beliefs work together as a group to combine with our desires to cause behaviours. In response to the first issue, he proposed to give an account of the content of a set of basic desires in terms of reinforcing satisfaction, which does not reference the content of beliefs. In response to the second, he proposed to first give an account of the content of sets of beliefs as the conditions that guarantee the success of the actions that they jointly motivate, and to second understand the content of a given belief as what is common to the content of all possible sets of beliefs with which it could combine to motivate actions. Whyte's proposal constitutes a promising naturalistic account of the broad content of our beliefs and desires, which, in addition to the above-mentioned virtues, seems to further develop Dennett's original holistic insight about what it takes for a system to have genuine beliefs and desires by offering a plausible explanation of what exact kind of complex, interacting states a system must have to count as having intentional mental states.

Despite all of these advantages, Whyte's general account does seem to face problems. While Whyte argued that the truth of a set of beliefs is sufficient to guarantee the success of the actions it motivates, it seems, as Brandom (1994) suggested, that an agent can have all true beliefs but fail because of being ignorant of certain impediments. Nor will it due to weaken the connection between the obtaining of the content of an agent's beliefs and success to any causal relation, since as Godfrey-Smith (1994) observed, agents' true beliefs are not necessary for success, because agents can have false beliefs but still succeed if they get lucky. Moreover, Whyte's attempt to explain the content of a set of "basic" desires in terms of reinforcing satisfaction faced the problem that, for the desires he considered as "basic," it seems possible for them to be reinforcingly satisfied without being fulfilled, and fulfilled without being reinforcingly satisfied. To solve these problems Whyte altered the account, identifying basic desires' contents with what would reinforcingly satisfy them under "normal" conditions, understood as conditions of arbitrary perceptual improvement. But as Hattiangadi (2007) observed, this seems circular, since what counts as a perceptual "improvement" seems to presuppose the content of the desire in question. Finally, various authors worried that success semantics as developed by Whyte might work for explaining the content of certain relatively

simple mental states, but cannot explain the content of more complex attitudes. Bermudez (2003) claimed that the account could not explain the sophisticated contents of the mental states of language-using creatures like us, which are often far removed from considerations of what would fulfill relatively basic desires. Nanay (2013), in order to solve the problems of insufficiency and non-necessity, proposes two ideas: (i) restricting the scope of success semantics to explaining only the contents of the immediate perceptual antecedents of action that can operate independently of our beliefs, which he calls "pragmatic representations," and (ii) replacing the idea of representations' contents guaranteeing success with their increasing the probability of success in a strong sense – i.e. independent of what else is going on in the agent's mind.

Inspired by Nanay's domain specific account, I argued, however, that we can still solve these problems with success semantic in a domain general account which makes it our best naturalistic account of what makes it the case that a state is a psychological mental state with a given broad content. First, I argued that we can apply Nanay's idea of raising the probability of success in the strong sense to solve the problems of insufficiency and non-necessity, not only in accounting for the content of pragmatic representations, but also in accounting for the content of beliefs which guide our conduct in conjunction with many other beliefs. When we recall the actual structure of Whyte's actual view, we see that it begins by explaining the content of entire sets of beliefs that combine with our desires to produce behaviour, the truth of which does indeed increase the probability of success in the strong sense. We can then apply Whyte's original idea of understanding the content of individual beliefs as what is common to the content of all possible sets of beliefs into which they could enter to guide behaviour.

Second, I argued that we can begin to solve the problem of circularity by being more careful about the set of desires the content of which we seek to explain in terms of reinforcing satisfaction. I argued that, for immediate desires like an agent's desires for her own phenomenal states, there is no gap between what tends to reinforcingly satisfy them and what actually fulfills them. We can thus explain the content of these immediate desires in terms of what tends to reinforcingly satisfy them without needing to reference the content of any of the agent's beliefs. We can then give a non-circular explanation of an initial set of "level-1" beliefs in terms of the fulfillment conditions of the immediate desires with which they interact. What it is for a set of states to be a set of level-1 beliefs with content C is for them to combine with the agent's

immediate desires to cause behaviours that would increase in the strong sense the probability of the fulfillment (i.e. reinforcing satisfaction) of those desires if C. We can then once more follow Whyte's idea of identifying the content of a specific level-1 belief as what is common to the content of all of the possible sets of level-1 beliefs with which it could combine to guide behaviour.

Finally, I argued that we can complete the solution to the problem of circularity and the problem of complex attitudes by using the foregoing accounts together with one additional idea to give a recursive account of all of an agent's beliefs and desires, from immediate or level-1 desires and level-1 beliefs to desires and beliefs of arbitrary additional levels. We use as the base clause the foregoing account of the fulfillment conditions of immediate desires in terms of what would tend to reinforcingly satisfy them. We then use as our first two recursion clauses my probabilified version of Whyte's strategy of explaining the content of beliefs in terms of the contents of desires, generalized to explain the content of level-n beliefs in terms of that of level-1 through level-n desires (and also level-1 through level-n-1 beliefs). What it is for a set of states to be a set of level-1 through level-n beliefs with the content that C is for them to tend to combine with the agent's level-1 through level-n desires to cause behaviour that increases in the strong sense the probability of the fulfillment of these desires. What it is for a state to be a level-n belief that P is for P to be what is common to the content of all sets of level-1 through level-n beliefs with which it could combine to guide behaviour. Finally, we add to this the idea that what it is for a state to be a *non*-immediate desire that O of level n+1 is for the state to tend to combine with the agent's beliefs of level-1 through level-n to cause behaviour that increases in the strong sense the probability that O if those beliefs were true.

I believe that this defense of success semantics opens up important avenues for further work, and has important practical applications. First, as I have argued, success semantics is best understood as a naturalistic account of the broad content of mental states which forms one part of a "two factor" theory with a naturalistic version of conceptual role semantics as the account of narrow content. It would be interesting to further explore how these factors of the two factor account might interact and inform each other. Are there problems for naturalistic versions of conceptual role semantics (beyond those Block raises for Harman's single-factor version) that can be better solved when we bear in mind that it is being paired with success semantics as a

theory of broad content? Similarly, are there further issues for success semantics that can be resolved or informed when we bear in mind that it is being paired with conceptual role semantics as a theory of narrow content?

Second, as I indicated, the most obvious candidates for immediate desires are an agent's desires for her own phenomenal states, and phenomenal states seem to be intentional, or capable of representing the world in a sense that allows for misrepresentation. It would be important, in order to give a thoroughly naturalistic account of the intentionality of all mental states, for there to be a successful naturalistic account of the intentionality of phenomenal states. It would be interesting to explore in further work the extent to which a success-semantics-plus-conceptual-role-semantics account could be extended to explaining or used to help explain the content of phenomenal states. For instance, could these states' content be understood in terms of their causal tendencies along analogous lines – or at least in part in terms of their tendencies to cause or interact with beliefs and desires with the content-conferring causal tendencies identified by success semantics and conceptual role semantics?

Finally, although I have sketched how success semantics can give a recursive account of the content of belief and desires of increasingly complex levels, it would be good to apply this abstract account to actual examples of complex beliefs and desires. For example, it would be interesting to work through exactly how this recursive account can help to explain the broad content of the various states that Bermudez (2003) speculated could not be explained by a success semantical account, such as beliefs and desires about other agents' desires for states of affairs, other agents' perceptions and beliefs that one can think of as mistaken, necessity and possibility, the abstract notion of past and future, universal and existential quantification, and entities in ways that are maximally abstracted from their contextual features. It would also be interesting to explore how success semantics can give an account of beliefs about causation and essence, which might be important for explaining the differences between beliefs about surface features as opposed to underlying kinds, such as watery stuff in general as opposed to H_2O or XYZ in particular (see e.g. Chalmers 1996).

I believe that success semantics can help us correctly understand what sorts of systems can have mental states with their distinctive sort of intentionality, which is capable of genuine error. As I mentioned in subsection 1.3, this seems important for evaluating the success of

various projects in artificial intelligence, and attributing mental states to various non-human animals. Moreover, as I pointed out in the same section, if intentional mental states are an important prerequisite for morally important phenomenal states, or have moral importance in their own right, this view may be able to bring significant insights into what sorts of future AI systems, non-human animals, and extremely young or disabled humans (e.g. late term fetuses or adults between a persistent vegetative state and a minimally conscious state) are moral patients. As I have discussed, success semantics develops in more detail Dennett's general suggestion about the essence of genuine intentional mental states as consisting in the right sort of complex, holistic pattern of dispositions for internal states to interact with each other. It thus may take us closer to understanding what precise workings of AI systems, various non-human animals and very young or disabled humans constitute the presence of intentional mental states of various kinds.

Bibliography

Adams, Frederick and Kenneth Aizawa. 1994. Fodorian Semantics. In Steven Stich and Ted Warfield (eds.), *Mental Representation: A Reader*. Oxford: Blackwell.

Adams, Fred and Kenneth Aizawa. 2017. Causal Theories of Mental Content. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2017 Edition), URL = <https://plato.stanford.edu/archives/sum2017/entries/content-causal/>.

Armstrong, David M. 1981. The Causal Theory of the Mind. In David Armstrong, *The Nature of Mind and Other Essays*. Ithaca, NY: Cornell University Press.

Baars, B. 1988. *A Cognitive Theory of Consciousness*. Cambridge, UK: Cambridge University Press.

Basl, John. 2014. Machines as Moral Patients We Shouldn't Care About (Yet): The Interests and Welfare of Current Machines. *Philosophy and Technology* 27 (1): 79-96.

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

Bargh, John and Melissa Ferguson. 2000. Beyond Behaviorism: On the Automaticity of Higher Mental Processes. *Psychological Bulletin* 126 (6) 925-945.

Baumeister, R.F., E. Bratslavsky, M. Muraven, and D.M. Tice. 1998. Ego depletion: Is the active Self a Limited Resource? *Journal of Personality and Social Psychology* 74: 1252-1265.

Bermudez, Jose Luis. 2003. *Thinking Without Words*. Oxford: Oxford University Press.

Berkeley, George. 1710. *Treatise concerning the Principles of Human Knowledge*. Kenneth Winkler ed., Indianapolis IN: Hackett Publishing Company, 1982.

Berkovitz, Joseph. 2007. Action at a Distance in Quantum Mechanics. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition), URL = <https://plato.stanford.edu/archives/spr2016/entries/qm-action-distance/.

Blackburn, Simon. 2005. Success Semantics. In Hallvard Lillehammer & D. H. Mellor (eds.), *Ramsey's Legacy*. Oxford: Oxford University Press.

Block, Ned. 1978. Troubles with Functionalism. *Minnesota Studies in the Philosophy of Science* 9: 261-325.

Block, Ned. 1981. Psychologism and Behaviorism. *The Philosophical Review* 90(1) (Jan., 1981): 5-43.

Block, Ned. 1986. Advertisement for a Semantics for Psychology. *Midwest Studies in Philosophy* 10(1): 615-78.

Block, Ned. 2005. Two Neural Correlates of Consciousness. *Trends in Cognitive Sciences* 9(2) (February 2005): 46-52.

Boly, Melanie, Anil K. Seth, Melanie Wilke, Paul Ingmundson, Bernard Baars, Steven Laureys, David B. Edelman, and Naotsugu Tsuchiya. 2013. Consciousness in Humans and Non-Human Animals: Recent Advances and Future Directions. *Frontiers in Psychology* 4: 1-20.

Braddon-Mitchell, David and Frank Jackson. 2007. *Philosophy of Mind and Cognition: An Introduction*. Malden, MA: Wiley-Blackwell.

Braithwaite, Victoria. 2010. *Do Fish Feel Pain?* Oxford: Oxford University Press.

Brandom, Robert B. (1994). Unsuccessful Semantics. *Analysis* 54 (3):175-178.

Brentano, Franz. 1874. The Distinction Between Mental and Physical Phenomena. In Franz Brentano, *Psychology from an Empirical Standpoint*, transl. by A.C. Rancurello, D.B. Terrell, and L. McAlister, London: Routledge, 1973 (2nd ed., intr. by Peter Simons, 1995).

Brown, Curtis. 2016. Narrow Mental Content. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2016 Edition), URL = <http://plato.stanford.edu/archives/sum2016/entries/content-narrow/>.

Butterfill, Stephen. 2004. Thinking without Words by José Luis Bermúdez. *Mind* 113(452) (Oct., 2004): 733-736.

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

Chalmers, David. 2015. Panpsychism and Panprotopsychism. In Torin Andrew Alter (ed.) *Consciousness in the Physical World: Perspectives on Russellian Monism*, Oxford: Oxford University Press.

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

Clark, Andy. 1996. Dealing in Futures: Folk Psychology and the Role of Representations in Cognitive Science. In R. N. McCauley (ed.), *The Churchlands and Their Critics*. Oxford: Blackwell.

Davidson, Donald. 1987. Knowing One's Own Mind. *Proceedings and Addresses of the American Philosophical Association* 60(3): 441-458.

Degrazia, David. 1996. *Taking Animals Seriously: Mental Life and Moral Status*. Cambridge: Cambridge University Press.

Dennett, Daniel C. 1978. *Brainstorms*. Montgomery, VT: Bradford Books.

Dennett, Daniel C. 1981. True Believers: The Intentional Strategy and Why It Works. In A. F. Heath (ed.), *Scientific Explanation: Papers Based on Herbert Spencer Lectures Given in the University of Oxford*. Oxford: Clarendon Press, 150-167.

Dennett, Daniel C. 1991. Real Patterns. *Journal of Philosophy* 88(1): 27-51.

Descartes, René. 1637. *Discourse on Method*. In Donald A. Cress (ed., trans.) Discourse on Method and Meditations on First Philosophy, Indianapolis, IN: Hackett Publishing Company 1999.

Descartes, René. 1641. *Meditations on First Philosophy*. J. Cottingham, (ed., trans.), Cambridge: Cambridge University Press, 1985.

Descartes, René. 1649. *The Passions of the Soul*. In J. Cottingham, R. Stoothoff, and D. Murdoch (eds., trans.), The Philosophical Writings of Descartes, Volume 1, Cambridge: Cambridge University Press, 1996.

Dokic, Jérôme and Pascal Engel. 2002). *Frank Ramsey: Truth and Success*. New York: Routledge.

Dokic, Jérôme Pascal Engel. 2005. Ramsey's Principle Re-Situated. In Hallvard Lillehammer and D. H. Mellor (eds.), *Ramsey's Legacy*. Oxford: Oxford University Press.

Dretske, Fred. 1981. Knowledge and the Flow of Information. Cambridge MA: MIT Press.

Dretske, Fred. 1986. Misrepresentation. In R. Bogdan (ed.), *Belief: Form, Content, and Function*, Oxford: Oxford University Press, 17-36.

Dretske, Fred. 1988. *Explaining Behavior*. Cambridge MA: MIT Press.

Dretske, Fred. 1994. If You Can't Make One, You Don't Know How It Works. *Midwest Studies in Philosophy* 19(1): 468-482.

Elwood, Robert W. 2011. Pain and Suffering in Invertebrates. ILAR [Institute for Laboratory Animal Research] Journal 52(2): 175-184.

Feldman, Fred. 2004. *Pleasure and the Good Life: Concerning the Nature, Varieties and Plausibility of Hedonism*. New York: Oxford University Press.

Field, Hartry. 1977. Logic, Meaning, and Conceptual Role. *Journal of Philosophy* 74(July): 379-409.

Fodor, Jerry A. 1987. *Psychosemantics*. Cambridge MA: MIT Press.

Fodor, Jerry A. 1990. *A Theory of Content and Other Essays*. Cambridge MA: MIT Press.
Fodor, Jerry A. 2003. More Peanuts: Review of Jose Luis Bermudez, `Thinking without Words.' London Review of Books 25(19) (10/9/2003).

Frankfurt, Harry G. 1971. Freedom of the Will and the Concept of a Person. *The Journal of Philosophy* 68(1) (Jan. 14, 1971): 5-20.

Gibbard, Allan. 1990. *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Cambridge, MA: Harvard University Press.

Godfrey, Peter -Smith. 1994. A Continuum of Semantic Optimism. In Stephen P. Stich and Ted A. Warfield (eds.), *Mental Representation: A Reader*. Oxford: Blackwell.

Godfrey, Peter -Smith. 1996. *Complexity and the Function of Mind in Nature*. Cambridge: Cambridge University Press.

Grice, H.P. 1957. Meaning. *The Philosophical Review* 66(3) (Jul., 1957): 377-388.

Grice, H.P. 1969. Utterer's Meaning and Intention. *The Philosophical Review* 78(2) (Apr., 1969): 147-177.

Harman, Gilbert. 1982. Conceptual Role Semantics. *Notre Dame Journal of Formal Logic* 28(April): 242-56.

Hattiangadi, Anandi. 2007. *Oughts and Thoughts: Rule-Following and the Normativity of Content*. Oxford: Oxford University Press.

Jackson, Frank. 1982. Epiphenomenal Qualia. *The Philosophical Quarterly* 32(127): 127-136.

Knutsson, Simon. 2015. *The Moral Importance of Invertebrates Such as Insects*. Master's Thesis in Practical Philosophy, University of Gothenburg, Department of Philosophy, Linguistics and Theory of Science.

Kriegel, Uriah. 2013. Two Notions of Mental Representation. In U. Kriegel (ed.), *Current Controversies in Philosophy of Mind*. New York: Routledge, 161-179.

Leibniz, Gottfried Wilhelm. 1686. *Discourse on Metaphysics*. In Roger Ariew, and Daniel Garber (eds., trans.), G.W. Leibniz, *Philosophical Essays*, Indianapolis, IN: Hackett Publishing Company, 1989.

Lewis, David. 1970. An Argument for the Identity Theory. *Journal of Philosophy* 63(2): 17-25.

Luper, Steven. 2016. Epistemic Closure. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition), URL = <https://plato.stanford.edu/archives/spr2016/entries/closure-epistemic/>.

Mather, Jennifer A. 2008. Cephalopod Consciousness: Behavioural Evidence. Consciousness and Cognition 17: 37-48.

McMahan, Jeff. 2002. *The Ethics of Killing: Problems at the Margins of Life*. New York: Oxford University Press.

Merker, Bjorn. 2005. The Liabilities of Mobility: A Selection Pressure for the Transition to Consciousness in Animal Evolution. *Consciousness and Cognition* 14: 89-114.

Millikan, Ruth G. 1989. Biosemantics. *The Journal of Philosophy* 86(6) (Jun., 1989): 281-297

Nanay, Bence. 2013. Success Semantics: The Sequel. *Philosophical Studies* 165(1): 151-165.

Nisbett, Richard and Timothy Wilson. 1977. Telling More than We can Know. *Psychological Review* 84: 231-259.

Panksepp, Jaak. 2005. Affective Consciousness: Core Emotional Feelings in Animals and Humans. *Consciousness and Cognition* 14: 30-80.

Papini, Mauricio R. 2008. *Comparative Psychology: Evolution and Development of Behavior*, 2nd Edition. New York: Psychology Press.

Putnam, Hilary. 1967. The Nature of Mental States. In W.H. Capitan & D.D. Merrill (eds.), *Art, Mind, and Religion*. Pittsburgh: Pittsburgh University Press, 37-48.

Putnam, Hilary. 1975. The Meaning of 'Meaning'. *Minnesota Studies in the Philosophy of Science* 7: 131-193.

Prinz, Jesse J. 2002. *Furnishing the Mind: Concepts and Their Perceptual Basis*. Cambridge MA: MIT Press

Quine, W.V.O. 1951. Main Trends in Recent Philosophy: Two Dogmas of Empiricism. *The Philosophical Review* 60(1) (Jan., 1951): 20-43.

Ramsey, F. P. (1927/1990). Facts and Propositions. Reprinted in his *Philosophical Papers* (Ed. D. H. Mellor). Cambridge: Cambridge University Press.

Rosenthal, David M. 2002. Explaining Consciousness. In David J. Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings*, Oxford: Oxford University Press.

Searle, John R. 1980. Minds, Brains and Programs. *Behavioral and Brain Sciences* 3(3): 417-57.

Singer, Peter. 1975. *Animal Liberation*. New York: Harper Collins Books. 2nd edition, New York Review/Random House, 1990.

Smart, J.J.C. 1959. Sensations and Brain Processes. *The Philosophical Review* 68(2) (Apr., 1959): 141-156.

Stampe, Dennis W. 1977. Towards a Causal Theory of Linguistic Representation. *Midwest Studies in Philosophy* 2(1): 42-63.

Stich, Stephen P. 1978. Autonomous Psychology and the Belief-Desire Thesis. *The Monist* 61: 573-591.

Tanney, Julia. 2004. On the Conceptual, Psychological, and Moral Status of Zombies, Swamp-Beings, and Other 'Behaviourally Indistinguishable' Creatures. *Philosophy and Phenomenological Research* 69(1): 173-186.

Tang, Weng Hong. 2014. Success Semantics and Partial Belief. *Journal of Philosophical Research* 39: 17-22.

Tononi, Giulio. 2008. Consciousness as Integrated Information: a Provisional Manifesto. *The Biological Bulletin* 215: 216-242.

Tye, Michael. 1995. *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. Cambridge MA: MIT Press.

Tye, Michael. 1997. The Problem of Simple Minds: Is there Anything it is Like to Be a Honey Bee? *Philosophical Studies* 88: 289-317.

Wegner, Daniel M. and Thalia Wheatley. 1999. Apparent Mental Causation: Sources of the Experience of Will. *American Psychologist* 54: 480-492.

Whyte, J. T. 1990. Success Semantics. *Analysis* 50(3): 149-157.

Whyte, J. T. 1991. The Normal Rewards of Success. *Analysis* 51(2): 65-73.

Whyte, J. T. 1992. Review of Paul Horwich, Truth [1990]. *British Journal for the Philosophy of Science* 43(2): 279.

Whyte, J. T. 1993. Purpose and Content. *British Journal for the Philosophy of Science* 44(1): 45-60.

Whyte, J. T. 1997. Success Again: Replies to Brandom and Godfrey-Smith. *Analysis* 57(1): 84-88.