Michael Hegarty

Minds, Materialism and Mental Representation

Can naturalistic accounts of mental representation explain intentionality without circularity?

Summary

This essay argues that naturalism about mental representations is a failure: the matrix of explanatory requirements, ontological commitments and intuitions in naturalist accounts fails to result in a self-consistent notion of representation. Mental representations are posited to explain an intentional agent's behaviour and this explanatory role depends crucially on what the representation is about. Therefore it is necessary that representations have determinate content in virtue of which they cause behaviour. Naturalist accounts try to combine these explanatory requirements with a physicalist ontology in which intentional properties must be reduced to, or be supervenient on, the physical. Moreover, it is often demanded that two intuitions are respected: that representations are interpreted, and that there is a strong dividing line between mental and non-mental which should be recoverable from an account of mental representation. I argue that no consistent notion of representation can balance all these demands. The most labile commitments are the pre-theoretical intuitions, so I suggest either we radically alter these in light of theoretical results to persevere with a physicalist ontology, or we keep them and accept that physicalism cannot do them justice. Finally, I present a reason for choosing the latter position. I argue that the leading naturalist accounts still fail to yield determinate content in virtue of which representations are used to cause behaviour (i.e., they fail William Ramsey's job description challenge (JDC)). I suggest one plausible solution to this ('Representation as'), but which would be unacceptable to a naturalist. If there are no other options for naturalist accounts to secure the required determinate content other than 'Representation as' it follows that no naturalist account can pass the JDC. I suggest this results from the naturalist starting point that representations are subpersonal entities: only entities at the personal level are equipped to pass the JDC as Ramsey lays it out. Therefore, true representations are only at the personal level. Subpersonal posits in explanations of cognition may have explanatory value, but there is little relevant similarity between them and non-mental representations so as to validate thinking of them as representational as such.

Chapter 1 spells out a framework in which mental representations are posits in theories of cognition specified completely by their parent theories' structure and ontological commitments, the desiderata they seek to meet, and a set of pre-theoretical commitments about the nature of mind and representation. Chapter

2 outlines how naturalism and physicalism are related in the ontological commitments of naturalist theories of mental representation, and also introduces explaining intentionality as an important desideratum. Chapter 3 draws out two prominent pre-theoretical intuitions on mental representations, and examines how and why two leading philosophical theories of mental representation fail to yield determinate content or pass the JDC. Chapter 4 diagnoses a problem common to naturalistic theories and advocates a solution which is incompatible with naturalism, concluding that the explanatory goals of mental representations are ultimately unachievable in a naturalistic framework.

Chapter 1

1.1 Introduction

It is a fact of our mental lives that we think thoughts about particular things. My thought that it is a sunny day is about the condition of the weather today. Another commonplace observation is that what our thoughts are about guides our behaviour. A cliched example is that my belief that there is beer in the fridge, in conjunction with my desire to drink a beer, can be used to explain my going to the fridge to get a beer. Elementary reasoning like this — 'folk psychology' — is ubiquitous in everyday explanations of behaviour. Beliefs and desires, when talked about like this, are propositional attitudes. The proposition 'there is a beer in the fridge' is the 'content' of the propositional attitude of belief.

For folk psychological reasoning to work those propositions must be determinate. A belief-desire explanation would be unsatisfactory if the disjunction 'there is beer in the fridge or Paris is in France' were the content of my belief. The point is that to satisfactorily explain behaviour the content of propositional attitudes must be relevant to the behaviour explained, and also determinate, i.e. not indeterminate between different, non-disjunctive propositions. Folk psychological explanation, and indeed our experience, suggests we operate with thought contents (the term 'thought' here covers beliefs, desires etc.) that are determinate and recognisably so. Some have dubbed such explanations 'cognitivist' (Haugeland, 1978: 215),[1] and I will follow this for expediency. So a basic principle I will hold is: cognitivist explanations require that what a relevant thought is about plays a causal-explanatory role in explaining behaviour (Braddon-Mitchell and Jackson, 2007: 188).

A third preliminary observation about thoughts is that they can, and frequently do, concern objects and states of affairs which are absent or nonexistent. Our thoughts can be about, or directed upon, things not being currently perceived (electrons, Abraham Lincoln) and also those which we could never hope to perceive (unicorns). Yet thoughts are reliably about such things, and not others, in ways that figure prominently in cognitivist explanations of our actions. How? One popular view is that the mind is a representational device. When we have thoughts our minds have representations which bear some relevant relation to the actual things the thoughts are about.[2] A propositional attitude's content is a representation on this view. Thoughts are about things because the mind can represent those things. I will call all thoughts and

[1] According to Haugeland, cognitivism is "the position that intelligent behaviour can (only) be explained by appeal to internal 'cognitive processes', that is, rational thought in general".

[2] Relations only obtain between existing objects or properties so 'representation' cannot *strictly* be a relation as we frequently represent things which do not exist. I will not go further into this issue. Sceptics can view my use of 'relation' as a placeholder term for whatever sort of thing it turns out to be, if no understanding of relation strictly fits.

propositional attitudes 'mental states'. The representational view of mind then supposes that some mental states are, or involve, representations.

Representational mental states thus have contents which represent what the states are about. Determining how mental states get their contents is the job of a theory of mental content. This leads us to a simple schematic of how mind, world and thought relate. Let $\mathfrak{R}(p)$ be a mental representation with the content *p*. A content is generally presumed to be a proposition. The representation is about some object or state of affairs, and from this derives its particular content. Though my thought 'The Netherlands is a good place to live' is about the Netherlands, the content is more specific than simply 'The Netherlands'. For now it suffices to note that for $\mathfrak{R}(p)$ to represent some object or state of affairs, $\mathfrak{R}(p)$ must stand in a certain relation to the object or state of affairs which the mental state is about.[3] Elucidating the nature of this relation is a central problem in philosophy of mind.

So the picture is this: when some thinking subject, Q, thinks about something they token a mental representation $\mathfrak{R}(p)$ of that thing. $\mathfrak{R}(p)$ is about some object or state of affairs, $\mathcal{O}$, though $\mathcal{O}$ need not exist. I will call $\mathcal{O}$ the 'represented object' or 'represented'. The content, *p*, is related to, and derives from, $\mathcal{O}$. In Chapter 2 I will explain this relationship further. A representation stands in for the represented object in Q's thoughts. Filling out the details of this simple picture will be my task here. The nature of the representing relation — what it means for a mental state to represent something — is still a puzzle. I aim to formulate a definition of mental representation and assess whether or not naturalistic approaches to the mind can accommodate it. To do so requires a clear starting point on how to conceive of representation and its role in cognition. I will follow the dominant view that "representing is a functional *status* or *role* of a certain sort, and to be a representation is to have that status or role" (Haugeland, quoted in Ramsey, 2007: 23). In this chapter I will lay the groundwork for specifying what this role is. This requires stripping the term 'representation' of much theoretical baggage and outlining a framework within which to specify how its role is defined.

1.2 Towards a Theory-Neutral Conception of Representation

The concept of mental representation has accrued a lot of theoretical baggage by featuring in diverse theories of mind which invoke representations. For example, Jerry Fodor's language of thought (LOT) hypothesis

[3] There are disagreements over whether representations represent objects like 'that cat', properties like 'being red', or whole states of affairs like 'the red apple on the picnic bench'. My formulation is intended to be neutral on this issue, and I will not take a stand on it here.

(1975), and the computational view of mind in which LOT features, have fostered strong associations between mental representations, folk psychology, and the semantic and syntactic requirements of symbols in computational views of mind.

Yet influential work starting with Stephen Stich (1992), and continuing with William Ramsey (2007), challenges the essentiality of these associations. Peter Godfrey-Smith (2006: 43), for example, argues you needn't be a realist about propositional attitudes to have a notion of mental representation. He attempts to abstract the notion of 'mental representation' from the different theoretical and quasi-theoretical contexts in which it is enmeshed. I will use his work to form as theory-neutral a view about mental representations as possible.

Separating the concept of mental representation from folk psychology is perhaps most important for attaining this theory-neutral perspective. The naturalistic philosophical work on mental representations which I will be discussing is largely aligned with contemporary cognitive science. Cognitive science is concerned with explanations of cognition on the level of brain processes, thereby introducing a reason why folk psychological talk and mental representation should be separated.

Daniel Dennett observed that explanations of cognition and behaviour proceed along parallel lines he distinguished as the 'personal' and 'subpersonal' levels of explanation. He describes the personal as "the explanatory level of people and their sensations and activities" and the subpersonal level of explanation as "of brains and events in the nervous system" (1969: 93). Personal-level explanations reference things like intention, belief and desire — just like folk psychology. But cognitive science looks for subpersonal-level explanations by giving a causal-physical analysis in terms of neurological or computational mechanisms (Ramsey, 2016: 4). There is no place, according to Dennett, for personal-level phenomena like believing or seeing in explanations aimed at the subpersonal level, and vice versa. Dennett wants to prevent the "contamination of the physical story with unanalysable qualities or 'emergent phenomena'" (1969: 96). In his view a personal-level term like 'pain' does not refer to anything — to talk about pain in terms of neural processes is to describe a different phenomenon. A physical explanation of minds will thus not include personal terms at risk of committing a category error.

I think it is important to bear the distinction in mind as it imposes some order in the discussion of naturalistic theories of mental content, which often seek to explain personal-level phenomena in terms of subpersonal processes. I will follow Ramsey and other contemporary naturalistic philosophers of mind in separating folk psychology from mental representation. To engage with the naturalistic discussion I will for no align myself with their starting point that mental representations are subpersonal entities. I will also

follow Ramsey and Stich in holding that representations should be treated as posits within theories of mind with the goal of explaining cognition. This means that the nature of a mental representation is constrained in part by the particular commitments of the theory of cognition in which it is embedded.

My goal is to give an account of the representing relation. In the next section I summarise an important consideration raised by Ramsey in this regard.

1.3 Ramsey's Job Description Challenge

I mentioned above that mental representation is a functional role, and said that one of my core principles is that a representation's content must play a causal-explanatory role. One of Ramsey's important contributions to the debate is to pull these two aspects apart. He argues — I think convincingly — that a full theory of mental representation should have two parts: an account of how mental representations get their content, and a description of a representation's functional role (Ramsey, 2007: xv). Ramsey claims there has been too much focus on theories of content at the expense of getting clear on the functional role (2007: xii; 29). The main message of his book *Representation Reconsidered* is that, where mental representations are invoked in theories of cognition, the theory must be justified in claiming the posits labelled 'mental representation' are actually playing representational roles. He argues that many contemporary theories of cognition invoke mental representations which, on closer inspection, fail to actually play representational roles. For clarity, I will call such posits 'pseudo-representations'.

Clearly, judging if a posit plays a representational role requires an account of what it means for something to function as a representation. Ramseys' project is to specify this functional role, or in other words to give a 'job description' of a posit for it to qualify as a representation. For this reason he terms the challenge he lays down to theories of cognition the 'job description challenge' (JDC). Ramsey explains the JDC: "There needs to be some unique role or set of causal relations that warrants our saying some structure or state serves a representational function… I'll refer to the task of specifying such a role as the '*job description challenge*.'" (2007: 27) Any posit which is called a representation in the language of the theory, but fails the JDC, is merely a pseudo-representation.

As a result of the causal-explanatory role representations are supposed to play in cognitivist explanations, it is clear that part of the motivation for positing representations is because of their explanatory value. The breakdown of cognition into, for example, computational states and representational tokens —as in the computational theory of mind — is only motivated because it serves an explanatory goal. If the theory posits entities which add no explanatory value then — aside from considerations like coherence with other

theories — there is no reason to keep those entities in the theory. Explanatory relevance is one of Ramsey's central motivations (2007: 28). The idea can be expressed like this:

> *Explanatory Purchase*: if we remove mention of 'representation' in a theory of cognition and the theory's explanatory power is undiminished, then the notion of representation in the account in question fails to be truly representational.

A posit's functioning as a representation has to be essential to the explanatory ambit of the theory in which that posit is embedded. If not then it is a pseudo-representation. Ramsey's aim in putting forward the JDC is to bring out the functional role of a representation in a theory by examining how it achieves *Explanatory Purchase* on cognition.

What does this functional role consist in? Ramsey invokes a 'use' condition: representations should represent a state of affairs and thereby allow an agent to use the representation to cognise or behave in relation to that state of affairs. The representation must be used *as* a representation for the state to function as a representation. As we have seen, representations have particular contents. It is in virtue of having a particular content that a representation represents what it does.

But having a particular content and being used by a system in a certain way to cause behaviour are logically separable. John Searle first made the point in his Chinese Room thought experiment (Searle, 1980):

> *Chinese Room*: an English-speaking man who knows no Chinese is locked in a room with a set of Chinese symbols and a rule book. The book pairs sets of Chinese symbols with each other. When certain Chinese symbols are passed into the room — in the form of questions from Chinese-speaking external observers — he uses the book to select and pass the corresponding paired symbols out again. It seems to the Chinese speakers that the questions are being answered correctly by a fluent Chinese speaker.

Searle's originally argued against strong varieties of artificial intelligence which suggested that running the right computer program was sufficient for something to have a mind. He exploits the intuition that since the man (playing the role of a CPU) understands no Chinese a computer program does not either. Mere formal manipulation of the symbols can never convey understanding of Chinese to the man. The idea is that

successful syntactic manipulation is insufficient to recover semantic interpretation of the symbols manipulated.

Ramsey notes that Searle's argument also counts against a representational interpretation of the symbols operated on in the computational theory of mind (2007: 48-49). *Chinese Room* presents a case of someone behaving exactly as if they do in fact understand the meaning of the Chinese symbols, though they do not. Nevertheless the symbols do have meaning independently because a Chinese speaker can understand them. So *Chinese Room* shows a case where a symbol is used in a role that seems representational, has the appropriate content, and causes behaviour in the representation user commensurate with their recognising that content. The content of a Chinese symbol would be correct for the causal-explanatory role of the representation to be played in line with a cognitivist explanation, though it seems clear that the symbol is not used as a representation *in virtue* of that content. That is, the symbol has not played a representational role, though all other conditions for it being a representation are apparently met.

This means that "to be a representation, a state or structure must not only have content, but it must also be the case that this content is in some way pertinent to how it is used" (Ramsey, 2007: 27) .Conjoining this insight with *Explanatory Purchase* we can say that a state only qualifies as a representation if it possesses a content that is explanatorily relevant and the system uses the representation in virtue of that content. Ramsey writes: "*real* mental representations — things like our thoughts and ideas — intuitively interact with one another and produce behaviour by virtue of what they are about" (2007: 48).

So we see that both $\mathfrak{R}(p)$ having the particular content *p*, and $\mathfrak{R}(p)$ being used as a representation by the system are necessary conditions for $\mathfrak{R}(p)$ being a representation. However, they are jointly insufficient. What must be the case for $\mathfrak{R}(p)$ to pass the JDC is that $\mathfrak{R}(p)$ must be used by Q in virtue of having the content *p*. I propose *JDC Pass* — which is a specific way of cashing out *Explanatory Purchase* — to establish what conditions must be met for passing the JDC:

> *JDC Pass*: a theory of cognition, C, which posits a representation, $\mathfrak{R}(p)$, with content *p* passes the JDC if and only if (i) a representation user, Q, within the architecture of C uses $\mathfrak{R}(p)$ as a representation in virtue of the content *p*; and (ii) *p* is determinate and explanatorily relevant.

The functional role of representing is still unexplained. All *JDC Pass* specifies is that any representational theory has to have *Explanatory Purchase*.

To summarise, I have set out Ramsey's job description challenge (JDC). Some apparently representational posit in a theory of cognition must pass the JDC (by fulfilling the conditions in *JDC Pass*) to qualify as a representation, according to Ramsey. If it fails, it is a mere pseudo-representation and does not play a recognisably representational role in the theory: the posit's being thought of as a representation is irrelevant to the theory's explanatory ambit.

1.4 Three Kinds of Constraint on a Theory of Mental Representation

I am following Ramsey in his conception of a representation as a theoretical posit, the utility of which is measured by the explanatory power it brings to the phenomena the theory aims to explain (Ramsey, 2007: 5). Thus, in trying to specify the functional role of a representation we must look at the theory in which it is embedded. As I see it, the commitments of that theory form one of three kinds of constraint on the functional role of a mental representation when viewed as a theoretical posit in this way. These three kinds are: the internal structure and background commitments of a theory of cognition, the desiderata the theory seeks to meet, and the pre-theoretical intuitions behind the theory.

Each theory has particular commitments and background assumptions which will affect the nature of the representational posit. For example, many theories — especially the naturalistic ones I consider here — are set within a physicalist ontology. Also, the internal structure of the theory affects how we interpret the nature of the representational posit. A computational theory of mind takes a representation to be a symbol token with particular semantic content which adheres to certain syntactical rules. In connectionism, representations are distributed and constructed from nodes with different weights of activation (Garson, 2016). I will say that these commitments around internal structure and background assumptions (including ontology) specify the 'form' of the theory. The theory's form is then one constraint on the nature of the representation posited in that theory.

The second constraint is what the theory aims to explain, or its desiderata. Mental representations are posits in theories of cognition: theories that explain our mental capacities. Therefore, a theory of cognition should explain those mental capacities. Its success at meeting those desiderata in turn is a measure of its success as a theory. For example, Fodor's LOT is successful because it succeeds in explaining the compositionality, productivity and systematicity of thought.

Theory form and desiderata are both constraints on mental representations internal to particular theories of cognition. The third constraint which I have identified is crucial, and also the only extra-

theoretical constraint: our pre-theoretical intuitions about what representing is. One example of how pre-theoretical intuitions encroach on the notion of mental representation is the prevalence of folk psychology in the discussion. For example, LOT aims to explain belief and desire talk in computational terms, thereby taking our pre-theoretical view that there are beliefs and desires that serve in cognitivist explanations and building it into the theory from the outset. The pre-theoretical constraint is brought out by Ramsey's JDC: we can only judge a posit to be a representation if we already have some idea of what representations are. The preponderance of metaphor and analogy around mental representations — viewing them as like maps or pictures — further supports this view. Thus our pre-theoretical intuition is a constraint that must be met by a representation-invoking theory because otherwise there is no justification for talking about minds as representational. It is only by analogy and metaphor with non-mental representations that we have any purchase on what a mind being representational means. This therefore entails that if we think of minds as representational we are already helping ourselves to some kind of preexisting, non-mental notion of representation.

If we view mental representations as posits in theories of cognition, then the form of the theory, the theory's desiderata, and the pre-theoretical intuitions adhered to jointly specify the properties and functional role of the mental representation. For the rest of the chapter I will spell this out in more detail and try to derive a basic, theory-neutral notion of representation with which to work for the rest of the essay. This will rely on conceiving of mental representations as mental models used in reasoning.

1.5 Mental Models and Mental Representations

The idea of a mental representation is, at its most basic, that of an internal state related to some object or state of affairs which it represents. From this starting point I will try to derive a general notion of representation which, suitably augmented, could fit into the major theories like computationalism, connectionism etc.

Following Godfrey-Smith, I take the 'mental model' conception to embody the most pure notion of mental representation at the heart of these theories. This way of thinking about representations traces its roots to Kenneth Craik (1943), and has been popularised in contemporary thought in works by Godfrey-Smith (2006), Chris Swoyer (1991), Ramsey (2007), and Philip Johnson-Laird (1989). The idea is essentially that representations are 'stand-ins' or 'surrogates' for those objects or states of affairs which they represent. An agent can utilise those surrogates in planning and executing actions in virtue of the relevant similarity the model bears to the domain which the agent is considering.

'Standing in for' is implicit or explicit in many philosophical discussions of mental representation. Ruth Millikan follows it in her theory of mental content (Millikan, 2009: 397-398). Moreover, according to Andy Clark, activities like dreaming of Paris or planning a future holiday "require the brain to use internal *stand-ins* for potentially absent, abstract, or non-existent states of affairs" (Clark, 2001: 29). Thus we see how the mental model conception can help explain our ability to reason about objects even in their absence, as was mentioned in §I.i.

Craik's original idea uses the observation that workers like engineers use models to test out features of a future construction. A civil engineer may construct a scale model of a bridge to test its performance under various conditions. This is a form of reasoning in the absence of what is being reasoned about (a currently non-existent, hypothetical bridge) to test out theories and solve problems. Craik's insight is that mental representations allow minds to do similar tasks. So on this view a mental representation is an internal model ('stand-in', or 'surrogate') of whatever is being reasoned about (1943: 61). Godfrey-Smith puts it well: "a representation is one thing that is taken to stand for another, in a way relevant to the control of behavior or some other decision… when a person decides to control their behavior towards one domain, Y, by attending to the state of something else, X. The state of X is 'consulted' in working out how to behave in relation to Y." (2006: 45)

Craik's original presentation built in a resemblance between model and state of affairs. However, the core idea of a mental model remains in place without having to account for the relevant similarity of model and modelled through resemblance, or isomorphism. Those constitute just two of the options for how a mental representation 'stands in for' its represented object. For clarity, I will use the colourless term 'stand in for' rather than, say, 'represent' or 'mean'. My intention is that 'standing in for' is an idea prior to any of these other terms; it could be explicated in terms of resemblance, for example. But any such claim must be argued for.

I follow Godfrey-Smith and Ramsey in holding that the mental model conception captures the sense of representation while standing prior to developed theories. According to Ramsey: "Computational explanation often appeals to mental models or simulations to account for how we perform various cognitive tasks. Computational symbols serve as elements of such models, and, as such, must *stand in for* (i.e., represent) elements or aspects of that which is being modeled." (2007: xiv) Godfrey-Smith gives the example of a person using a map: they consult the map as a guide to a particular target domain which the map models. He calls this the 'basic representational model' and notes how the starting point is with the "basic, everyday sense" of 'representation'.

In the next section I will synthesise what has been learned about mental representations through the JDC, the mental models conception, and the three families of constraint to derive a relatively theory-neutral notion of mental representation.

1.6 A Schematic Theory of Representation

So far I have laid out my support for a particular view of mental representations which treats them as posits in theories of cognition. I have followed Godfrey-Smith and others in taking the mental model conception to be the foundational notion common to all representational theories. Through plugging this notion into any theory of cognition, and examining how the theory's form and desiderata constrain the role the representation plays I contend we can specify the representational role for any given theory. My ultimate goal is to use this way of thinking to construct a solid foundation for thinking about mental representations, and then build on the conception of mental representation that falls out of this approach to assess whether or not intentionality can be explained naturalistically.

In general, then, what a mental representation is like (what properties it has, the nature of its functional role) is determined by the role it plays in the theory — its explanatory role. At root, the explanatory role is always that the representation is a 'stand in' for an object or state of affairs cognised. However, this is further augmented by the nature of the specific theory. Thus, the following questions arise, given any particular theory of cognition:

(A): What is the form of the theory?

(B): What explanatory role do mental representations play in that theory?

(C): What does a mental representation need to be like to play that explanatory role?

The answers are interdependent. The answer to (B) will be determined by (A): what the assumptions, commitments and structure of the theory in question are, and the desiderata it aims to meet. Furthermore, the answer to (C) is entirely dependent on specifying the explanatory role the representation plays in the theory. In the remainder of this chapter I will seek to use this way of thinking to derive a basic notion of mental representation, building on the 'mental model' conception. A first attempt at this might be:

(*MR*): $\mathfrak{R}(p) =_{Def}$ an internal state, $\mathcal{S}$, that stands in for some object or state of affairs, $\mathcal{O}$, by having a content p.

Mental representations are, by definition, internal to agents. (*MR*) includes this internality requirement and captures the basic insight of the mental model view that representations 'stand in for' their represented objects. 'Standing in for', when suitably explicated, really amounts to the representation's functional role.

(*MR*) can be further refined. $\mathfrak{R}(p)$ has *p* as its content. Following Ramsey's bifurcation of a theory of representation into a theory of content plus an account of functional role I take it that, given a suitable theory of content, *p* could be derived appropriately from $\mathcal{O}$.[4] Nevertheless, *Chinese Room* shows it is logically possible for a state with a determinate content to cause behaviour in a way that would make a representational explanation appropriate, yet for the theory to still lack *Explanatory Purchase*. In other words, it seems to be a representation and functions indistinguishably from a representation, but fails to be *used as a representation by the system*. These considerations highlight two shortcomings of (*MR*). Firstly, part of the role of $\mathfrak{R}(p)$ is to be capable of causing behaviour. Secondly, $\mathfrak{R}(p)$ must do so in virtue of being about, or standing in for, $\mathcal{O}$. So a modified version of (*MR*):

(*MR2*): $\mathfrak{R}(p) =_{Def}$ an internal state, $\mathcal{S}$, of Q that stands in for some object or state of affairs, $\mathcal{O}$, (by having a content, *p*) such that $\mathcal{S}$ is capable of causing behaviour of Q in virtue of standing in for $\mathcal{O}$.

I have postponed detailed discussion of technical terms like 'content' for the moment because a proper understanding requires a fuller conception of the aboutness of minds. I will introduce this discussion in the next chapter, and further develop (*MR2*) by introducing intentionality as the desideratum constraint.

[4] For example, it could be the function of whatever brain state $\mathfrak{R}(p)$ corresponds with to indicate the presence of $\mathcal{O}$, following a simplified version of Fred Dretske's indicator semantics. Another option is Ruth Millikan's biosemantics. Both positions will be discussed in later chapters.

Chapter 2

In this chapter I will finish laying the groundwork for my argument by setting out the metaphysical stakes in the contemporary philosophy of mind debate. Most of the theories I consider are naturalistic, and thus are situated within physicalist ontologies. This figures in the theory form constraint from Chapter 1. I will then introduce explaining intentionality as a desideratum for a theory of cognition. From there I will be able to explore whether the account of mental representation that emerges is consistent with the physicalist background and explanatory aims of such a theory. I will conclude that it is not, and that this shows some of the commitments among theory form, desiderata and pre-theoretical intuitions are in conflict.

2.1 Physicalism, Dualism and Naturalism

The form of a theory includes the ontology within which the theory is embedded. All the theories of mental representation which I consider in this essay are formulated against a background physicalist ontology. To understand how this ontology impacts the notion of mental representation it is necessary to set out physicalism. Physicalism is in competition with dualism as the primary ontology of mind. Therefore, I will explicate physicalism in relation to dualism.

Theorising about the mind begins with talk of things like ‘thoughts’, ‘beliefs’ and ‘intentions’. Though these ways of speaking are successful, via folk psychology, this need not imply mental entities and propositional attitudes are part of our ontology (Dennett, 1969: 9-14). It is important to bear in mind this distinction between ways of talking about the mental and what there is. This leaves it an open question whether or not there are such things as beliefs. Moreover, even if mental talk is useful because it refers to actual entities this still need not imply there is a world of mental things independently of the physical world. One can be a physicalist and still acknowledge the necessity of “non-physical aspects of existence and non-physical truths” in a system of explanation (Poland, 1994: 12).

To a first approximation, physicalism says that everything that there is can ultimately be explained in terms of the entities of physics (Crane and Farkas, 2004: 603-604). Three versions of this general position are useful to consider here: reductive physicalism, non-reductive physicalism and eliminativism.

Eliminativists hold that the physical exhausts what there is and physics says all there is to say about the world (Poland, 1994: 12). To put this in terms of mental talk versus mental being, the eliminativist holds that mental terms are not referring or even useful (Churchland, 1981). They will ultimately be replaced by accurate physical descriptions.

Reductive physicalists hold that there are mental things, like mental properties such as ‘pain’, but that these can be reduced to some physical properties or physical states. Unlike eliminativism, this position does not denigrate the use of mental talk. The paradigmatic statement of this sort of view is identity theory (Kim, 1996: 52-60). A mental state like being in pain simply is identical with a particular physical state, like having one’s C-fibres firing. This is an ‘ontological reduction’ where mental properties like being in pain are reduced to physical properties. The hallmark of the reductive physicalist is that they hold there are no non-physical properties (Kim, 1996: 212).

The most widely held physicalist view, and the most important for my discussion, is non-reductive physicalism. In contrast to reductive physicalists, non-reductive physicalists hold that there are non-physical properties, including all the commonsense mental properties like being in pain. Though these are supposed to be irreducible to physical properties, all mental properties are held to be ultimately dependent on the physical. Fixing all the physical properties in the world would simultaneously determine all the mental properties (Poland, 1994: 16). To introduce some terminology from Jeffrey Poland, non-reductive physicalists hold that everything is ontologically grounded in the physical domain. Poland conceives of a “hierarchically structured system of objects grounded in a physical basis by a relation of *realisation*” (1994: 18). To say all attributes are realised by physical attributes is just to say that for any attribute we can specify it by specifying the arrangement(s — for there may be many different physical arrangements which entail that the attribute of being brittle is instantiated, for example) of underlying physical constituents. Thus non-reductive physicalism claims that mental properties exist, but that they *supervene* on physical properties.

Further, non-reductive physicalism claims that all phenomena are explanatorily grounded in the physical. This means that, though some mental phenomena can be explained in higher-level psychological terms for example, nevertheless the psychological phenomena are traceable to ultimate physical facts, the obtaining of which then explains those phenomena. Psychological facts supervene on the physical. Phenomena and regularities in physics ultimately form the bedrock on which explanations at all levels of generality rest, according to the physicalist (Poland, 1994: 21). However, this is not to say we might be able to explain, say, psychological laws — if such there are — in the language of physics.

So non-reductive physicalism is the position that everything is “dependent on, supervenient upon, and realised by physical phenomena”(Poland, 1994: 22). When I talk about physicalism in this essay I will refer to this non-reductive variant, unless otherwise specified:

> *Physicalism*: (i) there are non-physical, 'mental' properties which are explanatorily valuable and physically non-reducible but, (ii) all mental properties supervene on the physical and (iii) all mental properties are realised by physical facts.

Physicalism is compatible with the view that there are non-mental properties, so long as those non-mental properties are shown to depend upon, supervene on and be realised by the physical. It is only this realisation and supervenience requirement which separates it from a form of dualism which combines substance monism with non-reducible, foundational mental properties (Kim, 1996: 228). This is Property Dualism, the view that mental properties are 'emergent' from the physical, such that they can vary independently of their physical basis. Emergentism says that the correlations between physical states and mental properties are brute facts not reducible to, or explainable in virtue of, their physical bases (1996: 52).

The close similarity between *Physicalism* and Property Dualism leaves a grey area for mental properties. As Jaegwon Kim notes, emergentists might deny that *Physicalism* adequately explains the relationship between mental properties and their physical bases if, say, it claims C-fibres firing just is (or causes) pain. The real explanatory work needed is to say why the sensation of pain is how it is rather than a tickle (1996: 52-53).

Most philosophers I consider in this essay claim to offer naturalistic theories of the mind. Naturalism is less well-defined than *Physicalism*, though it is broadly taken to fit alongside some form of physicalism, and to deny dualism. To best characterise the ontological commitments of these philosophers I will try to explicate what I think they have in mind: broadly, that there is no place in the world for supernatural entities. Much naturalistic philosophy relies on the principle of causal closure under physics: the idea that all physical effects have sufficient physical causes (Yalowitz, 2014). This in turn means that anything that has physical effects must be physical too. Naturalists generally affirm causal closure alongside side the view that all events are ultimately supervenient on the physical. They are then in the business of searching for physical causes of, and explanations for, effects we speak of as mental.

It is usually supposed that mental representations are both intentional (see §2.2) and bear semantic properties, and that neither intentionality nor the basis of semantic properties is to be found in nature. Thus the naturalist seeks to recover semantic and intentional properties by explaining how representations acquire them from more basic, non-intentional and non-semantic physical things in the world (Ramsey, 2007: 18). As Fodor (1984) puts it:

> The worry about representation is… that the semantic/intentional properties of things will fail to supervene upon their physical properties. What is required to relieve the worry is therefore, at a minimum, the framing of *naturalistic* conditions for representation. […] [W]hat we want at a minimum is something of the form '*R represents S' is true iff C* where the vocabulary in which condition C is couched contains neither intentional nor semantical expressions. (p 232)

This means that explicating condition C in purely physicalist, naturalist terms (without using intentional or semantic terms) is the aim of a naturalistic theory of mental representation. This is the link between naturalism and *Physicalism* for my purposes. Henceforth, I will take naturalistic approaches to affirm *Physicalism* and refer to this position generally as 'representational naturalism'.

With the these definitions in hand I can detail what role they play in my argument. My claim is that a naturalistically conceived C can never make 'R represents S' true unless some of the shared pre-theoretical commitments we hold about minds, representations and intentionality are sacrificed. The choice is ultimately between intentionality and *Physicalism* because accounting for some of the intuitive commitments around representation is incompatible with representational naturalism due to the demands of adhering to *Physicalism*.

This requires arguing that something we take to be central to representation will not submit to representational naturalism. This amounts, I think, to specifying a counterexample to *Physicalism*. As Poland writes, if there are "higher-level phenomena that admit of non lower-level explanation… [this] provides direct counter-examples to physicalism and must be avoided if the programme is to be successful." (1994: 23) I think the following necessary condition must be fulfilled for some phenomenon to be a counterexample:

> *Counterexample*: if there exists some x — where x is an object, property or relation — and (i) x does not supervene on the physical, or (ii) x cannot be explained in lower-level physical terms, then x is a counterexample to *Physicalism*.

However, in principle one can argue that *Counterexample* is necessary but not sufficient to identify a true counterexample to *Physicalism*. Maybe our currently incomplete physics is incapable of explaining a mental phenomenon which is a *Counterexample*. This is not enough to prove future physics will never be able to explain it. One response to this is to argue that the mental phenomenon in question is not something physics might hope to explain. As we shall see, intentionality is often taken to be one such phenomenon (Fodor, 1987: 97).

Thus a *Counterexample* seems unable to definitively refute *Physicalism*. But by the same token defenders of *Physicalism* are prevented from giving knock-down arguments that a *Counterexample* will ultimately be explained in physicalist terms. A *Counterexample* therefore at best lends weight to the claim that there are foundational mental phenomena in line with Property Dualism, to the detriment of *Physicalism*, but without being decisive against it.

In the next section I will set out one crucial desideratum — explaining intentionality — which a theory of cognition positing mental representations should meet. I will ultimately claim that the failure to explain intentionality means it is a *Counterexample* to *Physicalism*, and that recognising this leaves us with a straight choice between abandoning representational naturalism, or reevaluating our pre-theoretical intuitions.

2.2 Intentionality

So far I have described thoughts and representations as 'being about' their represented objects by having some content. The idea of representational content, or 'aboutness', is captured by the technical term 'intentionality' (Crane, 2003: 30). Though some philosophers argue intentionality is better construed as 'directedness' I will think of it primarily as 'aboutness', as do the naturalistic philosophers I focus on here (Crane, 1998a: 233). Following Ramsey, I will assume that representations are the bearers of intentionality. One of the explanatory goals of a theory of cognition is to explain how mental states are intentional. Representations are posited to do this. Thus I will take explaining intentionality as one desideratum a representational theory should meet, in line with the framework established in Chapter 1. I will examine whether representational naturalism can support a consistent notion of representation when aiming at this explanatory goal.

As shown in the Fodor quotation in §2.1, intentionality cannot be treated as basic under *Physicalism* — it must be explained in physical terms. Being intentional is a property of things. Therefore it is one of those properties which either supervenes on, or is emergent from, the physical.

Intentionality was introduced into contemporary philosophy by Franz Brentano who claimed that every mental phenomenon was characterised by 'intentional inexistence' (Brentano, 1995: 68). "Every mental phenomenon includes something as object within itself, although they do not all do so in the same way. In presentation something is presented, in judgement something is affirmed or denied, in love loved, in hate hated, in desire desired and so on," he wrote in explanation of 'intentional inexistence'. For Brentano intentional inexistence — or simply 'intentionality' — was "characteristic exclusively of mental phenomena. No physical phenomenon exhibits anything like it".

Brentano's claim has been quoted and discussed widely, and 'intentional inexistence' forms the basis for the study of intentionality in contemporary philosophy. 'Inexistence' here means, roughly, an existence *in* something else. In this case, objects or states of affairs thought about have intentional inexistence in that they exist *in* the act of thought. Though the things one thinks about need not exist in the real world because the object thought about and the content of the thought are distinct, the *content* of one's thought always exists.

I will follow Tim Crane and distinguish 'intentional object' and 'object of thought'. The intentional object is what one thinks, while the object of thought is what that thought is about (Crane, 2001: 22). My thought about beer in the fridge is about the beer and the fridge — i.e., the beer itself is the object of thought — while my mind is, to speak metaphorically, directed on the intentional object which represents the beer in the fridge but is not itself beer in a fridge. Crane also holds that intentional objects are always presented in some particular way in thought. He calls this the 'aspectual shape' of the intentional object. Fred Dretske (1995) captures this idea well:

> In thinking about a ball I think about it in one way rather than another — as red not blue, as round not square, as stationary not moving. These are the aspects under which I think of the ball… Our mental states not only have a reference, an aboutness, an object (or purported object) that forms their topic; they represent that object in one way rather than another. When an object is represented, there is always an aspect under which it is represented. (pp30-32)

I can think about the beer in a number of different ways — under different aspectual shapes — but I must always think about it in *some* way. Irrespective of the aspectual shape, my thought about the beer always corresponds with the same beer.

Brentano is usually taken to defend what is known as 'Brentano's Thesis': all and only mental phenomena, or states, exhibit intentionality (Crane, 1998b: 819). Brentano's view was that the mental is irreducibly intentional (Crane, 2001: 12). In analytic philosophy, Brentano is understood to be claiming that intentionality is a criterion for distinguishing between entities in the world: anything exhibiting intentionality cannot be a physical entity (Crane, 1998b: 817-818). However, if we understand intentionality in terms of aboutness there are myriad examples of non-mental things which are still intentional: words, pictures and maps are three common non-mental things with apparent intentional properties. These simple counterexamples suggest Brentano's Thesis must be wrong. In answer to these well-motivated concerns, it has become commonplace to distinguish between original and derived (or derivative) intentionality. It is a popular idea that physical objects like maps and roadsigns are about things in a way derivative on the contents of thoughts. According to Alex Byrne: "A thing has derivative intentionality just in case the fact that

it represents such-and-such can be explained in terms of the intentionality of something else; otherwise it has original intentionality." (Byrne, 2006: 408)

The original/derived intentionality distinction raises another important issue. Maps and roadsigns represent and succeed in being about things other than themselves only by virtue of being interpreted by agents with some knowledge or understanding of how to take those symbols as standing for something else. Without this knowledge no representation can occur; reading a map without a key is impossible. But as interpreters seem essential to the representing abilities of symbols with derived intentionality, this means there must be agents with minds to bestow derived intentionality on those symbols.

So long as we can appeal to an interpreter with a mind we can satisfactorily explain how symbols function as representations. This view makes representing a 3-place relation between representation, object represented and interpreter.[5] However, this model will not do for explaining original intentionality. If we import the 3-place relation directly into minds it loses its explanatory power. We set out to explain the intentionality of minds, a property only minds were supposed to have. Positing an interpreter with a mind to interpret the internal intentional symbol (the mental representation) leads to circularity because the internal interpreter seemingly has to have a mind — with intentional capabilities — to perform its interpretative role. This leads to a vicious 'Homunculus Regress'. Avoiding this is a major obstacle to any theory of mental representation:

> *Homunculus Regress*: any theory of mental representation which requires positing an agent with a mind to interpret the representation leads to vicious infinite regress.

Closely linked with *Homunculus Regress* is another pitfall for theories of intentionality and, concomitantly, of theories of mental representation. No theory of intentionality or mental representation can invoke intentional terms in its account or definition of representation or intentionality, or will also run in a circle.

What is an 'intentional term'? This is simply a term in language which is used to label an (original) intentional state. I introduced this discussion by saying that thoughts and other mental states are held to be intentional. Intentional terms are linguistic terms used to refer to or report such mental states. Examples include 'believes', 'knows', and 'perceives'.

---

[5] A view discussed at length by von Eckardt (Barbara von Eckardt, *What is Cognitive Science?*, (Cambridge, MA: MIT Press, 1993), 147-195).

Roderick Chisholm (1974) presents an influential discussion of intentional talk. He recasts Brentano's Thesis into a linguistic form and claims that reports of intentional states all result in sentences that are intensional in the logical sense. One of Chisholm's motivations in doing this is to recreate in a linguistic form the fact that intentional states are about things which may or may not exist, just like intensional sentences can be about things without their subjects having to exist.

Chisholm's work is important because contemporary analytic philosophy has largely followed his interpretation of intentionality (Crane, 1998b: 818). It is also useful because of his detailed discussion of intentional terms. No definition or account of intentionality can include an intentional term at risk of being circular. By extension, as all mental representations are intentional, no definition or account of mental representation can include an intentional term for the same reason. Ultimately, I think this comes to the same thing as the *Homunculus Regress*: intentional terms are all agent-involving terms, so any mention of them in a definition of mental representation thereby invokes some full-blown agent with a mind. Thus the definition fails to be informative.

For the representational naturalist an acceptable account of mental representation or (original) intentionality would explain the phenomenon in terms of lower-level and more fundamental capacities. Note that intentional talk is all personal-level vocabulary. This means that an explanation of intentionality, and hence mental representation, cannot appeal to persons or phenomena on the personal level. *Homunculus Regress* occurs when explanations or definitions invoke personal-level concepts and vocabulary when a subpersonal explanation is sought. Preventing this mixing of levels of explanation is precisely why Dennett introduced the personal/subpersonal distinction in the first place (1969: 93-97).

If, as Ramsey and others hold, mental representations are subpersonal posits in theories of cognition which are supposed to explain intentionality, it is clearly a major error to invoke personal-level concepts like 'seeing that' or 'believing that' in a theory of mental representation. Subpersonal explanations are supposed to operate below the level of persons and are reductive explanations which should explain what we observe at the level of persons in terms of things which are at a more basic level than persons. I will ultimately question whether this strategy can work for mental representation.

As a final important point, Chisholm used his linguistic formulation of intentionality and argued that intentional phenomena like believing and perceiving cannot be specified in non-intentional terms (1974: 180ff). This supports Brentano's Thesis that mental phenomena are irreducibly intentional. Chisholm concluded from this that, as the language of physics has no room for intentional terms, then reduction of

intentional phenomena to physical phenomena can never succeed. In other words, the intentional by its nature resists reduction to the physical, and so reductive physicalism is false (Crane, 1998b: 818).

In *Word and Object*, W. V. Quine makes a similar point to Chisholm but reaches a different conclusion. In contrast to Chisholm, who held that the indispensability of intentional ways of speaking suggested intentionality was a 'mark of the mental' (1974: 181), Quine protested that the apparent irreducibility of intentional ways of speaking was an argument for the unreality of intentionality. He wrote: "One may accept the Brentano thesis either as showing the indispensability of intentional idioms and the importance of an autonomous science of intention, or as showing the baselessness of intentional idioms and the emptiness of a science of intention." (Quine, 1960: 221) Chisholm's and Quine's work on intentionality thus presents a dilemma (Crane, 1998b: 818):

> *Chisholm-Quine Dilemma*: intentionality cannot be reduced to physical processes or talk, hence reductive physicalism can never explain it. Therefore either intentionality is real and reductive physicalism is false, or intentionality is illusory.

Recall that *Physicalism* has three parts: (i) there are non-physical, 'mental' properties which are explanatorily valuable and physically non-reducible, (ii) all mental properties supervene on the physical, and (iii) all mental properties are realised by physical facts. As I characterised it here, intentionality is a mental property. If we lack an adequate account of how intentionality supervenes on and is realised by the physical then intentionality appears to be a *Counterexample*. As I pointed out in the last section, a *Counterexample* does not demonstrate that the non-reductive *Physicalism* is false, but does strongly suggests a stalemate between *Physicalism* and Property Dualism.

As Crane notes, most of the early work on mental representation and content concentrated on finding a way between the horns of the *Chisholm-Quine Dilemma* (1998b: 818). The popular strategy has been to try to reconcile intentional realism and cognitivism with representational naturalism. Rather than side with Quine against intentional talk, or with Chisholm for the irreducibility of the intentional, work focused on a naturalistic account of intentionality via explaining the content of mental representations. Therefore the stakes are high in finding a physical basis for intentionality: fail and *Physicalism* and Property Dualism become almost equally plausible ontologies of mind.

The remainder of this essay will argue that representational naturalism has failed. There are two reasons for this. Firstly, theories of content fail to yield content sufficiently determinate to play the causal-

explanatory role required of it in cognitivist explanations. Secondly, those theories also fail to convince that they have *Explanatory Purchase*: they fail the JDC. I will suggest one amendment to (*MR2*) which seems to satisfy *JDC Pass* and thereby solve this problem. However, this amendment introduces intentional talk into the definition and thereby fails to be the naturalistic explanation sought to navigate the *Chisholm-Quine Dilemma*. I will suggest that much criticism of theories of representation ultimately stems from a failure to appreciate the incompatibility of the irreducible intentionality buried deep in our intuitions about the mental, and the constraints of *Physicalism* and hence of representational naturalism.

Chapter 3

To recap, in Chapter 1 I set out a framework viewing mental representations as posits in theories of cognition constrained by the theory's form, the desiderata it seeks to explain, and pre-theoretical intuitions. For a cognitivist explanation of behaviour to succeed, representational content must be determinate to play a causal-explanatory role. Moreover, according to Ramsey's job description challenge (JDC) a representation must be used as a representation — whatever that functional role amounts to — in virtue of that particular content. In Chapter 2 I drew out the *Chisholm-Quine Dilemma* and argued that intentionality seems to serve as a *Counterexample* to *Physicalism* in the absence of a convincing naturalistic account of mental representation. Such an account would require explaining mental representation in non-intentional terms or face a vicious *Homunculus Regress*. This chapter will consider two families of naturalistic theory which attempt to navigate the C*hisholm-Quine Dilemma*. I will argue they ultimately fail to recover content sufficiently determinate to figure in cognitivist explanations, and that in any case they use only pseudo-representations because the theories lack *Explanatory Purchase* and fail the JDC. I begin in §3.1 with a discussion of what common pre-theoretical intuitions constrain naturalistic theories of mental representation.

3.1 The Intuition Constraint on Mental Representation

'Representation' is not a purely technical term as 'intentionality' is. Rather, our notion of mental representation is influenced by our pre-theoretical understanding of non-mental representations. This derives from many sources. Portraits, statues and other works of art often aim to represent by resembling their represented objects. Sheet music represents the symphony through a rule-based relationship between the symbols on the page and the notes produced by the interpreting musician. I have considered the theory-specific constraints on mental representation — theory form and desiderate — already. The third — pre-theoretical intuitions — comes from our experience of everyday representations like portraits and sheet music. This section will draw out two important such intuitions, and subsequent sections will show how theories of representation rely on them for better or for worse. My main contention is that, given their status as intuitions, they are not sacrosanct and we have the choice to reject them in the face of theories which are inconsistent with them yet possess explanatory power.

Many analyses of mental representation begin by looking at everyday representations. Fred Dretske compares representation to 'natural signs' such as smoke indicating fire (1986). Fodor's LOT (1975) models mental representation on natural language. Each everyday representation we are familiar with is non-mental and, as I suggested in the last chapter, they often require interpreters with minds to make them meaningful.

Meaning is always, in a sense, relative to some agent which a representation has meaning *for*. This gives us one important pre-theoretical intuition:

> (*PT1*): (mental) representations are intuitively representations for some interpreter.

Many philosophers subscribe to this view. For example, Ruth Millikan writes: "The notion of a sign makes intrinsic reference to a possible interpreter… There are no signs without potential interpreters." (Millikan, 1984: 118) Others who endorse this include C. S. Peirce (1931) and the early Robert Cummins (1983). Deniers of (*PT1*) include Daniel Dennett (1978) and the later Robert Cummins (1996).

(*PT1*) is best captured by conceptualising representing as a 3-place relation involving representation, represented object, and interpreter. But this opens up such an account to the *Homunculus Regress*. Strategies to overcome this centre around making the internal interpreter of the mental representation sufficiently 'dumb' (non-intentional) so that it does not have all the properties of a full-blown mind. Attempts at following this idea include homuncular functionalism (Dennett, 1978; Lycan, 1981),[6] Millikan's Teleosemantics (1984; 1989), and Ramsey's 'mindless strategy' (of which more will be said later).

Ramsey's discussion of the JDC shows that representations must have determinate content and be used in virtue of that content to play their causal-explanatory roles. In the next section I will explain how accounts of mental representation which accept (*PT1*) — by conceptualising representing as a 3-place relation — face severe problems in balancing sufficient content determinacy with avoiding the *Homunculus Regress*.

Another pre-theoretical intuition which dominates accounts of mental representations uses the same basic idea of Brentano's Thesis that all and only minds have intentionality. The predominant feeling among philosophers is that there is something special about minds which entails a sharp separation between things with minds and things without minds. It is supposed that any adequate account of intentionality must capture how it is that minds are different in some privileged way:

> (*PT2*): minds have some property over and above non-minds which allows them to be about other things in a way more fundamental than artefacts like maps are about things.

[6] For criticism see (Cummins, 1983: 92).

Note that (*PT2*) is not simply a restatement of the concept of intentionality. It is a logical possibility that minds are on one end of a sliding scale of intentionality, and that they simply have more of whatever intentionality is than something minimally intentional like a lower animal's mental state. Nevertheless, philosophers have long advanced the intuition in (*PT2*) that minds do have some special property, and that any adequate theory of mind must capture how they are importantly different from non-minds. Philosophers who endorse (*PT2*) include Brentano, John Haugeland (1981), Alex Morgan (2015), and Jerry Fodor (1987). Philosophers who deny (*PT2*) include Dennett,[7] Dretske (1981),[8] Millikan (1984),[9] and, I suggest, also Alex Morgan (2015).

(*PT2*) implies that non-mental things cannot be intentional in a strong sense (Jacob, 2014). This corresponds more or less with the distinction between original and derived intentionality. This seems to be in tension with the very essence of the naturalist, physicalist approach to navigating the *Chisholm-Quine Dilemma*. As Crane (1998b) puts it: "Some philosophers want to locate the basis of intentionality among certain non-mental causal patterns in nature." 'Locating' intentionality in the physical world is a useful metaphor for the representational naturalist project. But a moment's reflection suggests this is hard to reconcile with (*PT2*). On Brentano's view, intentionality is a property supposed to somehow be the preserve of minds alone. Locating the basis of intentionality in nature is, by definition, grounding intentionality in things that need not be mental. For example, Dretske (1981: vii) invokes natural information and causation in his approach to grounding intentionality.[10] This means that there is nothing stopping non-minds from participating in relations which are supposed to yield intentionality. This is in tension with the spirit of (*PT2*).

But it is hard to see how naturalising intentionality could proceed any other way. To get rid of supernatural-seeming unanalysed mental concepts or terms in an explanation the only option is to describe them in physical terms (unless you eliminate them altogether). But this entails opening up the possibility that non-mental things are intentional. Representational naturalism seems to erase the strict dividing line between mental and physical by putting mental properties on a continuum of things with the physical. This is a very important point, so for ease of reference I will dub it the *Naturalistic Intentionality Thesis*:

---

[7] See Fodor's reconstruction of Dennett's criticisms in his (1987).

[8] Arguably: "… *all* information-processing systems [including, for Dretske, thermostats and other obviously non-mental systems] occupy intentional states of a certain low order." (Dretske, 1981: 172)

[9] Also arguably. Millikan places thought on a spectrum of all intentional icons, including words and natural signs in her (1984).

[10] He claims no interpreter is needed. Information is "independent of its actual or potential use by some interpreter" (1981: vii).

> *Naturalistic Intentionality Thesis*: representational naturalism cannot take identity between physical and intentional states as a primitive, or make use of unanalysed mental or intentional concepts in explaining intentionality. Consequently it surrenders the power to account for the privileged difference between minds and non-minds by placing intentionality on a continuum with the physical.

This point comes back to the common ground between *Physicalism* and Property Dualism. It may be that certain physical things with certain compositions just do instantiate the property of being intentional, and this is supervenient on their physical constitution. But in the absence of an explanation why this arrangement of physical things does have that mental property there is still room for the Property Dualist to argue such properties are simply emergent from the physical, and in fact irreducible and fundamental.

Some representational naturalists wishing to locate intentionality in the world take it to be a 2-place relation between representation and represented object. This avoids the *Homunculus Regress* but at the cost of requiring some objective, naturalistic notion of meaning. Moreover, it clearly denies (*PT1*) — though this is a price many deem worth paying. It is at best unclear if the 2-place relation view can maintain (*PT2*), though it seems that taking representation to be a 2-place relation requires locating meaning or intentionality as an objective quantity somewhere in the physical world. Therefore it seems strongly committed to the *Naturalistic Intentionality Thesis*.

I contend that many philosophers seek to incorporate one or both of (*PT1*) and (*PT2*) in their accounts of mental representation. Many accounts are criticised for not adhering to one or other. Failure to properly mark the difference between minds and non-minds captured in (*PT2*) is a frequent form of criticism, as I shall detail. I think both intuitions are in tension, if not outright contradiction, with representational naturalism. Moreover, I think the prevalence of both intuitions is at the root of the *Chisholm-Quine Dilemma*.

In the coming sections I will consider Fred Dretske's and Ruth Millikan's attempts to negotiate the *Dilemma* and argue that both suffer problems of content indeterminacy and lack of *Explanatory Purchase*, and so ultimately fail the JDC. Both approaches weaken intentionality so as to break radically with (*PT2*), while Dretske's approach denies (*PT1*) altogether. Millikan's approach pays lip service to (*PT1*) but fails to either solve the content indeterminacy problem or pass the JDC without supposing the interpreter is intentional, thereby succumbing to the *Homunculus Regress*. Thus, my discussion will criticise Dretske and Millikan along these lines.

However, alongside this criticism I will also defend their approaches against a common, but unwarranted, criticism. It seems to me that any naturalistic account of mental representation must embrace

the *Naturalistic Intentionality Thesis*. I will argue that a corollary of this is that any such theory cannot be criticised for failing to meet (*PT2*). Explaining intentionality and representation naturalistically must involve locating intentionality — and hence minds — on a continuum with non-minds. To pursue a naturalistic approach requires accepting that a privileged division between mental and non-mental might not be recovered. To presume otherwise is a confusion I will criticise some philosophers for succumbing to. Thus some attacks on Millikan, Dretske and other naturalistic accounts are ultimately incoherent.

3.2 Dretske's Indicator Semantics

Dretske develops his naturalistic theory of mental content across books and articles published between 1981 and 1988, with modifications on a general theme based on two notions: the transmission of information, and indicator function. The essence of Dretske's proposal is that some brain state, *A*, represents some object or state of affairs in the external environment, *B*, in virtue of the relation that obtains between *A* and *B*. This relation depends on *A* having a function to represent, or indicate, the presence (or state) of *B*. *A* then is tokened reliably given the presence of *B* — it is reliably caused by *B* — and, by virtue of this causal relation, *A* comes to represent or indicate *B*.

The variations across the different iterations of Dretske's account concern how to account for *A*'s function being to indicate or represent *B*. In 1981 the function of *A* is set during a 'learning period' for the organism. Exposure to instances of *B* is, loosely, supposed to result in the training of *A* to respond differentially to *B*, thereby settling *A*'s function. In 1986 an appeal to teleology replaces the 'learning period': Dretske supposes that *A* comes to have the function of indicating *B* through a process of natural selection: the function of indicating *B* evolved to aid the organism's survival. The notion of *A* responding to the information relayed by the presence of *B* is a common thread in Dretske's accounts.

The above is a brief summary of Dretske's key ideas. Important to recognise is how objective natural information permits the conceptualisation of representation as a 2-place relation without need of an interpreter. *A* acquires the function of indicating the presence of *B* because *B* carries natural information — for example, 'smoke means fire' — which amounts to a lawful causal covariance (Godfrey-Smith, 1989: 543). Fixing *A*'s indicating function by either teleology or learning results in a lawful covariance of *A* with *B*.

Dretske tries to avoid the *Homunculus Regress* through use of the 2-place relation. This locates intentionality in the world via the notion of natural information, thereby eroding the division between mental and non-mental in line with the *Naturalistic Intentionality Thesis*. Consequently, Dretske's approach denies (*PT2*) and opens a gap between 'strong' and 'weak' intentionality. It is not obvious how 'strong' intentionality can be recovered (Ramsey, 2007: 123-124). This consequence is highlighted in Dretske's 1981

account, where he explicitly distinguishes a continuum of intentionality all the way from the "first order of intentionality" which is common to "all information-processing systems" including thermostats and voltmeters, to the "third order of intentionality" exhibited by states like belief and knowledge (1981: 172-173).

Is the explanatory payoff worth it? I think it is not; Indicator Semantics struggles with problems of content indeterminacy and doesn't pass the JDC. To see this, consider an expanded picture of how the account works. Dretske proposes that signals carry information about the state of something they lawfully indicate. It is this lawful indication which is fixed by teleology or learning. For example, it is lawful that a (correctly functioning) petrol gauge only points in the red when the petrol tank is empty. The signal *s is F* = 'the needle is in the red' indicates that *t is G* = 'the tank is empty'. However, as Godfrey-Smith points out, there is no reason to suppose the meaning is as determinate as this (1989: 545). For example, if *s is F* this also lawfully indicates that *t is H* = 'the tank is full of air'.

The problem stems from the indeterminacy of natural information. A signal — say, the light incident on my retinas from a red tulip — carries far more information than is plausibly encoded in a belief state. The tulip is red, and the signal carries that information, but also much more: the size, shape, orientation and many other properties of the tulip are part of the information carried in the signal. In contrast, intentional states like belief have much more determinate contents. Thus, a convincing version of Indicator Semantics should be able to explain how unconstrained information signals result in the determinate content of intentional states without becoming circular by invoking intentional concepts.

To overcome this, Dretske's 1981 account makes a seemingly circular appeals to 'background beliefs' (1981: 43; 65). The information of a signal is constrained by the beliefs of the 'receiver' of the signal.[11] For example, a radio football commentator's statement (an example of a 'signal' in Dretske's sense) "Someone has just scored in the 3pm game" carries the information that 'either an Ajax or a Feyenoord player has scored', contingent on the hearer knowing that the 3pm game is Ajax versus Feyenoord. 'Background belief' is an intentional term and so leads to circularity in specifying intentional content. Further, as beliefs are intentional and personal-level entities, Dretske's requirement that background beliefs figure in determining intentional content essentially involves smuggling in an interpreter after all by tacitly taking representation to be a 3-place relation. Therefore Indicator Semantics both fails to yield determinate content and succumbs to the *Homunculus Regress* after all by invoking intentional terms to try and

[11] Dretske's original formulation is in terms of information theory and is phrased in terms of the constraint of probabilities given background beliefs. For another statement of my worry, see (von Eckardt, 2012: 36-37).

individuate content. This means that, without invoking 'background beliefs', Indicator Semantics remains naturalistic but fails to yield determinate content. However, to yield sufficiently determinate content it invokes 'background beliefs' and thereby forfeits its naturalistic credentials.

The problem can be stated in more formal terms. Intentional content is characteristically fine-grained so that co-extensive properties of the same particular thing are readily available as the content of intentional states. One can have a belief about the colour or the shape of the tulip alone, for example. Indicator Semantics fails to recover this fine-grainedness because only by invoking intentional concepts can indicator function seem to distinguish between co-extensive properties. The same signal carries the information that, say, 'x is F and x is G ', whereas intentional states are ascribed more determinate content, like simply 'x is F'. To pick 'x is F' from the signal requires 'background beliefs' and thereby causes a *Homunculus Regress*.

So it is clear Indicator Semantics has a problem with content indeterminacy which puts pressure on its naturalistic credentials. Another related problem is that it fails the JDC. Ramsey devotes a full chapter to arguing this, and I will only briefly restate his main argument here. He claims Indicator Semantics is an example of what he calls a 'receptor' representation: a state that reliably responds to an obtaining state of affairs through lawful covariance. He writes "… a structure can be employed *qua* nomic dependent or *qua* reliable-respondent without being employed *qua* information-carrier or, more to the point, *qua* representation." (2007: 138) The main point is that Dretske presumes the information carried in the relation that allows the internal state to covary with the external state of affairs to be essential to that covarying, whereas Ramsey argues the informational content is irrelevant for the nomic dependency. For example, the state of a thermostat's bimetallic strip lawfully covaries with the ambient temperature and thereby causes the thermostat to switch on or off, but the bimetallic strip is not representational as it simply reliably influences the temperature setting of the thermostat. There is no reason to suppose this reliable influence proceeds via representation (2007: 136). This is to say Indicator Semantics gains no *Explanatory Purchase* by invoking representations.

In summary, Dretske's use of the 2-place relation denies both (*PT1*) and (*PT2*). It also fails to recover determinate content without circularly invoking intentional terms, and it fails the JDC. I have focused on the 1981 discussion, but as the later amendments replace the 'learning period' with teleology, leaving the indicator function and natural information notions intact, I think my criticism applies equally to the latter accounts. In the next section I will consider Ruth Millikan's Teleosemantics and assess what stance it takes towards (PT1) and (PT2), and whether it passes the JDC and the determinacy of content requirement.

3.3 Millikan's Teleosemantics

Ruth Millikan develops a complex naturalistic picture of intentionality, representation and mental content in her 1984 book *Language, Thought, and other Biological Categories*, and further in her 1989 article 'Biosemantics'. Though there are many striking differences between Dretske's view and Millikan's approach, in this section I will suggest that there are deep similarities. Moreover, both accounts fail to pass Ramsey's JDC as they gain no *Explanatory Purchase* from invoking representations: they fail to show that representations cause behaviour in virtue of their contents.

A general presentation of Millikan's account can be given with a limited number of essential concepts. Firstly, as the name suggests, mental content is ultimately individuated according to a mental state's teleology: states mean what they mean because it is essential for fulfilling their purpose, or function, to mean that. Secondly, this purpose is selected for during evolutionary history. Biological mechanisms which use intentional content have particular functions because — under conditions which are 'historically Normal' (a technical term which captures the historical environmental conditions which allowed the performance of the function) — they were naturally selected for. Together, these notions combine to yield the 'teleofunction' of a given biological mechanism: its function to perform some job which has been selected for through the course of evolutionary history, relative to the obtaining of historically Normal background conditions. "To have a teleofunction is to have emerged from a certain sort of history, one involving some form of selection," she writes (1990: 152). For example, a frog's 'fly-detecting' visual-motor mechanism evolved in the Normal condition that airborne dark spots in the visual field were reliably correlated with flies, hence the frog's tongue-snapping behaviour reliably led to feeding the frog.

Thirdly, Millikan's account depends on the distinction between a 'producer' and 'consumer' (or 'interpreter') of a representation. This elastic notion supposes that intentional systems decompose into two parts which evolved to work together. The producer's function is to produce a representation for the consumer to consume, or interpret, in aid of performing its function (1990: 157). This is highly abstract, and Millikan claims that while producer and consumer can be two internal components of the same system — as would seem to have to be the case with mental representations — they can also be entirely separate entities. A favourite example of Millikan's is the bee dance, in which one bee's 'dance' represents the location of nectar to another bee. Here the dancing bee is the producer, the dance is the representation, and the watching bee is the consumer.

This idiosyncrasy occurs because Millikan's theory, especially as presented in 1984, is intended to be highly general. One of her main contentions is that mental representations are simply one subcategory of 'intentional icons' (1984: 12). These include sentences, pictures, bee dances and more or less any other

meaningful token. This approach clearly exemplifies the naturalistic strategy of dissolving the barrier between mental and non-mental by locating intentionality in the world, which is precisely the denial of (*PT2*).[12] However, interestingly Millikan seeks to preserve — albeit in a highly naturalistic way — (*PT1*). The consumer/producer distinction allows Millikan to model representation as a 3-place relation (1984: 85, 95; 2009: 396). This naturally introduces the *Homunculus Regress* threat once more, but I will pass over this for now.

What makes an intentional icon a mental representation? Millikan explains: "… what distinguishes representations from more primitive intentional icons is that the mapping values of the elements of representations are supposed to be identified." (1984: 239) 'Identify' means: "roughly, understand the reference of" (1984: 96). Though this notion of identification is delivered in intentional language ('understand'), suggesting circularity, Millikan ultimately attempts a naturalistic account of identification which depends on the correspondence between world and icon for the correct performance of selected-for functions. She talks about the identity between two different representations of the same real-world state of affairs: an "act of correct identification" is performed by some interpreting device when performance of its function requires the device use the icons jointly to successfully perform the function. She writes "the interpreting device will be able to accomplish what good it does Normally only *because* these elements map the same" (1984: 242).

I am now in a position to state how content is individuated under Teleosemantics. Intentional icons are reliably correlated with how the world is through teleofunctions. Teleofunctions have been selected for because they have been historically successful in allowing their consumers to function effectively for the proliferation of the organisms that use them exactly because the correlation of the icon with the state of the world obtains reliably. Given this obtaining correlation — which Millikan often refers to as 'semantic rules' — the content of the icon is just what the icon must mean, and must always have meant and have been selected to mean, so that the consumer function properly.

> […] the producer produces a sign that will be true or satisfied only if it maps onto some affair or affairs… in the world in accordance with certain 'semantic' rules… of correspondence between signs and world affairs that have been instantiated in the past when the consumer and producer or their ancestors have succeeded in performing their cooperative function(s). (2009: 397)

Semantic rules hold because they had to have held for evolutionary success. Intentional icons "the referents of which are supposed to be identified" are representations. Mental representations are thus inner intentional icons which are 'supposed' to have their referents identified in the course of the Normal explanation. For

[12] Millikan overtly denies the existence of original intentionality — or 'basic' intentionality in her terms (1984: 89-90).

example, when tying a knot the action's success depends upon the agent's identification of the visual and tactile representations of the end of the rope as the same actual object (Millikan, 1984: 240).

Teleosemantics suffers from an indeterminacy of content just like Indicator Semantics. One problem afflicting its appeals to teleology is that it is difficult to specify biological function in a fine-grained enough way to get the sort of content imputed in intentional explanations. Consider the example of the frog snapping its tongue at dark spots. There are numerous contents which may correspond with the dark spots, such as 'fly' or 'food'. Yet sticking strictly to teleology would seem to yield a content perhaps as general as 'metabolism-satisfying object', and significantly less specific than is required for most intentional states. This is another instance of representational naturalism's difficulty in dealing with coextensive properties. As Richard Hall (1990: 195) puts it: "Darwin cares how many predators you avoid but not what description you avoid them under". Moreover, appeal to the function of the consumer is not as straightforward as it appears. Karen Neander worries the system that benefits from the visual representation of the dark spots might be some mental capacity in the frog, some aspect of its motor control system, the digestive system that digests the food, or even the circulatory system that distributes the nutrients (Neander, 2018). An inability to accurately specify the consumer entails an inability to specify content determinately.

Indeterminacy of content is one issue, but a bigger problem is that content ascription actually seems irrelevant to behavioural explanation. The frog cannot discriminate flies from dark spots to cause its tongue-snapping, but it does not need to. So long as there are dark spots, the frog reacts. Thus, claiming the dark spots are intentional icons at all seems to add nothing to the explanation, contravening the JDC. However, as I mentioned, Millikan distinguishes between intentional icons and full-blown representations through the identity condition. Thus to show that her proposal *qua* mental representations fails the JDC I must show that her appeal to identification cannot reconstruct 'strong' intentionality from the 'weak' variety intentional icons possess.

In Chapter 2 I mentioned that the aspectual shape of an intentional content is something that should be explained by a theory of content. This is captured by the fine-grainedness of intentional content extending to coextensive properties of the same object. I think that Teleosemantics is unable to explain the aspectual shape of intentional content even with Millikan's appeal to identification. Consider the knot-tying example. Identifying the visual and tactile representations of the rope as having the same referent is certainly necessary for successfully tying the knot. But to fully explain intentional content then the naturalistic description would still — much like the frog example — have to convince us that the agent takes the referent of the representation to be precisely the knot in question, not some other kind of object and certainly not

nothing in particular. An agent could feasibly succeed in doing the tying so long as its systems are satisfied it can feel and see its hands on the same object. The actual recognition of the object's identity need never enter into the explanation of the action beyond this minimal requirement and yet still yield successful knot tying. Yet our experience is of taking the object at hand to be one particular thing and not another.

To satisfy *JDC Pass*, describing the knot-tying in representational terms must be explanatorily beneficial and the system must use the representations in virtue of their content in causing behaviour. Even supposing the teleofunction (whatever that is) used in the knot tying is fine-grained enough to yield the right content, if there is no plausible need for the agent to take the representations to represent the knot rather than something else, then the representational description is explanatorily otiose. Thus Teleosemantics fails the JDC and is only pseudo-representational.

I mentioned that Millikan's use of a 3-place relation threatens a *Homunculus Regress*. She contends that so long as the interpreter need not token a further intentional icon when it performs its 'interpretative' (in, I presume, a loose, non-intentional sense) role — if representations simply cause non-intentional processes in the consumer — no regress begins (1984: 90; 2009: 396). For example, if a bee interpreting the dance of another responds by simply flying to the location represented, regress is avoided. But this suggests the question: why suppose the 'representation' was actually representational? Described like this, Millikan's position seems vulnerable to the same criticism Ramsey aimed at Dretske: the so-called 'representation' functions just like a 'receptor' in Ramsey's sense. Yet Millikan is clear that "the part of the system which consumes representations must understand the representations proffered to it. […] [T]here must be something about the consumer that *constitutes* its taking the signs to indicate, say, *p*, *q*, and *r* rather than *s*, *t*, and *u*." (1989: 286)

Thus we observe a tension in Millikan's thought. On one hand she admits that a representation consumer should take a representation to have a particular content. On the other hand, I argued that Teleosemantics is incapable of individuating such determinate content and that this content is explanatorily otiose. The overall goal of this essay is to suggest that, because of the *Naturalistic Intentionality Thesis*, it is simply incoherent to expect representational naturalism to simultaneously meet both conditions. Sticking to representational naturalism entails that, if representing is conceived of as a 3-place relation, then content can never be sufficiently determinate. The interpreter cannot be intentional, on pain of regress, yet without an intentional interpreter content seems doomed to be indeterminate. Naturalising the interpreter leads to the situation reminiscent of Searle's Chinese Room: a representation *seems* to be being used as if it has a particular content, but it is not used *in virtue of* that content.

To summarise, in contrast to Indicator Semantics, Teleosemantics seeks to preserve (*PT1*) in a naturalistic framework. However, both have in common the naturalistic strategy of denying (*PT2*) and both thereby necessarily adhere to the *Naturalistic Intentionality Thesis*. Neither Dretske nor Millikan see these concessions as problematic for their overall goals. However, I have argued that neither can succeed in yielding content sufficiently determinate for cognitivist explanation, or succeed in convincing that representations are used in virtue of that content.

3.4 The Distinctively Mental Criticism

The previous two sections detailed how neither conceiving of representation as a 2-place relation nor a 3-place relation allows us to adequately individuate representational content, or convince that representations are used in virtue of that content, to pass the JDC. Further, both Dretske's and Millikan's naturalistic approaches deny (*PT2*). In this section I will suggest this denial gives rise to a common criticism that naturalistic approaches label too many intuitively non-mental things as mental. I will use the exposition of the problems facing Dretske's and Millikan's approaches to show that such criticism is largely ill motivated. From the *Naturalistic Intentionality Thesis* it follows that some seemingly non-mental things are at risk of being admitted as mental. Any criticism along these lines is incoherent if the critic also endorses the core of a naturalistic project. Explaining why this is allows me to derive my central claim: either we revise some of our intuitions or we must drop representational naturalism.

If a theory of representation denies (*PT2*) then the privileged dividing line between mental and non-mental is dissolved. Thus critics can claim that a theory "casts the net of representationhood too wide, over areas on which it has no explanatory purchase", in the words of one such critic: Alex Morgan (2015). Morgan cites criticism of Millikan that interactions between trees are rendered contentful, and that saliva represents food as examples where our intuitions are contradicted (2015: 219). Similarly, Godfrey-Smith reconstructs an argument against Millikan from Paul Pietroski that there are counterintuitive cases where the biologically-determined content of a creature's representation is something it would in fact be unable to perceive, and thus unable to represent (2006: 63). Ramsey criticises Indicator Semantics in the same vein, arguing that indicators are pseudo-representational 'receptors' which reliably respond to environmental factors and hence cause behaviour, but not in virtue of representational content, thereby failing the JDC (2007: 118-141). Finally, in a critical article, Morgan argues that Ramsey's favoured 'S-representation' is not sufficient for defining representationhood because non-mental creatures use representations in this manner. Morgan's example is of circadian clocks in plants (2015: 238). Such cases emphasise results which break

from our intuitions about which things are mental representations, and hence which activities we count as mental.

I think these criticisms are underwritten by the critics' commitment to this claim: a necessary condition on a theory of mental representation is that it should allow us to distinguish the mental from the non-mental. Critics seem to take the mental/non-mental distinction to be an empirical fact which any good theory of cognition should account for. However, given the role of pre-theoretical intuitions in constraining theories of representation, I suggest this expectation reflects not an incorrigible fact but a manifestation of (*PT2*). Let's call this the *Distinctively Mental Criticism*:

> *Distinctively Mental Criticism*: No theory of mental representation that labels non-mental things as mental, or labels non-representational things as representational, or labels non-intentional things as intentional, is correct.

From the above examples of criticism it seems that many agree that if any theory of mental representation is vulnerable to the *Distinctively Mental Criticism* it must be flawed. However, I will show that arguments using it against theories of mental representation are in contradiction if they also affirm *Physicalism*.

Both Dretske's and Millikan's naturalistic approaches deny (*PT2*). I argued this denial is a necessary part of representational naturalism, which locates intentionality in the world. This is the *Naturalistic Intentionality Thesis*. Further, I claimed this leads to a seemingly unbridgeable gulf between 'strong' and 'weak' intentionality. For representational naturalism to succeed, intentionality must be described in physical terms. Therefore it is a strange criticism to claim that the fact such theories attribute intentionality to the non-mental counts against the theory because of the alleged fact — really the intuition (*PT2*) — that non-mental things aren't intentional in a strong way. If representational naturalism involves attenuating intentionality so that it is found in things we don't consider mental, we cannot simultaneously criticise theories that attempt this strategy for failing to return as outputs all and only those things which we want to characterise as mental.

I argue this is because the *Distinctively Mental Criticism* relies on the assumption that we can pick apart the mental from the non-mental. But (original) intentionality is often used as the means of distinguishing mental things from non-mental things. Thus, if a theory of mental representation is viewed as a way of explaining intentionality, then appealing to a supposed distinction between mental and non-mental as a means to tell if the theory is successful or not is circular.

Coming up with an infallible 'mark of the mental' is difficult (Kim, 1996: 15-16). Intentionality is commonly taken to mark the difference between mental and non-mental, as Brentano's Thesis shows. However this remains problematic as the need to distinguish original and derived intentionality demonstrates. The mark of the mental then becomes *original* intentionality. Is this not somewhat *ad hoc*? The aim was to find some criterion to distinguish between the mental and non-mental. To meet objections that intentionality does not apply to all and only mental phenomena, original intentionality is posited. But how do we decide which things have original intentionality? It is built into the definition of original intentionality that only minds have it, but we lack independent purchase on which things have minds as our original question was to find a way to distinguish minds from non-minds. Therefore claiming all and only minds have original intentionality is of no help. Thus, in the absence of an independent mark of the mental, there are no grounds for the *Distinctively Mental Criticism* because only a theory of representation could allow us to distinguish minds from non-minds — but this distinguishing ability is *presupposed* in making the *Criticism* in the first place.

Consequently, barring appeal to an intuition like (*PT2*), any critic of a theory of mental representation cannot simultaneously complain the theory is open to the *Distinctively Mental Criticism* and also support a naturalistic theory adhering to *Physicalism*. (*PT2*) and *Physicalism* are incompatible (because *Physicalism* just is locating intentionality in the world) and the *Distinctively Mental Criticism* relies on (*PT2*). Adopting representational naturalism just dissolves the mental/non-mental barrier, which is a denial of (*PT2*). The *Distinctively Mental Criticism* presupposes (*PT2*), so any argument using it is inconsistent unless it denies *Physicalism*.

Nevertheless, many instances of the *Distinctively Mental Criticism* — including Morgan's version against Ramsey — are made by philosophers who also adhere to representational naturalism. If my analysis is correct there seem to be two options: either (i) we abandon the allure of pre-theoretical intuition and allow the explanatory value of theories of cognition in a physicalist ontology to guide us, or (ii) we put faith in the strength of the intuition and accept that the *Chisholm-Quine Dilemma* was correct all along: intentionality and P*hysicalism* are incompatible. At best *Physicalism* and Property Dualism are equally plausible.

I have claimed that any argument against a theory of representation using the *Distinctively Mental Criticism* is only valid when conjoined with a denial of *Physicalism*. Representational naturalism affirms *Physicalism*, and the work of Dretske and Millikan suggests an acceptance of (i). Any philosopher holding on to (*PT2*) while pursuing representational naturalism fails to see their position is inconsistent. Morgan's criticism of Ramsey's S-representation is an example of this confusion. In this next chapter I analyse their

dispute and suggest the motivation for Morgan's criticism reflects an underlying deficiency with representational naturalism, which is a reason to choose (ii) over (i).

Chapter 4

4.1 Morgan's Criticism of S-representation

To pass the JDC a theory must show that representations are used in virtue of some explanatorily relevant content. Ramsey claims to identify one notion of representation which passes the JDC: S-representation. S-representations are structurally isomorphic with their represented objects and work in tandem with the mental model conception (2007: 193). A mental model is able to model some domain because it is composed of elements which, together, are structurally isomorphic with that domain. In a map of a park the individual resembling elements — the lines representing roads, rivers etc. — enable the map as a whole to function as a model by being elements in a representational structure isomorphic with the real park. S-representation takes this idea and supposes it works equivalently inside one's head. Using an internal map an agent plans their route by using the representation. This is the 'surrogate' reasoning mentioned in Chapter 1. For Ramsey, "components of the model *become* representations when the isomorphism is exploited in the execution of surrogative problem-solving" (2007: 96).

To avoid the *Homunculus Regress*, Ramsey employs what he calls the 'mindless strategy' to show that S-representations are used by systems which need not have full-blown minds: "we can have something that functions as a representation in a physical system, even if there is no sophisticated built-in learner or inference-maker that it serves as a representation *for*… [S]uch a consumer is little more than a mechanical process or device that the representation effects" (2007: 192-193). Ramsey gives an example of a driverless car navigating a track using an internal S-representation. First, imagine a car with blacked out windows and a driver inside who possesses a map of the track. The map is isomorphic with the track so functions like a model of the track to guide the driver's behaviour navigating the track. The map in this case is the S-representation.[13]

To employ the mindless strategy, suppose we replace driver and map with a groove inside the car which is isomorphic with the shape of the track. Some mechanical system could be set up with a rudder tracing the groove, the deflections of which cause appropriate manipulations of the steering wheel. Ramsey thinks this amounts to a mindless system using an S-representation. The map is essentially transformed into the groove, and the driver's interpretative role is replaced by a mechanical system which 'reads' the groove and thereby steers the car. Ramsey claims "the car is exploiting the isomorphism between the groove and the track in much the same way that the driver did [with the map], even though the process is now fully

[13] I think Ramsey's exposition misses out the importance of the agent's being able to constantly check their progress with relation to their surroundings during navigation through their perceptual faculties. For the sake of argument I will ignore this complication.

automated and mindless" (2007: 199). The automatic system still exploits an isomorphism between groove and track, and this is sufficient for a representational explanation that meets *JDC Pass* while also avoiding the *Homunculus Regress*.

Morgan argues S-representation is necessary, but not sufficient, for *mental* representation. His argument is a *Distinctively Mental Criticism*: there exist counterexamples which meet the criteria to be considered S-representations but which we would not want to call representational. "While [S-]representations might count as genuine representations, they aren't distinctively mental representations, for they can be found in all sorts of non-intentional systems such as plants," Morgan writes (2015: 213).

He cites circadian clocks in plants, which he claims are functionally isomorphic with external "daylight cues" which let them model the Earth's rotation period in line with the day-night cycle (2015: 233). The clock interfaces with a plant's 'motor control systems' to 'influence behaviour'. They apparently use these clocks 'offline' — something we associate with mental representations — to orient their leaves overnight (in the absence of the daylight cues) to maximise exposure to morning sunlight. The clocks are thus functioning isomorphs and should be considered representations on Ramsey's account. But Morgan argues they "surely *don't count* as *mental* representations, i.e. the vehicles of cognitive processes like episodic memory, planning, and mental imagery" (2015: 240). Here it is plain that Morgan expects an account of mental representation to mark the difference between mental and non-mental. Ramsey's S-representation fails to do so and should be rejected: precisely the structure of the *Distinctively Mental Criticism*.

In the previous chapter I argued that if one is a representational naturalist — necessarily affirming the *Naturalistic Intentionality Thesis* and rejecting (*PT2*) — then one has no grounds to make a *Distinctively Mental Criticism*. Morgan is a representational naturalist so his argument against Ramsey fails. This means that Morgan's issue with Ramsey can be traced back to their stance on (PT2): Morgan affirms it while Ramsey denies it. Thus their disagreement reflects not a substantive problem with Ramsey's theory, or even anything about the merits of Morgan's criticism, but simply is emblematic of an irreconcilable difference in their starting assumptions.

Ramsey's view is that a suitable explication of how representations are used by mindless systems is sufficient for explaining mental representations. In contrast, Morgan thinks this explication is insufficient if there is no 'distinctively mental' element in such an account. But Ramsey simply denies there *is* anything more to being a mental representation than being an S-representation in the head. Morgan references private correspondence from Ramsey which appears to confirm this: "Ramsey has agreed that circadian clocks in

plants might count as [S-]representations; he emphasizes that what's essential is whether a system is used as an internal surrogate, not whether it's distinctively mental." (2015: 238) Giving up claim to the 'distinctively mental' is, I think, just the price one pays for affirming *Physicalism*. Morgan's intuitive view of the mind is simply incompatible with his preferred ontology.

I believe this discussion brings out a conflict of starting assumptions which is generally mischaracterised as a difference in explanatory value of equally plausible theories. A *Distinctively Mental Criticism* is only valid when raised by a denier of *Physicalism*. Therefore, such criticisms raised at naturalists by naturalists are really manifestations of philosophers' differences in intuition about the nature of the mental. In other words, to make such a criticism amounts to criticising the way the nature of the mental is characterised. But the way the mental is characterised is built into one's theoretical paradigm, thus one cannot make changes to that characterisation without rejecting or radically reworking the basic claims of the paradigm itself. As mentioned in the last chapter, one is faced with a hard choice between revising intuitions to fit within a physicalist worldview, or accepting that *Physicalism* may not be a suitable ontology for capturing 'the mental' — however one construes this. In the next section I will consider why this might be.

4.2 What Representational Naturalism Lacks

As I see it, the force of criticisms like Morgan's against Ramsey come from the fact that representational naturalism misses out on an important, but crucial, element of representation. This is something like the representation user's recognition of the representation as representing something in particular.

To support this speculation, I return to some of the material discussed earlier. Firstly, according to Crane, intentionality has an aspectual shape. A thought is about something and is presented in one's mind in a particular way. Multiple possible ways of presentation are all still ultimately directed on the same object of thought, and can involve different coextensive properties of the object. Thus any intentional content-determining account should have a way of picking between coextensive properties of an object. Secondly, as Millikan notes, identification of the referent of a representation is crucial: what seems special about mental representation in us is that we are able to identify what is being represented. I argued that Teleosemantics lacks the resources to allow the identification of intentional content under such aspectual shapes, as seems necessary for a satisfying account of intentionality. Thirdly, Dretske also identifies aspectual shape as something to be accounted for in our experience of representation (1995: 30-32). Fourthly, Dretske's early attempts at naturalising representation using a 2-place relation ran into difficulties when picking between coextensive properties. I showed that he even had to tacitly invoke intentional terms like 'belief' to make up this shortfall in his 1981 account. All of this does not prove that some sense of first-person recognition or

awareness of what a representation represents is necessary. However, I do take it to indicate that this is at least a strongly felt intuition.

Even Ramsey's account suffers from the same problem. On the surface, Ramsey claims representation is a 3-place relation with an interpreter. We saw in Chapter 3 that Millikan adopted the same strategy but could not see it through because her notion of mental representation failed to pass the JDC. There was seemingly nothing about the explanation which required it to be representational by being used in virtue of a particular content. I think Ramsey succumbs to the same problem with his 'mindless strategy'. He claims that an S-representation has a particular content and is used in virtue of that content because the reasoning done on the mental model could only successfully cause behaviour if the representation did in fact represent the thing in the world reasoned about. Thus, use of the representation ensures the isomorphism exploited is the correct one. And this only happens because of the state of the world in which the subject finds itself. However, as Morgan notes, this merely pushes the question of how the representation has *just that* content deeper into the act of directing the mental model on that state of affairs in the first place (2015: 224). To get around this, Ramsey makes an appeal to what the explanandum in a cognitive explanation actually must be in any given situation (2007: 94). For example, a cognitive explanation of the driver's successful navigation only makes sense if the driver actually did model *that* particular track navigated. In the automated car case it only makes sense that isomorphism of the internal groove with the track navigated makes the groove a representation if the groove has been designed to be isomorphic with that particular track. It is possible that the groove is accidentally isomorphic with a different track and would also succeed in navigating that track. Yet without the intentional designing element to 'aim' the isomorphism at that particular state of affairs then successful navigation is only accidentally achieved through the isomorphism. This then seems insufficient to qualify the groove as a representation of one particular track and not another. But, as Morgan notes, this appears to introduce a radical observer-dependence to the content of the representation, which evidently just pushes the *Homunculus Regress* further down the explanatory food chain (2015: 224). In a similar manner to Millikan and Dretske before him, Ramsey fails to secure determinate content or to pass the JDC by having to smuggle in an intentional notion to fix representational content on one state of affairs rather than another.

A way out that would solve the indeterminacy of content problem and allow the JDC to be passed, while also easing 'distinctively mental' worries, would be to introduce some way to account for the representation user's identification of the referent of the representation. Mindless systems can always plausibly use representations without requiring them to have any particular meaning, as *Chinese Room*

suggests. For this reason it appears unlikely that any automated, mindless strategy could pass the JDC. But if we try to avoid this criticism by introducing a condition that only things with minds can use mental representations, we need to appeal to things being minds in the account. Nothing without a mind can identify or recognise a representation as having a particular content in the appropriate sense. This strategy can then only proceed if representation is a 3-place relation with an interpreter. An interpreter functions by *seeing that* or *understanding that* the representation represents this state of affairs. How can we build this idea into an account of mental representation?

The shared failing of naturalistic accounts is that the content ascribed to representations is irrelevant for playing a causal-explanatory role, i.e. naturalistic accounts fail the JDC. My solution to ensure content does play such a role will be incompatible with *Physicalism*.

4.3 'Representation as'

Morgan's idea of the 'distinctively mental' lacking in Ramsey's account is not analysed further. He thinks that S-representation cannot be correct because non-mental plants appear to be using S-representations. He does not propose any condition that would separate the usage of S-representations by things with minds and things without minds. In this section I will propose such a condition — 'representation as'. I think it would avoid the grounds for Morgan's criticism, though I do not think Morgan himself had this in mind or would necessarily endorse it because it is overtly anti-naturalist. However, my argument so far in this chapter has been that criticisms such as Morgan's against Ramsey are strictly incoherent when combined with representational naturalism. Thus, in even raising the question Morgan is at least tacitly dissatisfied with representational naturalism. As I said, either one sticks with naturalism and bites the bullet that one's distinctively mental intuitions must be renounced, or one accepts the intuitions and considers that naturalism may be unable to support them. I adopt the latter option in the following.

'Representation as' attempts to integrate the aspectual, recognitional aspect of mental representation into the picture. It can be formulated, as far as I can see, in intentional terms only and therefore runs immediately into the *Homunculus Regress* unless representational naturalism is abandoned.

Dretske mentions something like what I have in mind in his paper 'Misrepresentation'. He discusses sea-dwelling 'magnetotactic' bacteria which use magnetic 'sensors' to orient their movements using the direction of the Earth's magnetic field. Oxygen-rich water is toxic for them, so they evolved the magnetic sensors to keep them directed towards deeper, oxygen-poor waters. Southern and northern varieties of the bacterium exist. Northern bacteria align with the direction of geomagnetic north, while southern bacteria align against the direction of geomagnetic north to guide them to deep water. You can lure a northern

bacterium to its death by placing a bar magnet such that it orients itself against geomagnetic north and towards oxygen-rich water. Does this mean that the bacterium misrepresents the direction of geomagnetic north, or of oxygen-free water? Dretske (1986) wonders:

> The most that might be claimed is that there is some cognitive slip (the bacterium mistakenly 'infers' from its sensory condition that *that* is the direction of oxygen-free water). This sort of reply, however, begs the question by presupposing that the creature *already* has the conceptual or representational capacity to represent something *as* the direction of oxygen-free water […]. (p 30)

Dretske does not argue that the bacteria are representing in the strong intentional sense required for mental representations in human cognition. He wants to build his account on an instance of 'natural misrepresentation' using the natural information the bacterium uses. But the problem is again one of content indeterminacy: the bacterium's sensors can pick out the direction of geomagnetic north, but what license do we have to say that the sensory state indicating this direction has the content 'oxygen-free water', 'direction of geomagnetic north', 'direction of deep water', or any of various other coextensive properties? Indicator Semantics does not give us the resources to pick any of these over any other. Teleosemantics would at least yield the content 'the direction of oxygen-free water' because this is what the bacterium's ancestors needed it to mean to proliferate. But, as I argued in the last chapter, in neither account is there reason to even require *any* content to explain the bacterium's 'behaviour'. There is no explanatory benefit to supposing the bacterium represents anything at all, rather than simply responding automatically to the magnetic field like a thermostat responds to temperature changes. What would ensure this need? A situation where it is essential to the explanation that the system uses the representation as a representation: the necessity of recognition of the representation's referent for the successful functioning of the bacterium, in this instance.

Thus it seems finding a way to distinguish when representational explanations are appropriate from when they are not is needed. Jerry Fodor has addressed exactly this question. He considers what differences there are between intentional and non-intentional creatures in what he calls a 'primal scene': some creature, A, 'sees' (broadly understood: the bacterium 'sees' the magnetic field) some particular object, x, that is F (instantiates the property of Fness), leading to A's behaviour, C. I will talk about this in terms of the stimulus, x, received by the agent and the relevance of the properties of x in explaining C — properties of the stimulus which are then invoked in behavioural explanations.

Fodor claims only organisms which can respond to stimulus properties such that the relation between behaviour and stimulus property is 'non-nomic' warrant intentional explanations (1986: 9-10). This is a careful way of saying that the capacity to respond selectively to coextensive properties in non-automatic

ways (unlike the magnetotactic bacterium) shows when intentional explanation is appropriate for the creature in question.

Crucial to this is Fodor's distinction between nomic and non-nomic properties. A brief gloss: nomic properties are those which cause a certain behaviour in the 'automatic' sense I have spoken of already, e.g. the bacterium's movements lawfully covary with the magnetic field. Non-nomic properties are all the rest: properties of stimuli which are invoked in behavioural explanations but do not lawfully covary with the behaviour. Fodor takes physical laws as the paradigm for nomic properties. The bacterium moves with the direction the magnetic field identically to how iron filings move in response to a magnetic field. Stimulus properties connected by physical laws to behavioural properties are then 'nomic properties', while those not so connected are non-nomic. Fodor's point is that any object has indefinitely many properties, but only some subset is subject to laws linking them with the behaviour of agents. I presume Fodor would agree that talk of 'behaviour' and 'agent' here is understood loosely so as not to question-beggingly tie non-nomic properties with all and only intentional agents like people. The aim is to find an outside purchase on which things (including rocks, bacteria and people) plausibly have intentional states.

If F is non-nomic then "there is no law that relates [x's] being [F] to A's behaviour coming to be C" (Fodor, 1986: 14). As an example he offers a human responding to a 'crumpled shirt': "I see a thing that has this property and respond to it in a way that is explained by reference to the fact that the thing has the property and that I see it" (1986: 13). This might be a verbal response affirming the crumpled shirt's presence. Unlike the bacterium's movement in response to the magnetic field, the human's verbal assent to the crumpled shirt is not nomically connected to the shirt's crumpledness.

Fodor claims that a stimulus property being implicated in the explanation of behaviour in the absence of a lawful connection between property and behaviour is a "puzzle that motivates the representational theory of mind" (1986: 14). For him, explaining "selective response to non-nomic properties" is why we posit representations in the first place. It follows that those representations must actually represent those properties (determinately, and not others, as we have seen from the JDC) for representational explanation to succeed.

In the bacterium case the magnetic field indicates both the direction of geomagnetic north (x is F) and the direction of oxygen-free water (x is G). To explain the bacterium's 'behaviour' it is irrelevant whether x is F or x is G. All that matters is that the bacterium *responds to x*, not its Fness or Gness — neither are relevant to the explanation of behaviour. To justify claiming that the bacterium represents 'the direction of geomagnetic north' then it would need to respond selectively to the magnetic field in virtue of that

property alone, and not simply in virtue of its being a magnetic field. But there is no lawful relation between 'indicating geomagnetic north' and the bacterium's behaviour. The lawful relation obtains between the magnetic field's being a magnetic field and the bacterium's magnetic sensor. The bacterium lacks the capacity to respond selectively to the non-nomic properties of the magnetic field, hence we should not ascribe an intentional explanation in this case.

In contrast, if I lend money to some x where x is a person, then my lending that individual money is not explained merely by their being x but rather by their having the property of (say) being my friend. In that case it is the *Fness* of x which is explanatorily relevant, not just the stimulus being x. It is not the case that my lending money is explained merely by the debtor being a person. I lend money only to people who possess properties like 'being a friend'. Moreover, the explanation requires that I recognise the individual as a friend to make sense of the fact that I do lend them money. In other words, I must represent this person as being a friend for the explanation to work because there is no nomic property of people which I respond to by lending them money.

Fodor's insight allows us to see that representational explanations are needed only when there is no lawful connection between stimulus property and behaviour. Only representing the Fness of x (which cannot be accounted for by causal covariance because no natural law connects Fness to the behaviour) can explain this and this must be accounted for in mental representation. Contrast this picture with representational naturalism which — following the primacy of physics in *Physicalism — just is* seeking lawful connections between mental states and the things they are about. If we only require representations to explain selective response to non-nomic properties then it seems an intentional explanation of behaviour is *ipso facto* unrecoverable from a naturalistic starting point.

We are now in a position to define 'representation as':

> *'Representation as'*: a representation user, Q, tokens a representation $\mathfrak{R}(p)$ which represents x as F to Q by having content p = x is F. The Fness of x is essential to the causal-explanatory role of $\mathfrak{R}(p)$, and F is a non-nomic property.

'Representation as' serves to codify the fact that to ensure a representational explanation is needed Q must identify x as F in just the way Millikan and Dretske (1995) required. It is only through this that we can ensure that the Fness of x is essential to the explanation by building in that Q must recognise that x is F in

order to use the representation in line with the JDC: in virtue of the fact that $\mathfrak{R}(p)$ represents x as F. The problem is ensuring that x's Fness is essential to causing Q's behaviour. I argued that coextensive properties which are non-nomic in Fodor's sense can't be adequately differentiated between according to naturalistic approaches. Now I will suggest that there is no way to ensure this without introducing intentional terms like 'recognises that' into my definition of representation, which was:

> (*MR2*): $\mathfrak{R}(p) =_{Def}$ an internal state, $\mathcal{S}$, that stands in for some object or state of affairs, $\mathcal{O}$, (by having a content, *p*) such that $\mathcal{S}$ is capable of causing behaviour in virtue of standing in for $\mathcal{O}$.

$\mathcal{S}$ is capable of causing behaviour in virtue of standing in for $\mathcal{O}$ only if $\mathcal{S}$ does stand in for $\mathcal{O}$ by having the right content for the representational explanation, and if $\mathcal{S}$ is used by (the representation user or system) Q in virtue of having that content. The difficulty is ensuring that Q does use $\mathcal{S}$ in virtue of that content — passing the JDC. I argue one intuitive way to ensure this is *'Representation as'*. So, when $\mathcal{O}$ is the state of affairs or object x which is F, $\mathcal{S}$ must represent x as F and Q must recognise that $\mathcal{S}$ represents x as F.

Clearly a definition of mental representation cannot mention 'represents' in the definiens on pain of circularity. Yet we need a way of building up that $\mathcal{S}$ causes action by standing in for $\mathcal{O}$ which ensures Q takes x to be F, or recognises that x is F. Is there a way to paraphrase this and still get across the sense of *'Representation as'*? Candidates for replacement phrases might be: S 'sees', 'knows', 'understands', 'recognises', 'judges', 'thinks', 'identifies', 'believes', 'accepts', or 'grasps' that x is F. However, a quick examination of these phrases shows that they are all intentional terms in Chisholm's sense. It seems any way of rendering *'Representation as'* will inevitably feature intentional terms. But this is not unexpected, and fits with Chisholm's branch of the *Chisholm-Quine Dilemma*. Accepting this for now yields (*MR3*):

> (*MR3*): $\mathfrak{R}(p) =_{Def}$ an internal state, $\mathcal{S}$, of Q which (i) stands in for some object or state of affairs, $\mathcal{O}$, which is that x is F; (ii) has the content *p* = 'x is F'; (iii) causes Q's behaviour because Q recognises that $\mathcal{S}$ has the content 'x is F'.

I have chosen 'recognises that' in (iii) to reflect '*Representation as'*, though other intentional locutions might do just as well.

My aim was to identify a deficiency of representations defined through representational naturalism and trace this to a common cause. I think the irreducible intentionality of *'Representation as'* underlines this common cause. It is simply that the *Chisholm-Quine Dilemma* has not been successfully navigated. The fundamental incompatibility of naturalistic approaches and intuitions (*PT1*) and (*PT2*) has not been properly recognised. If I am correct, then intentionality is a *Counterexample* to *Physicalism*. Representational naturalism invites the *Distinctively Mental Criticism* because it fails to bear out our intuitions about the mind. A related criticism is that representational naturalism fails to yield determinate content or to convince that the content is essential for the explanation. A way to solve all these problems is available through something like *'Representation as'*, but this is irreducibly intentional and thus leads to a *Homunculus Regress* under representational naturalism, hence the two are flatly incompatible. This exposes a difficult choice: either we reevaluate our intuitions in light of the success of our theories, or we should accept that accounting for intentionality leads to the equal plausibility of *Physicalism* and Property Dualism. To recap this point: intentionality cannot be explained satisfactorily under representational naturalism, suggesting Property Dualism is the correct ontology. However, this is not decisive because a proponent of *Physicalism* can respond — albeit weakly — that a complete future physics may be able to give a physical explanation for intentionality. However, my view is that, in the absence of a strong argument to believe future physics can do this we are more justified in holding Property Dualism.

One might object that intentional talk cannot be included in (*MR3*) because mental representations are subpersonal theoretical posits, as Ramsey and Morgan explicitly hold. Recognition and identification of the referent of a representation is an act on the personal level; there is no 'seeing that' or 'taking that' at the subpersonal level. As a first response, note that accounts like Indicator Semantics and Teleosemantics do not explicitly address the distinction. Moreover, discussions of mental representation in visual perception frequently blur the lines themselves by talking about 'seeing', 'identifying' and 'taking' objects to be this or that (Fodor, 1987: 197-108; Dretske, 1995: 30-32). Thus any analysis of mental representation involving the early and important work of Dretske and Millikan cannot avoid engaging in the same conceptual unclarity around personal/subpersonal which their work exhibits.

What all this means is unclear. I think it points to two orthogonal possible starting points on the mind. If we choose the picture that accepts (*PT1*) and (*PT2*) then I argue any naturalist account fails. This is because without *'Representation as'* it does not pass the JDC, but with *'Representation as'* it is engulfed by

the *Homunculus Regress*. If we choose the picture that denies (*PT1*) and (*PT2*) then a naturalist account can succeed but at the expense of determinate representational content and also failing to pass the JDC. On this interpretation of my results what is at issue is still what we mean by 'representation' in the first place. We cannot do justice to either (*PT1*) or (*PT2*) by thinking of representation as subpersonal because the account collapses as circular. But to do justice to them, it follows that representational naturalism must be denied and that representations must be personal-level posits.

If we need to invoke personal-level properties to get representations to do what their causal-explanatory role requires — as has been my argument in proposing *'Representation as'* — then it follows that representations cannot be subpersonal. *'Representation as'* means applying personal-level properties to mental representations, i.e. entails that mental representations are at the personal level. To maintain, as Ramsey does, that representations are subpersonal then it is necessary to deny (*PT1*) and (*PT2*). If my reading of Ramsey's JDC and my solution through *'Representation as'* is right, it follows that no subpersonal theoretical posit actually could pass the JDC. The JDC itself is predicated on representations being the kind of phenomena which require personal-level notions like '*Representation as'* to be able to have sufficiently determinate, causally efficacious content.

From here all that can be said is that the logical conclusion of representational naturalism of the kind Ramsey endorses is that the term 'mental representation' is not really 'representational' in the pre-theoretical sense at all. These theoretical posits may play valuable roles in subpersonal explanations of cognition, but they no longer bear significant resemblance to their pre-theoretical namesakes.

4.4 Conclusion

I have argued that naturalistic approaches to mental representation have so far failed to adequately account for the phenomena. They fail to meet Ramsey's JDC because there is no reason to suppose a representation user uses the representation to cause behaviour in virtue of a particular representational content, in line with preserving the characteristics — aspectual shape, fine-grained individuation between coextensive properties — we ascribe to intentional contents. I proposed one solution to this problem of indeterminacy (*'Representation as'*) which is irreducibly intentional. Thus, I argued, intentionality serves as a *Counterexample* to *Physicalism*. If we admit the intuitions underpinning our conception of intentionality have force then we have no decisive arguments in favour of either a physicalist or a property dualist metaphysics of mind. Moreover, I suggested that Ramsey's JDC cannot be passed without invoking intentional terms and concepts when it comes to recognition of the referent of a representation. Therefore, representations must be at the personal level of explanation, and so any theory that treats them as

subpersonal is unable to pass the JDC and fulfil the causal-explanatory role mental representations are taken to play. Consequently, if *'Representation as'* remains the only way to pass the JDC, then representation is shown to be personal-level posit and subpersonal notions of representation — though possibly useful for explanation — should give up the claim to be significantly similar to everyday, non-mental representations.

Bibliography

Braddon-Mitchell, David and Frank Jackson. *The Philosophy of Mind and Cognition*. Blackwell, 2007.

Brentano, Franz. *Psychology From an Empirical Standpoint*. Edited by Oscar Kraus. New York: Routledge, 1995.

Byrne, Alex. "Intentionality." In *The Philosophy of Science: An Encyclopedia*. Edited by Sarkar, Sahotra, and Jessica Pfeifer. Routledge, 2006.

Chisholm, Roderick M. *Perceiving: A Philosophical Study*. Ithaca, NY: Cornell University Press, 1974.

Churchland, Paul M. "Eliminative Materialism and the Propositional Attitudes." *The Journal of Philosophy* 78, no. 2 (1981): 67–90.

Clark, Andy. "Reasons, Robots and the Extended Mind." *Mind & Language* 16, no. 2 (2001): 121–45.

Craik, Kenneth. *The Nature of Explanation*. Cambridge: Cambridge University Press, 1943.

Crane, Tim. *Elements of Mind: An Introduction to the Philosophy of Mind*. Oxford: Oxford University Press, 2001.

— "Intentionality". In *Routledge Encyclopedia of Philosophy*. Edited by Edward Craig. London: Routledge, 1998b.

— "Intentionality as the Mark of the Mental." *Royal Institute of Philosophy Supplement* 43 (1998a): 229-251.

— *The Mechanical Mind: A Philosophical Introduction to Minds, Machines and Mental Representation*. London: Routledge, 2003.

Crane, Tim, and Katalin Farkas. *Metaphysics: A Guide and Anthology*. Oxford: Oxford University Press, 2004.

Cummins, Robert. *Representations, Targets, and Attitudes*. Cambridge, MA: MIT Press, 1996.

— *The Nature of Psychological Explanation*. Cambridge, MA: MIT Press, 1983.

Dennett, Daniel C. *Brainstorms: Philosophical Essays on Mind and Psychology*. Hassock: Harvester Press, 1979.

— *Content and Consciousness*. London: Routledge & Kegan Paul, 1969.

Dretske, Fred. *Knowledge and the Flow of Information*. Oxford: Basil Blackwell, 1981.

— "Misrepresentation." In *Belief: Form, Content, and Function*, edited by R Bogdan. Oxford University Press, 1986.

— *Naturalizing the Mind*. Cambridge, MA: MIT Press, 1995.

Fodor, Jerry. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press, 1987.

— "Semantics, Wisconsin Style." *Synthese* 59, no. 3 (1984): 231–50.

— *The Language of Thought*. New York, NY: Crowell, 1975.

— "Why Paramecia Don't Have Mental Representations." *Midwest Studies In Philosophy* 10, no. 1 (1986): 3–23.

Garson, James. "Connectionism", *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition). Ed. Edward N. Zalta. URL = <https://plato.stanford.edu/archives/win2016/entries/connectionism/>.

Godfrey-Smith, Peter. "Mental representation, naturalism, and teleosemantics." In *Teleosemantics: New Philosophical Essays*. Edited by Graham Macdonald and David Papineau. Clarendon Press, 2006.

— "Misinformation." *Canadian Journal of Philosophy* 19, no. 4 (1989): 533–50.

Hall, Richard J. "Does Representational Content Arise from Biological Function?" *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1990, no. 1 (1990): 193–99.

Haugeland, John. “Representational genera”. In *Philosophy and Connectionist Theory*. Edited by William Ramsey, Stephen Stich, & David Rumelhart, 61-91. Hillsdale, NJ: Lawrence Erlbaum, 1991.

— “The Nature and Plausibility of Cognitivism.” Behavioural and Brain Sciences 1, no. 2 (1978): 215–26.

Jacob, Pierre. “Intentionality.” In *The Stanford Encyclopedia of Philosophy* (Winter 2014 Edition). Edited by Edward N. Zalta. URL = <https://plato.stanford.edu/archives/win2014/entries/intentionality/>.

Johnson-Laird, Philip. “Mental Models”. In *Foundations of Cognitive Science*. Edited by Michael I Posner, 469–99. Cambridge, MA, USA: MIT Press, 1989.

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

Lycan, William G. “Form, Function, and Feel.” *The Journal of Philosophy* 78, no. 1 (1981): 24–50.

Millikan, Ruth Garrett. “Biosemantics.” *The Journal of Philosophy* 86, no. 6 (1989): 281–97.

— “Biosemantics.” In *The Oxford Handbook of Philosophy of Mind*. Edited by Brian P McLaughlin, Ansgar Beckermann, and Sven Walter. Philosophy of Mind. Oxford : Clarendon Press, 2009.

— “Compare and Contrast Dretske, Fodor, and Millikan on Teleosemantics.” *Philosophical Topics* 18, no. 2 (1990): 151–61.

— *Language, Thought, and Other Biological Categories: New Foundations for Realism*. Cambridge, MA: MIT Press, 1984.

Morgan, Alex. “Representations Gone Mental.” *Synthese* 191, no. 2 (January 29, 2014): 213–44.

Neander, Karen. ”Teleological Theories of Mental Content”. In *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition). Edited by Edward N. Zalta. URL = <https://plato.stanford.edu/archives/spr2018/entries/content-teleological/>.

Peirce, C. S. *Collected Papers of Charles Sanders Peirce*. Edited by Charles Hartshorne and Paul Weiss. Cambridge, MA: Harvard University Press, 1931.

Poland, Jeffrey S. *Physicalism : The Philosophical Foundations*. Oxford : Clarendon Press, 1994.

Quine, W V. *Word and Objec*t. Cambridge, MA: MIT Press, 1960.

Ramsey, William M. *Representation Reconsidered*. Cambridge University Press, 2007.

— “Untangling Two Questions about Mental Representation.” *New Ideas in Psychology* 40 (2016): 3–12.

Searle, John R. “Minds, Brains, and Programs.” *Behavioural and Brain Sciences* 3, no. 3 (1980): 417–24.

Stich, Stephen. “What Is a Theory of Mental Representation?” *Mind* 101, (1992): 243–61.

Swoyer, Chris. “Structural Representation and Surrogative Reasoning.” *Synthese* 87, no. 3 (1991): 449–508.

von Eckardt, Barbara. “The Representational Theory of Mind.” In *The Cambridge Handbook of Cognitive Science*. Edited by Keith Frankish and William Ramsey, 29–49. Cambridge: Cambridge University Press, 2012.

—*What Is Cognitive Science?* Cambridge, MA: MIT Press, 1993.

Yalowitz, Steven. "Anomalous Monism.” In *The Stanford Encyclopedia of Philosophy* (Winter 2014 Edition). Edited by Edward N. Zalta. URL = <https://plato.stanford.edu/archives/win2014/entries/anomalous-monism/>.