What Are Neural Representations? A Cummins Functions Approach

Abstract

This paper introduces the Cummins Functions Approach to neural representations (CFA), which aims to capture the notion of representation that is relevant to contemporary neuroscientific practice. CFA shares the common view that 'to be a representation of X' amounts to 'having the *function* of *tracking* X', but maintains that the relevant notion of *function* is defined by Robert Cummins's account. Thus, CFA offers a notion of neural representation that is dependent on explanatory context. I argue that CFA can account for the normativity of neural representations, and defend its dependence on explanations.

Author Contact Information

Ori Hacohen
Department of Cognitive and Brain Sciences
Hebrew University of Jerusalem
Jerusalem 9190501, Israel
ori.hacohen@mail.huji.ac.il

Acknowledgements

Many thanks to Oron Shagrir for his support and guidance. I also thank Arnon Levy, Gal Vishne, Lotem Elber-Dorozko, the members of the neurophilosophy reading group at the Edelstein Center for the History and Philosophy of Science, and Chiara Elettra Ferrario for significant and enlightening comments. This work was supported by the Edelstein Center for the History and Philosophy of Science.

1. Introduction

At first pass, a physical representation is a vehicle of content or information. Within contemporary neuroscientific explanations, it is common practice to refer to internal neural states or structures as "representations", as carriers of information, or as coding a signal. And yet the nature of such "neural representations" is far from clear. How should we define the notion of representation that is relevant to contemporary neuroscience? Furthermore, what makes something a *neural representation* of some specific content X? A theory of content for neural representations should be able to answer these questions. Thus far, no existing theory has been widely accepted as achieving this task.

This paper introduces the Cummins Functions Approach to neural representations (CFA). CFA combines two well-known ideas, the first of which we will call "the shared conception", and the second is Cummins's (1975) notion of function. Both will be described in section 2. In short, "the shared conception" is the broadly defined view that 'to be a representation of X' is ultimately 'to have the *function* of *tracking* X'. Through one interpretation or another, this basic idea is agreed upon by many of the existing theories of content for neural representations. Where CFA diverges from existing views is by appealing to *Cummins functions* (as opposed to historical functions) in unpacking this shared conception. This will define the Cummins functions approach to neural representations, as described in section 3. Section 3 will also include a brief example of the way CFA can account for representations in neuroscience.

The rest of the paper will then be devoted to defending this approach, in the face of what I take to be its most significant possible objections. In section 4, I argue that Cummins functions can enable an account of the normativity of neural representations. Section 5 expands on CFA's appeal to explanatory context, clarifying the extent and the significance of this appeal. I then continue to defend CFA's dependence on explanatory context in section 6. I argue that it does not contradict the current neuroscientific practice, and that CFA's violation of the naturalistic constraint should not dissuade us from accepting it.

2. Preliminaries

2.1. The Shared Conception

While, as stated above, there is no consensus theory of content philosophers have subscribed to, we *can* point to a broad approach that is widely shared. Following Morgan and Piccinini (2017) we can define "the shared conception" of existing theories of content: "While this summary elides over a great many differences, proponents of naturalizing intentionality arrived at a shared conception of what a naturalistic theory of content should look like, and indeed of the very nature of intentionality. They conceptualized intentionality as a special kind of causal-informational relation between an internal state – a representation – and a distal entity, where the internal state has the function of responding to, or tracking, the distal entity. If the internal state fails to track that entity, it can be evaluated as incorrect or inaccurate." (p. 10)

In short, we define "the shared conception" as:

(*) A **representation** of X has the *function* of *tracking* X.[1]

Of course, for the shared conception (*) to become an actual theory of content for neural representations, one would have to properly explicate how the notion of *function* and the notion of *tracking* are best understood, and this is where different theories diverge. While the broadly defined "shared conception" is not committed to any specific definition of these notions, it still carries some basic understanding of both. As for *tracking*- it is generally assumed to be some causal-informational relation that (on its own) does not allow for mistakes. Following Neander (2017a, p.83) we can consider this as a commitment to some "natural-factive information relation", such as Grice's (1957) "natural meaning". And as for the notion of function- on the shared conception it must enable precisely this talk of normativity and *mis*representation. We can again follow Neander (2017a) and consider the shared conception as committed to a notion of "normal-proper function", where: "'normal-proper function' is here identified by ostension. The relevant notion of a *normal-proper function* is the one that most centrally underwrites talk of normal function, of systems functioning properly, of malfunction, dysfunction, abnormal functioning, impaired functioning, and functional deficits." (Neander 2017a, p. 52)[2]

[1] It is worth noting that this broad idea about the nature of representation goes beyond talk of neural representations or "natural representations". In fact, it is often our understanding of non-natural representations that best motivates this approach (see, for example, (Dretske 1988, pp. 59-62)).

[2] Identifying the notions of *tracking* and *function* along the lines of (Neander 2017a) also means that the shared conception (*) is entirely consistent with Neander's *broad* definition of informational teleosemantics (see Neander 2017a, p. 86).

But these broad, or ostensive, descriptions of *tracking* and *function* still have to be unpacked by actual theories of content. And if we take a look at how existing theories interpret the relevant notion of *function*, there seems to be a common thread among them- they all define neural representations by appealing to some *historical* notion of function. On the most popular approach the function of a representation, or a representational vehicle, is defined by what it was *selected for* through some evolutionary process of selection (e.g. Millikan 1984, 1989, 2004, Neander 1995, 2017a). Others have also appealed to a natural process of learning (e.g. Dretske 1988, 1995) or to selection in virtue of "contribution to an organism's persistence" (Shea 2018). While these theories might differ in pointing out the relevant type of historical process, they still all define representations through a historical notion of function. The aim of this paper is to promote an appeal to a different notion of function- that of Cummins's (1975) account.[3]

[3] This paper stays away from any debates over the relevant notion of *tracking*. But to state a few possible interpretations we can mention, for example, Dretske's (1988) "indication" relation, as well as his (1981) formal definition of information in terms of conditional probabilities. Stampe (1977) gives a causal analysis of the natural information relation, as does Neander (2017a). Shea (2018) talks of 2 kinds of "exploitable relations"- correlational information, defined in terms of conditional probabilities, and structural similarity, which can both be considered as a type of *tracking* for our current concerns (despite the fact that Shea's view isn't exactly on par with the shared conception as described above). The shared conception (*) generalizes over these and other definitions of *tracking*.

2.2. Cummins Functions

Debates over 'what is the relevant notion of *function*' are hardly limited to the issue of representations. There is a wider philosophical debate, about understanding the relevant notion of function for scientific explanations in general, and physiological explanations in particular. And within *that* debate *Cummins functions* play a significant role. In what follows, Cummins defines when is it the case that a certain internal component x (e.g. the heart), within a containing system s (e.g. the circulatory system), has the function ϕ (e.g. to pump blood). Note that the definition is relative an analytical account A (i.e. an explanation) of some capacity ψ that is performed by the containing system.

Cummins Functions:

> "x functions as a ϕ in s (or: the function of x in s is to ϕ) relative to an analytical account A of s's capacity to ψ just in case x is capable of ϕ-ing in s and A appropriately and adequately accounts for s's capacity to ψ by, in part, appealing to the capacity of x to ϕ in s." (Cummins 1975, p. 762)

On this view x has the function ϕ, not because it evolved to have that function, or it was selected for that function in some other manner. Rather x has the function ϕ because the fact that x is capable of ϕ allows us to explain how the containing system achieves a more general capacity (ψ). Importantly, this means that functions are defined relative to an explanation: “To ascribe a function to something is to ascribe a capacity to it which is singled out by its role in an analysis of some capacity of a containing system.” (Cummins 1975, p. 765)

We should note that *Cummins functions* are sometimes referred to as *role functions, causal-role functions, system functions,* or *systematic functions*. The latter is also how Cummins himself refers to such functions. There are also some variations on how we should define such functions, but we will focus mainly on Cummins's classic (1975) account. Accordingly, we will continue to refer to these as *Cummins functions*. And with regards to the general debate over how to understand functional attributions in the physiological sciences, Cummins functions are often regarded as a viable candidate.[4]

Importantly, though, this paper does not take a stance with regards to the interpretation of functional attributions in any general sense. I only wish to claim that Cummins functions can figure into an account of neural representations. Clearly, that would carry *some* commitment to Cummins's account of function, but that does not necessarily mean that *all* functional attributions in neuroscience should be understood in this manner, and it definitely does not mean that functional attributions in the biological or physiological sciences should always be understood as Cummins functions. In this paper, we restrict our commitment to Cummins functions only to those functions that are relevant to defining representations. And I believe that, in this context of neural representations, Cummins functions have been grossly overlooked. Let us now see how an appeal to Cummins functions can feature in a theory of content for neural representations.

[4] Neander (2017b) discusses the notion of function used in "explaining how bodies and brains operate" and states that "Cummins' (1975) notion, as originally defined [...] is often taken to be the clearly relevant notion for such an explanatory context." (p. 1147) She illustrates this with a wide variety of quotes from the relevant literature, before going on to argue against this apparent consensus.

3. The Cummins Functions Approach

3.1. Definition

The Cummins Functions Approach (CFA) simply combines the shared conception (*) with Cummins's (1975) account of function. This will define a notion of internal representation that is dependent on explanatory context, just as Cummins functions are. But since we are currently only interested in the *neural* representations that are relevant to *neuroscientific* explanations, we will restrict our definition accordingly. While Cummins's (1975) account can define the function of *any* internal component, relative to some explanation of its containing system, CFA's definition of *neural representations* will focus only on *neural* internal components, relative to some *neuroscientific* explanation of its containing system. Other than that, CFA directly follows Cummins's definition of function quoted above. There, Cummins explained when some capacity ϕ of an internal component x can be regarded as its function. Now, if we look only at those cases where ϕ is a capacity of *tracking*, and assume (*) that representations are defined by a *function* of *tracking*, then Cummins's definition of function also entails a notion of representation:

The Cummins Functions Approach (CFA) to *neural representations*:

Suppose that x is an internal neural component of s, that s has some capacity to ψ, and that A is a neuroscientific explanation of s's capacity to ψ. Then,

x is *a neural representation of* X, relative to A, just in case:

1. x is capable of *tracking* X, and
2. A appropriately and adequately accounts for s's capacity to ψ by, in part, appealing to the capacity of x to *track* X.

There are a couple of points worth stressing here, before moving on. First, CFA does *not* amount to a well-defined notion of neural representation, since we have not offered a definition of *tracking*. This is why we consider this as an *approach* to neural representations, and not as an actual theory of content. This approach generalizes over various possible definitions of tracking, and as such it defines a *family* of possible theories of content.

Second, on this approach neural representations are only defined *relative to a given explanation*. And following Cummins, we demand that the explanation *A* "appropriately and adequately accounts for *s*'s capacity to ψ." Yet, we will refrain from defining what makes an explanation "appropriate and adequate". CFA is not committed to, and not dependent on, any account of explanation in general, or neuroscientific explanations in particular. *Given* a neuroscientific explanation, CFA will enable our understanding of the neural representations that are considered within it, under the assumption that this explanation "appropriately and adequately explains *s*'s capacity to ψ". Furthermore, CFA is also not committed to *Cummins's* view of scientific explanations as functional analyses- whereby a complex capacity is explained by appealing to the systematic organization of simpler capacities. We *will* assume that the Cummins function of a particular component is defined by its contribution to the capacity of a containing system, relative to some explanation *A*, but we need *not* assume that the explanation *A* is itself a functional analysis.[5]

[5] It should also be clear that while this paper argues for a "Cummins functions approach" to neural representations, it does *not* convey Cummins's own approach to neural representations, which is far more focused on isomorphisms and structural similarity (see Cummins 1989, 1996).

3.2. Example

Let us now illustrate how CFA can account for the notion of representation found in neuroscientific explanations. We will consider the vestibulo-ocular reflex (VOR) which is the phenomenon of maintaining a steady eye gaze while the head moves. An explanation of this phenomenon, adapted from (Robinson 1989),[6] is very briefly described in figure 1.

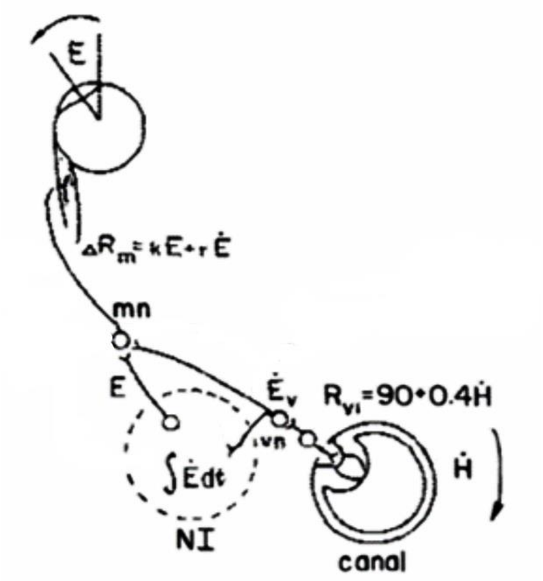

Figure 1. The explanation of VOR.

"On the right the canals transduce head velocity, $\dot{H}$, and report it, coded as the modulation of the discharge rate, R_{V1}, of primary vestibular afferents to the vestibular nucleus, vn. This signal becomes an eye velocity command for vestibular movements, $\dot{E}_v$, which is sent directly to the motoneurons, mn, and to the neural integrator, NI, to provide the needed position signal E. These signals provide those needed by the motoneurons modulating by ΔR_m." (Robinson 1989, p.35)

[6] This explanation of VOR is also discussed in (Bechtel & Shagrir 2015, Shagrir 2018).

For the purposes of this paper, we need not delve into the details of this explanation of VOR. Suffice to say that it describes the mechanism underlying VOR as an information-processing mechanism, whereby internal neural components are regarded as carrying content or information. But still, if we wish to claim that these internal components truly are *neural representations*, then we need an account of what *being a neural representation* amounts to. For example, this explanation clearly regards the primary vestibular afferents as carriers of a head velocity signal, coded by their discharge rate (R_{V1}). But in order to claim that these neurons are a *neural representation* of head velocity, we need to show that they meet some clearly defined criteria for *being a representation of head velocity*, and that is what theories of content aim to achieve.

CFA can account for the representational status of the primary vestibular afferents, though we will need to say something more about *tracking*. As defined above, CFA is uncommitted to *any* definition of *tracking*, but some assumption will be necessary in order to illustrate this approach on any specific example. In this case, we will assume that the fact that the discharge rate of the primary vestibular afferents correlates with head velocity (as described in figure 1) can be taken as meaning that the primary vestibular afferents *track* head velocity. This is a nontrivial assumption but also an uncontroversial one. It is consistent with existing possible definitions of *tracking* and I believe most philosophers would accept it. For us, the interesting question is not whether the vestibular afferents *track* head velocity, but whether it is their *function* to do so.

As stated, existing theories of content have dealt with such questions through some *historical* notion of function, such as claiming that the primary vestibular afferents have the function of tracking head velocity because they were naturally *selected for* this capacity in the evolutionary process. But now, let us fit this example to the definition of CFA given above. We have an internal neural component x- the primary vestibular afferents, of some system s- the brain[7], where s has a certain capacity ψ- to maintain steady eye gaze while the head moves. We also have a neuroscientific explanation A of how the brain maintains its gaze, i.e. an analytical account of s's capacity to ψ. A is the explanation of VOR found in (Robinson 1989), and described in figure 1. Thus, according to CFA, we can state that:

The primary vestibular afferents (x) are a *neural representation* of head velocity (X), relative to A, just in case:

1. The primary vestibular afferents are capable of *tracking* head velocity, and
2. A appropriately and adequately accounts for the brain's capacity to maintain steady eye gaze while the head moves (VOR) by, in part, appealing to the capacity of the primary vestibular afferents to *track head velocity*.

We have already assumed that condition 1 holds- that the primary vestibular afferents are capable of *tracking* head velocity. We also assumed that this tracking relation is identified by the neurons' discharge rate correlating with head velocity. Therefore, I believe it is quite clear that condition 2 holds as well. The fact that the vestibular afferents track head velocity seems to be the *only* capacity of theirs that the explanation of VOR appeals to. Of course, their basic ability to transfer a signal via neural excitation or inhibition must also be relevant to the explanation, but that would clearly be true for *all* components of the mechanism. And it is not *this* ability that

[7] We can also consider s as the specific subsystem of the brain which is comprised of the vestibulo-ocular mechanism. But regarding the brain as s is simpler and good enough for our needs.

accounts for the role of this specific component – the primary vestibular afferents – in enabling VOR. *Tracking head velocity* is what accounts for that. *Tracking head velocity* is the capacity of the primary vestibular afferents which, in part, allows us to explain the functioning of the containing mechanism and understand the phenomenon of VOR. Thus condition 2 would hold, and we can conclude that, according to CFA, the primary vestibular afferents are a *neural representation* of head velocity, relative to the explanation of VOR.

I contend that CFA offers the right path towards understanding the notion of representation that is at use throughout current neuroscientific practice. Let us now scrutinize this claim through some possible challenges.

4. Normativity

Some objections to CFA can be derived from existing objections to Cummins's (1975) account of function. The first is that Cummins functions do not allow for *mal*function. If a component's function is defined by *something it does* that contributes to a containing system, then if for some reason the component does not do this particular something, then it does not have a function. Thus, no room is left for the possibility of *mal*function. If that is indeed the case, then CFA would not be able to account for the normativity of representations. Furthermore, we defined CFA as committed to the shared conception (*), which is in turn committed to *some* notion of "normal-proper function"- defined by ostension as a notion of function that can underwrite talk of "normal functioning" and "malfunction". We must therefore show that Cummins functions can play this role.

Neander (2017b) shows that Cummins's original (1975) paper does not amount to an account of normal-proper functions. But while it is true that (Cummins 1975) did not clarify how it allows for malfunction, this has since been clarified elsewhere. As Godfrey-Smith (1993, p.200) states: "Although it is not always appreciated, the distinction between function and *mal*function can be made within Cummins' framework […]. If a token of a component of a system is not able to do whatever it is that other tokens do, that plays a distinguished role in the explanation of the capacities of the broader system, then that token component is malfunctional. The concept of malfunction is context-dependent on Cummins' view, just as the concept of function in general is."

Cummins himself has also explained the normativity of his account in a similar manner (see Cummins 1996, p. 116, Cummins & Roth 2010, p. 79). The idea is that function is assigned to a *type*, and based on a capacity of that type that contributes to the functioning of the containing system. Tokens have their function in virtue of their type. Once a specific token does not perform this capacity, and hence does not contribute accordingly to the containing system, that is a malfunction of that token. Neander (2017b, p. 1164) mentions this view, and acknowledges that it does "make space for token malfunction", she just doesn't consider this as Cummins's (1975) approach. Neander considers this as Godfrey-Smith's (1993) view of *modified* Cummins functions. But obviously, how we *call* the relevant notion of function, and who we attribute it to, isn't the main issue here. What matters is that we consider Cummins functions as allowing for malfunction and normative distinctions as illustrated above.

Accordingly, CFA *will* be able to account for *mis*representations. As representations are defined by a function of tracking, misrepresentations are defined by a *mal*function of tracking. Consider the example of VOR above. We assumed that there exists a tracking relation between the primary vestibular afferents and head velocity, and that it is the *function* of these neurons to track head velocity in this manner. The tracking relation, however it is ultimately defined, must hold between types, and there can be cases of tokens which, for some reason, do *not* track head velocity. If a person's vestibular afferents do not track head velocity, then they do not properly perform the tracking function of the vestibular afferents. They cannot contribute what they are meant to contribute to the analyzed capacity (VOR). That is a token malfunction of tracking, hence a misrepresentation.

The normativity of representations is thus dependent on the normativity of Cummins functions, which is enabled by the type-token distinction described above. But critics have claimed that Cummins functions will run into trouble once we try to describe how the function of a type is determined. Since some tokens of this type will exhibit the relevant effect, while others might not, how do we define this effect as the function of the type? Neander (2017b, pp. 1165-1166) raises this concern within a series of questions that she poses for the proponent of what she refers to as "modified" Cummins functions. When Garson (2019) makes this point to criticize Cummins functions he initially assumes that the type-token malfunction distinction must depend on statistical considerations (e.g. the percentage of tokens that exhibit the relevant effect) to assign a function to a type. "This is, at root, to rely on a statistical norm for making sense of dysfunction." (p. 1152) But that is overlooking one of the most significant features of Cummins functions- that they are defined *relative to an explanation of a more complex capacity*. An effect is defined to be the function of a type if a scientific explanation identifies this effect as contributing to the capacity of a containing system (the capacity which is being explained). And that is not necessarily dependent on the percentage of tokens that exhibit this effect.

Garson goes on to object to this proposal – by which the explanation plays a role in defining the function of a type – as well. He claims it amounts to a notion of function that is far too subjective. We will discuss this objection, and how it might affect CFA, in the next section. For now, what should be clear is that *if* we can define the function of a type by its role as described in a scientific explanation, then we can account for malfunction, relative to that explanation. And in this sense, we will be able to consider Cummins functions as normal-proper functions. Neander (2017b) seems to accept this, when she acknowledges that an "instrumentalist" approach *would* be capable of dealing with her questions of how function is defined for a type: "The instrumentalist can deny that there are general answers to be given, and can instead maintain that the answers are determined by the pragmatics of the explanatory context on a case-by-case basis." (p. 1166) Neander raises other concerns for such dependence on explanatory context, which we will soon discuss. But for our current interests, Neander does not seem to object to the claim that the proposed view – which defines the function of a type relative to an explanation and allows for token malfunction as described above – *can* be consistent with the ostensive definition of normal proper function.

Thus, we have shown the manner in which Cummins functions are normative and how that normativity can allow CFA to account for the normativity of representations. But as stated, there are further objections to this account's reliance on explanatory context. We turn to those now.

5. Understanding the Appeal to Explanatory Context

Cummins functions are defined relative to an explanation and we have just stressed how this allows us to determine the function of a type, and account for dysfunctions. We also mentioned the fact that Garson (2019) objects to this approach. He notes that both Craver (2001) and Hardcastle (2002) make statements along those lines (which define functions relative to explanatory interests), but he interprets those statements as saying that- "Simply put, dysfunction happens when a trait can’t do what we want." (Garson 2019, p. 1152) He goes on to explain that this makes no sense as functions cannot be reduced to preferences: "I’d prefer not to need sleep and water; I’d prefer if nobody had to go through the pain of childbirth or teething, either. But none of those things are dysfunctions. For that matter, I’d prefer if my hands were equipped with retractable adamantium claws. The fact that my hands can’t do what I want them to do doesn’t make them dysfunctional." (pp. 1152-3)

To start, though, it is worth emphasizing that Cummins functions aren't defined by what we *want* a type to do, they are defined by what a scientific explanation *says* they do. If there was a scientific explanation that would explain how people exhibit a certain capacity ψ by appealing to their hands' ability to produce adamantium claws, then *relative to that explanation*, the fact that Garson's hands cannot produce such claws *would* be a dysfunction, and that might explain why Garson cannot exhibit the capacity ψ. As stated in Godfrey-Smith's quote above- "The concept of malfunction is context-dependent on Cummins' view, just as the concept of function in general is."

And note that our appeal to scientific explanations also doesn't mean that scientists get to just *decide* the functions of components. Cummins's account of functions gave clear conditions- the function of a type must be a capacity that at least some tokens of this type can exhibit, and it must be the case that this capacity enables an appropriate and adequate explanation of a capacity of the containing system. At the very least, these are strong restrictions on what can count as a function. It would be hard to claim that there actually *is* an "appropriate and adequate" explanation of some human capacity ψ, that appeals to the hands' ability to produce adamantium claws. And until such an explanation is provided, no one can state that it is the hands' function to produce such claws. For that matter, scientists don't *decide* that it is the function of the primary vestibular afferents to track head velocity. They show that the primary vestibular afferents are capable of tracking head velocity, and they show, through some explanation *A*, that this capacity of the vestibular afferents enables a capacity of the containing system. Only if this is achieved, and assuming that *A* appropriately and adequately accounts for the capacity of the containing system, will we state that it is the *function* of the primary vestibular afferents to track head velocity, *relative to the explanation A*.

Thus, Cummins functions do not allow for the type of rampant subjectivity Garson (2019) discusses. But they still *are* dependent on explanatory aims. A function is only defined relative to a given explanation, and I do think explanatory aims play a role in determining the relevant explanatory context. Neander (2017b, p. 1155) also states that: "There are explanatory aims when anyone tries to explain complex or, for that matter, simple capacities. And which causal contributions ought to be mentioned in a given explanatory context will depend on one’s aims. But, on Cummins’ account, *if there are no relevant explanatory aims, then there are no functions*. Explanatory aims are *constitutive* for Cummins functions."

I am willing to accept this, and more importantly, accept what this entails for CFA. CFA will inherit Cummins functions dependence on explanatory context, and explanatory aims, and I would like to argue that there is nothing necessarily wrong with that. To start, though, let's try to better clarify the significance of the explanatory context *for CFA*. One way to do this is by briefly considering an altered variation of CFA, which takes the definition that was stated in 3.1, and attempts to rephrase it in a manner that does *not* appeal to a given explanation:

The Altered Definition[8]

Suppose that x is an internal neural component of s, and that s has some capacity to ψ. Then,

x is *a neural representation of X*, just in case:

1. x is capable of *tracking X*, and
2. *s*'s capacity to ψ is partly due to the capacity of x to *track X*.

Instead of defining neural representations relative to a given neuroscientific *explanation* (as CFA does in 3.1), this altered definition defines neural representations relative to a given neuroscientific *phenomenon* (the capacity ψ of the containing system). Does that mean that the explanatory context of CFA is actually unnecessary? I will offer two reasons to answer that question negatively, which correspond to two types of roles for the explanatory context. First, the explanatory context is necessary in order to single out *the* relevant phenomenon by which we define the representation. In fact, I would argue that the altered definition is itself implicitly committed to the explanatory context, since it only defines determinate representations relative to a

[8] I thank an anonymous reviewer for suggesting this altered definition.

specific neuroscientific phenomenon ψ. Let us illustrate this point by considering a type of content determination problem.

Consider a case where the same neural component x can enable one cognitive capacity, ψ_1, by tracking some property X_1, and enable a different cognitive capacity, ψ_2, by tracking a different property X_2.[9] What would the altered definition entail in this case? Initially, one might conclude that 'x is a neural representation of both X_1 and X_2'. But taking this route would leave us with an indeterminate notion of representation. Furthermore, to simply state that x represents both X_1 and X_2 misses the fact that the altered definition *does* distinguish the two contents- they each serve different cognitive capacities. A more accurate conclusion would therefore seem to be: 'Relative to ψ_1, x is a neural representation of X_1, and relative to ψ_2, x is a neural representation of X_2.' But what does this statement actually mean? First of all, it

[9] The extent to which such cases actually occur is of course an empirical question, that is also dependent on the definition of *tracking*. But the possibility that the same neural component can carry different kinds of information to enable different cognitive capacities, certainly seems consistent with contemporary neuroscience. To give one type of example, we can consider Hubel and Wiesel's (1962, 1968) famous findings of orientation sensitive neurons in V1, which are commonly referenced in discussions of neural representations. It is actually widely accepted that the same neurons in V1 that enable orientation detection, also enable contrast discrimination, and are simultaneously sensitive to both *orientation* and *contrast* (e.g. Gawne 2000, Tolhurst et al. 1981, Reich et al. 2001). There is also recent evidence that *color* and *orientation*, which have traditionally been regarded as encoded by distinct neurons, are actually jointly coded in V1 (Garg et al. 2019). It thus seems that these V1 neurons, which are often referred to as representations of *orientation*, also track other visual properties which are relevant to different visual capacities.

means that the objective state of affairs in the world, in which ψ_1 and ψ_2 are on equal footing, does not in itself define x as a determinate neural representation. Second, it means that if *we* focus specifically on ψ_1 (or ψ_2), *then* x is a neural representation of X_1 (or X_2).

The objective state of affairs in the world does not single out one specific neuroscientific phenomenon to define the determinate content of x. And yet, person A could *look* at the objective state of affairs in the world and determine that 1. x is capable of *tracking* X_1, and 2. The cognitive capacity ψ_1 is partly due to the capacity of x to *track* X_1. Thus, assuming the altered definition above, person A would correctly conclude that x is a neural representation of X_1. But in reaching this conclusion person A *herself* singled out the cognitive phenomenon ψ_1 (as opposed to say ψ_2), thus defining the neural representation relative to her own explanatory aims. This is the type of implicit commitment to *explanatory* context that is implied when a neural representation is defined relative to a *given* phenomenon. In other words, the neuroscientific phenomenon, while objectively real, isn't objectively *given*. Nothing objectively singles out this particular phenomenon from others, at least not to the extent that is necessary to define a determinate neural representation. Thus, when we define a neural representation *relative to a specific phenomenon,* we define it in a manner that is dependent on our explanatory aims. And whereas the altered definition introduces such explanatory dependence implicitly, CFA does so explicitly.

Moving on, besides its role in singling out the relevant neuroscientific phenomenon, the explanatory context also plays a role in singling out the relevant *tracking* relation. This can be illustrated by posing a different content determination problem for the altered definition above. We can show that even *after* we fix the relevant phenomenon ψ, the altered definition will likely be unable to define a determinate content. For example, consider the discussion in 3.2, where we appealed to CFA in order to define the primary vestibular afferents as representations of head velocity, relative to the explanation of VOR. Note that the altered definition can offer a similar conclusion. Condition 1 of the altered definition would demand that these neurons track head velocity, as we assumed they do in 3.2. And condition 2 would demand that this tracking relation enable the phenomenon of VOR. This was also illustrated in 3.2, thanks to Robinson's (1989) description of VOR mechanism. Hence, one could appeal to the altered definition to determine that, relative to the phenomenon of VOR, the primary vestibular afferents are representations of head velocity. But things get more complicated once we consider even a slightly more detailed description of the neural mechanism that underlies VOR.

In 3.2, we identified the *tracking* capacity of the primary vestibular afferents by the fact that their discharge rate is correlated with head velocity. But this discharge rate is also correlated with many other *causally relevant* properties. For example, take a brief look at the causal mechanism within the semicircular canals. As the head rotates to one direction, the endolymph fluid within the semicircular canal lags behind (due to inertia) and thus flows in the opposite direction. This fluid pushes on the cupula, tilting it opposite the direction of head movement. As the cupula is deflected, the hair cells are bent and are consequently depolarized/hyperpolarized (depending on the direction of the cupula's deflection). The depolarization/hyperpolarization of hair cells leads to the excitation/inhibition of the primary vestibular afferents.

Now, the fact that the discharge rate of the primary vestibular afferents correlates with head velocity is precisely due to the causal mechanism just described. But this means that the discharge rate of the vestibular afferents will also correlate with the velocity of the endolymph fluid, with the state of the cupula, and with the depolarization of hair cells. And note that the causal relations between these properties and the primary vestibular afferents are all *essential* to the mechanism that enables VOR. So it seems one would be correct to claim that the phenomenon of VOR is partly due to the relation between the discharge rate of the primary vestibular afferents and the velocity of the endolymph fluid, for example. If we assume (like we did with regards to head velocity) that this is a *tracking* relation, we will thus find that both conditions of the altered definition hold. The primary vestibular afferents *track* the velocity of the endolymph fluid, and the phenomenon of VOR is partly due to this tracking relation. Hence, the altered definition will also define the primary vestibular afferents as representations of the velocity of the endolymph fluid, relative to the phenomenon of VOR. And we can reach similar conclusions with regards to the state of the cupula and the depolarization of hair cells.

I suppose one might hold out hope that a proper definition of *tracking* could solve this type of content determination problem. But this seems highly unlikely. Teleosemantic theories of content have been attempting to deal with such problems for decades. And the consensus seems to be that *tracking* on its own is indeed indeterminate. Thus, teleosemantic theories appeal to *function* in an attempt to single out a specific *tracking* relation, and define a determinate representation. CFA does the same by appealing to Cummins functions. But that is *not* conveyed in the altered definition above. The altered definition only adds the condition that the tracking relation serve a more general neuroscientific capacity. And as illustrated in the VOR example, that condition is insufficient to define a determinate content. It cannot pick out one specific tracking relation that defines the representation.

CFA solves this problem by following Cummins (1975). Cummins didn't just define functions relative to a phenomenon, he defined them relative to an *analysis* of a phenomenon. Only *then* do we get a determinate function. (As quoted in 2.2: “To ascribe a function to something is to ascribe a capacity to it which is singled out by its role in an analysis of some capacity of a containing system.” (p. 765)) That is why we can state that, relative to the *explanation* of VOR, the function of the primary vestibular afferents is to track *head velocity*. And that is why, according to CFA, these neurons are representations of head velocity, and not of any other properties they might track. It is the *explanation* of VOR which singles out this particular capacity (tracking head velocity) as *the* relevant effect that these neurons contribute to the mechanism that enables VOR.

We have thus clarified the significant role of the *explanatory* context for CFA. Safe to say, I think, that philosophers are traditionally averse to the possibility of explanations, and explanatory aims, having *any* role in defining representations. But I would like to claim that when it comes to CFA, such aversion is unjustified. To that end, let us try and ease a few possible worries.

6. Defending the Appeal to Explanatory Context

6.1. Compatibility with Neuroscientific Practice

To start, the previous section raised the possibility that CFA will define the same neural component x as being a representation of X_1 in one context, and a representation of some different X_2 in another context. Perhaps some would consider this a problem, or even claim it means CFA offers an indeterminate notion of representation. This latter claim is simply false, since CFA will only define representations *relative to a given explanatory context.* Thus, the type of possibility we are discussing is *not* a case of an indeterminate representation. It is *not* a case where CFA defines a representation with two contents X_1 and X_2. Instead, this is a case where CFA defines the neural component x as a *determinate* representation of X_1 in one context, while also defining the same neural component x as a *determinate* representation of X_2 in a different context. And I believe that this is precisely the type of outcome we should want.

If CFA allows for a single neural component x to be a representation of two different contents (in two different contexts), then that can only occur in the following scenario. First, there is a single neural component x that has both the capacity to track X_1 and the capacity to track X_2. Second, there are two neuroscientific explanations A_1 and A_2, whereby A_1 explains some brain capacity ψ_1 by appealing to x's capacity to track X_1, and A_2 explains some brain capacity ψ_2 by appealing to x's capacity to track X_2. Finally, we assume that both A_1 and A_2 appropriately and adequately account for the capacities ψ_1 and ψ_2, respectively. Only under these conditions would CFA define x as having two different contents, relative to the different contexts. And again, I believe that under these conditions, we should *want* a theory that allows us to state that x is a representation of X_1 in one context, and a representation of X_2 in the other. This is the simplest, most straightforward option. And it also seems consistent with contemporary neuroscience, where scientists often regard the same neural component as carrying different kinds of information, or "representing" different things.[10]

[10] This is mentioned in footnote 9, but there are many other types of examples (e.g. Desimone et al. 1985, Pinel et al. 2004, Haxby et al. 2001).

In general, any fear that the context sensitivity of CFA will allow us to define neural representations that are in some sense "unwanted", seems to be entirely unfounded. We have already stressed that CFA's dependence on explanatory context does not allow for some rampant subjectivity where scientists do as they please. If CFA defines x as a neural representation of X, relative to some explanation A, that means that there is some successful (or "appropriate and adequate") neuroscientific explanation A according to which some phenomenon is achieved, in part, because x has the capacity to track X. Why, then, would we *not* want to consider x as a neural representation of X? Without assuming some alternative definition of neural representations, what reasoning could justify the claim that 'the fact that CFA identifies x as a representation of X is bad or counterintuitive'? It is normally precisely these cases – where a successful neuroscientific explanation appeals to an internal component that *tracks* some distal entity – which representationalists regard as evidence for the explanatory significance of representations. Thus, we should *want* a theory that defines this internal component as a representation, and that is what existing theories of content have always attempted to do.

Still, some might insist that CFA's dependence on explanatory context will necessarily mean that it is incompatible with the actual neuroscientific practice. That is because defining representations relative to explanations means defining them in a manner that is at least partially subjective and non-naturalistic. Meanwhile, neuroscientists are generally in the business of describing *real* and objective phenomena, and studies such as (Bechtel 2016), and (Thomson and Piccinini 2018) have shown that this also applies to representations.[11] Now, before I reply, it is worth

[11] I thank an anonymous reviewer for raising this point.

noting that others have drawn very different conclusions from investigations of neuroscientific explanations. Egan (2014), for example, concludes that content ascriptions *must* be pragmatic.[12] But I will not be making such a claim here. Nor do I intend to prove here that CFA is *the* correct view of neural representations, which offers the best possible account for representations in neuroscience. As I mention at the conclusion of this paper, that still remains to be seen. But in order to establish CFA as a viable approach to neural representations, I do wish to show that CFA certainly *can* be consistent with the naturalistic, objective practices of neuroscientists. To that end, let us highlight some aspects of this view that are usually associated with "realistic" approaches to representations.[13]

For starters, according to CFA, a tracking relation will only define a neural representation if it is *essential* to a neuroscientific explanation. Hence, neural representations are *essential* components of neuroscientific explanations. That is, CFA is *not* a deflationary view of representations. Additionally, I would also claim that CFA is consistent with vehicle realism. Shea (2018, p.15) states: "I will reserve the term ‘realism’ for accounts that are committed to there being real *vehicles* of content: individuable physical particulars that bear contents and whose causal interactions explain behaviour." CFA, I would argue, is consistent with this demand. The primary vestibular afferents, for example, are individuable physical particulars that bear contents and whose causal interactions explain behavior, hence *real* vehicles of content. Now, it is possible some would want to insist that "vehicle realism" should

[12] See also Egan's reply to (Bechtel 2016) in (Egan 2020).

[13] In doing so we also differentiate CFA from Egan's pragmatism, and some other existing deflationary or instrumentalist accounts of representation (e.g. Sprevak 2013, Chomsky 1995).

also imply that the vehicle has its contents *intrinsically*. I think that would be a mistake. But frankly, whether you want to call it "vehicle realism" or not, the important point is that CFA regards *real* individuable physical particulars as carriers of content.

And it's more than that. In defining a representation, CFA also demands that this real physical particular "track" some property X (its content). While we have not committed to any definition of "tracking", CFA is certainly consistent with the (very likely) possibility that this is an *objective* relation in the world. In that case, neural representations, according to CFA, are dependent on scientists finding an objective "tracking" relation between a real physical particular and some property X. Furthermore, according to CFA, this tracking relation will only define a representation if it plays an essential role in enabling a more complex phenomenon. And it is again highly likely that scientists show this by pointing at an *objectively real* causal structure, such as the physical mechanism that enables VOR. All in all, we find that CFA is consistent with the possibility that neural representations are *real* physical particulars that exhibit an *objective* "tracking" relation, within an *objectively real* causal structure. All of which, of course, must be discovered empirically by neuroscientists.

Still, as discussed in section 5, CFA also maintains the necessity of the explanatory context in *picking out* these specific (objective) phenomena. But I do hope that the discussion above helps clarify why this type of explanatory dependence does not necessarily clash with the objective practices of neuroscientists. Again, we will still need to check and see which available theory *best* conforms to the scientific practice. But it would be a mistake to dismiss CFA just because we have stated that it is dependent on explanatory context. I think there is reason to believe that CFA *can* capture the notion of representation that is relevant to neuroscience, and we have seen some evidence for the advantages of this context-dependent view. But we have yet to touch upon the fact that CFA violates the naturalistic constraint.

6.2. Against the Naturalistic Constraint

The naturalistic constraint is the demand that a theory of content be specifiable in non-semantic, non-intentional terms. Thus, CFA violates the naturalistic constraint by defining representations relative to a given explanation and with respect to scientists' explanatory aims. And since the project of accounting for neural representations is often considered synonymous with the project of finding a *naturalistic* theory of content, the possibility of violating the naturalistic constraint might seem atrocious to some. Obviously, I disagree.

Ultimately, the basic reasoning behind the naturalistic constraint is that if we are to have a theory that explains content, then it can't itself be dependent on content (or semantics, or intentionality). But importantly, CFA makes no claims towards explaining content in any general sense. It *only* aims to account for the notion of representation we find in current neuroscientific practice. And I suppose that in this regard CFA may have strayed from the norm. Traditionally, theories of content have been considered in significantly different contexts. Through theories such as the Representational Theory of Mind (RTM), or the Representational Theory of Intentionality, philosophers of mind have attempted to account for the nature of thought and mental intentionality by positing the existence of internal *mental representations*. And if *mental representations* are used to explain intentionality, then intentionality can't be used to explain *mental representations*. Thus, theories of content for such *mental representations* must be naturalistic.

Even theories of content which are specifically aimed at accounting *only* for subpersonal/nonconceptual/nonconscious representations (Shea 2018, Neander 2017a), are still normally regarded as ultimately serving some grander naturalization project in the philosophy of mind. As Shea puts it:

> "My overall philosophical strategy, then, is to start with the subpersonal and work upwards. […] If we are puzzled about how there could be space in the natural world for intentionality at all, then seeing how it arises in a range of cases in cognitive science will be a major step towards resolving the puzzle. Furthermore, seeing how representational content arises and earns its explanatory keep in these cases should prove a useful staging post on the way to tackling the more complex cases. So, an account of subpersonal representational content is part of a broader strategy for tackling the problem of intentionality." (Shea 2018, pp. 27-28)

Neander (2017a) offers the same strategy. Her theory accounts solely for subpersonal nonconceptual representations, but she also makes it abundantly clear that her *ultimate* goal is to account for mental intentionality in general.[14] Neander

[14] Neander's (2017a) is titled "A Mark of the Mental", and she immediately clarifies the book "aims to persuade readers that – while the theory it offers is limited in scope – it makes genuine progress toward a naturalistic account of mental representation." (p. 1) She also states that, if her theory is correct, then "what is left is the ramping-up problem, which is the problem of understanding how to get from a theory of content

even states that it *this* ultimate goal- accounting for mental intentionality, that should lead us to exclude Cummins functions from theories of content: "It won’t do, for example, to claim that the relevant functions are ontologically grounded in the explanatory aims of researchers (as Cummins does) and then explain intentional mental phenomena, such as the explanatory aims of researchers, as grounded in such functions. That would be circular." (Neander 2017a, p.86)

But unlike Neander (2017a) and Shea (2018), CFA is truly not concerned, either directly *or indirectly*, with the attempts to "explain intentional mental phenomena, such as the explanatory aims of researchers". CFA is not in any way intended to serve the naturalization project of thought, intentionality, or content. Not that I oppose such projects, but that's simply not the goal of CFA. CFA is offered with the sole aim of accounting for *neural representations*, which we defined in section 1 as the representations that are posited in contemporary neuroscience. We *only* care about successfully characterizing the notion of representation that is relevant to current neuroscientific practice. Why, then, must our theory of content adhere to the naturalistic constraint?

There is a common assumption that an account of neural representations *has to* also (somehow) serve a more general account of intentionality. As Sprevak states: "it is widely assumed that neural representations are more fundamental than, and ground, other representations. Neural representations ground, and are somehow responsible for, personal-level thoughts such as beliefs, desires, and intentions. Personal-level representations in turn ground conventional representations such as signs, maps, and

for nonconceptual representations to a theory of the referential power of sophisticated human thought." (p. 26) public language." (Sprevak 2013, p. 552) And this type of assumption can justify the naturalistic constraint. We cannot appeal to neural representations to ground personal level intentions if we appeal to the intentions of scientists to ground neural representations.

But I reject this line of thought. To be precise, I reject the assumption that neural representations *must* be able to play this type of 'grounding' role. Again, we identified *neural representations* as the representations that are posited in contemporary neuroscience. Now, I have no intention of opposing RTM, or the idea that subpersonal representations can somehow ground personal level thoughts. But to assume that the representations that are posited in contemporary neuroscience *are* these subpersonal representations that will ultimately serve a naturalistic account of intentionality, is a hopeful hypothesis at best. And a hopeful hypothesis cannot define a necessary constraint.

There might be good reasons for *wanting* neural representations to enable a naturalistic account of intentionality, but that doesn't mean there are good reasons to assume that has to be the case. We cannot preemptively constrain a theory of *neural representations* to be naturalistic, just because we *want* it to be the case that the same representations that neuroscientists appeal to, will also somehow ground the intentionality of thought. If we are looking for an account of *neural representations*, then that means we must look for the theory that *best accounts for the notion of representation that is relevant to neuroscience*. And whether or not a theory successfully accounts for representations in neuroscience, is *not* dependent on whether or not this theory can also help us account for the intentionality of thought. Once the task of accounting for *neural representations* is adequately detached from "the problem of intentionality", the naturalistic constraint loses its justification.

To be clear, it's only the *constraint* that loses its justification. It's only the claim that a theory of content for neural representations must *necessarily* be naturalistic that I reject here. I am not arguing against naturalistic theories of content. I am just saying that, when it comes to theories of *neural representations*, being *naturalistic* is not a necessary constraint. It could still be viewed as an advantage. I certainly agree, for that matter, that it would be nice to have a theory of neural representations that can also ground mental intentionality and content in general. And obviously CFA, at least as it is described in this paper, does not seem to achieve that. It *is* worth noting, though, that neither does any other existing account. No one has been able to actually make good on this promise to explain mental intentionality by grounding it in a naturalistic theory of content. As Sprevak notes: "Unfortunately, and despite a large investment of effort, an adequate theory of natural representation has not been forthcoming. Many contemporary philosophers suspect that representation simply cannot be naturalized." (Sprevak 2013, p. 547). So perhaps trying a different approach is not such a bad idea. Sprevak (2013) and Egan (2012, 2014), for example, have also offered alternative theories of content for neural representations that do not accept the naturalistic constraint as defined above.[15]

Much more importantly though, and the point I have been trying to make throughout this section, is that it is wrong to let external considerations restrict our account of *neural representations*. The only question we should care about is whether or not CFA best accounts for the notion of representation that is relevant to

[15] Though we should note that Egan *does* consider her theory as advancing the program of naturalizing intentionality (see Egan 2014, pp. 130-131, or Egan 2018, p. 256). Nevertheless, it is clear that Egan's theory of content does not abide by the naturalistic constraint as defined above.

contemporary neuroscientific explanations. And if it turns out that it does, we must conclude that *neural representations* really are what CFA says they are. The implications CFA might somehow carry for the naturalization program in the philosophy of mind, whether we like them or not, change nothing about that.

7. Conclusion

I believe that Cummins's notion of function holds the key to understanding *neural representations*, and that this has been wrongly overlooked in the philosophical debate thus far. This paper proposes the Cummins Functions Approach (CFA) to neural representations. On CFA, neural representations are defined by a *function* of *tracking*, and the notion of *function* is understood by Cummins's (1975) account. I illustrated how CFA can account for the notion of representation that is relevant to neuroscientific explanations, and defended it from a number of possible challenges. We saw how this approach can account for the normativity of representations, and discussed its appeal to explanatory context. While philosophers tend to oppose theories of content that appeal to explanations, I attempted to show why, as it relates to CFA, we should be willing (and perhaps happy) to accept its dependence on explanatory context.

Of course, this is only the beginning. The obvious next step is to commit to a *specific* theory of content. CFA generalizes over different possible definitions of *tracking*, and an actual theory of content for neural representations must commit to one such definition. Once we have a theory of content, along the lines of CFA, we can compare it to other accounts with the aim of proving that it truly captures what neural representations *are*. But while that still remains to be settled, I do hope this paper successfully illustrates the promise of taking the Cummins Functions Approach towards neural representations.

References

Bechtel, W. (2016). Investigating neural representations: the tale of place cells. *Synthese*, 193:1287–1321.

Bechtel, W., & Shagrir, O. (2015). The Non-Redundant Contributions of Marr's Three Levels of Analysis for Explaining Information-Processing Mechanisms. *Topics in Cognitive Science*, *7*(2), 312-322.

Chomsky, N. (1995). Language and nature. *Mind*, *104*(413), 1-61.

Craver, C. (2001). Role functions, mechanisms, and hierarchy. *Philosophy of Science*, *68*, 53–74.

Cummins, R. (1975). Functional analysis. *The Journal of Philosophy*, *72*(20), 741–765.

Cummins, R. (1989) *Meaning and Mental Representation*. Cambridge, MA: MIT Press.

Cummins, R. (1996). *Representations, targets, and attitudes*. Cambridge: MIT press.

Cummins, R., & Roth, M. (2010). Traits have not evolved to function the way they do because of a past advantage. *Contemporary Debates in Philosophy of Biology, Oxford, Reino Unido, Wiley/Blackwell*, 72-88.

Desimone, R., Schein, S. J., Moran, J., & Ungerleider, L. G. (1985). Contour, color and shape analysis beyond the striate cortex. *Vision research*, *25*(3), 441-452.

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

Dretske, F. (1988). *Explaining Behavior: Reasons in a World of Causes*. Cambridge, MA: MIT Press.

Dretske, F. (1995). *Naturalizing the Mind*. Cambridge, MA.: MIT Press.

Egan, F. (2012). Representationalism. In E. Margolis, R. Samuels, and S. Stich (Eds.), *The Oxford Handbook of Philosophy of Cognitive Science* (pp.250-272). Oxford University Press.

Egan, F. (2014). How to Think about Mental Content, *Philosophical Studies 170*(1), 115-135.

Egan, F. (2018). The Nature and Function of Content in Computational Models, in *The Routledge Handbook of the Computational Mind, M.* Sprevak and M. Colombo (eds.), Routledge (2018), 247-258.

Egan, F. (2020). A deflationary account of mental representation. In J. Smortchkova, K. Dolega, & T. Schlicht (Eds.), What are mental representations? (pp. 26-53) New York: Oxford University Press.

Garg, A. K., Li, P., Rashid, M. S., & Callaway, E. M. (2019). Color and orientation are jointly coded and spatially organized in primate primary visual cortex. *Science*, *364*(6447), 1275-1279.

Garson, J. (2019). There Are No Ahistorical Theories of Function. *Philosophy of Science*, *86*(5), 1146-1156.

Gawne, T. J. (2000). The simultaneous coding of orientation and contrast in the responses of V1 complex cells. *Experimental brain research*, *133*(3), 293-302.

Godfrey-Smith, P. (1993). Functions: Consensus without unity. *Pacific Philosophical Quarterly*, *74*(3), 196–208.

Grice, H. P. (1957). Meaning. *The philosophical review*, *66*(3), 377-388.

Hardcastle, V. G. (2002). "On the Normativity of Functions." In Functions: New Essays in the Philosophy of Psychology and Biology, ed. A. Ariew, R. Cummins, and M. Perlman, 144–156. Oxford: Oxford University Press.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425-2430.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, *160*(1), 106-154.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, *195*(1), 215-243.

Millikan, R. (1984). *Language, Thought and other Biological Categories*. Cambridge, MA: MIT Press.

Millikan, R. (1989). Biosemantics. *Journal of Philosophy*, 86: 281–97.

Millikan, R. (2004). *Varieties of Meaning*. Cambridge, Mass: MIT Press.

Morgan, A., & Piccinini, G. (2017). Towards a Cognitive Neuroscience of Intentionality. *Minds and Machines*, 1-21.

Neander, K. (1995). Misrepresenting & malfunctioning. *Philosophical Studies*, *79*(2), 109-141.

Neander, K. (2017a). *A mark of the mental: In defense of informational teleosemantics*. MIT Press.

Neander, K. (2017b). Functional analysis and the species design. *Synthese*, *194*(4), 1147-1168.

Pinel, P., Piazza, M., Le Bihan, D., & Dehaene, S. (2004). Distributed and overlapping cerebral representations of number, size, and luminance during comparative judgments. *Neuron*, *41*(6), 983-993.

Reich, D. S., Mechler, F., & Victor, J. D. (2001). Temporal coding of contrast in primary visual cortex: when, what, and why. *Journal of neurophysiology*, *85*(3), 1039-1050.

Robinson, D. A. (1989). Integrating with neurons. *Annual Review of Neuroscience*, 12, 33-45.

Shagrir, O. (2018). The brain as an input–output model of the world. *Minds and Machines*, *28*(1), 53-75.

Shea, N. (2018). *Representation in Cognitive Science*. Oxford University Press.

Sprevak, M. (2013). Fictionalism about neural representations. *The Monist*, *96*(4), 539-560.

Stampe, D. (1977). Towards a causal theory of linguistic representation. *Midwest Studies in Philosophy*, 2, 42-63.

Thomson, E., & Piccinini, G. (2018). Neural representations observed. *Minds and Machines*, *28*(1), 191-235.

Tolhurst, D. J., Movshon, J. A., & Thompson, I. D. (1981). The dependence of response amplitude and variance of cat visual cortical neurons on stimulus contrast. *Experimental brain research*, *41*(3), 414-419.