The Problem with Appealing to History in Defining Neural Representations

--Final draft. Forthcoming in *European Journal for Philosophy of Science*--

Ori Hacohen

Hebrew University of Jerusalem
Jerusalem, Israel
ori.hacohen@mail.huji.ac.il

Abstract

Representations seem to play a major role in many neuroscientific explanations. Philosophers have long attempted to properly define what it means for a neural state to be a representation of a specific content. Teleosemantic theories of content which characterize representations, in part, by appealing to a historical notion of function, are often regarded as our best path towards an account of neural representations. This paper points to the anti-representationalist consequences of these accounts. I argue that assuming such teleosemantic views will deprive representations of their explanatory role in computational explanations. My argument rests on the claim that many explanations in cognitive neuroscience are *entirely* independent of any historical considerations. In making this claim, I will also offer an adapted version of the famous Swampperson thought experiment, which is better suited to discussions of subpersonal neural representations.

Keywords

neuroscientific explanations; neural representations; theories of content; teleosemantics; swampperson

1. Introduction

For many explanations in cognitive neuroscience, *representations* seem to play an important role. In explaining a cognitive phenomenon, scientists regularly refer to internal neural states or structures as vehicles of content or carriers of information. Often they will also explicitly treat these internal components as "representations" of some distal properties. This apparent prevalence of representations in neuroscience is treated by many philosophers as evidence for the *representationalist* view of cognition- the view that "postulating representational (or 'intentional' or 'semantic') states is essential to the theory of cognition" (Fodor & Pylyshin 1988, p.7).

Still, some have argued that despite appearances, there actually *isn't* a necessary explanatory role for representations in many neuroscientific explanations. Perhaps the most influential such argument comes from Ramsey (2007). Ramsey argues that many neuroscientific explanations which seem as if they are invoking representations, or 'information-carrying' states, are in fact *not* appealing to representations in any explanatorily relevant sense. Thus, "a wide range of theories that claim to be representational in character, like many theories in the computational neurosciences, actually aren't." (Ramsey 2007, p. 223) Naturally, this type of view is consistent with an *eliminativist* approach, which holds that representations have no place in explaining cognitive phenomena (e.g. Chemero 2009, Hutto & Myin 2013). Ramsey himself concludes that "cognitive science is secretly (and non-deliberately) moving in a direction that is abandoning representationalism." (Ramsey 2007, p. 188) Representationalists have replied to such claims in a variety of ways (e.g., Morgan 2014, Shagrir 2012, Sprevak 2011).

Amongst representationalists, there is also a *different* debate regarding how to best understand the relevant notion of representation. If 'neural representations' are essential to cognitive neuroscience, then there must be a *theory of content* that can properly define what it means for a certain neural structure or state to be a representation of some specific distal content. There is no single agreed upon theory of content that gets this job done, but there does seem to be some broad agreement among many philosophers that we should account for neural representations by appealing to some historical notion of function. Such theories, which are usually identified as *teleosemantic* theories of content, have largely become the mainstream view of neural representations. In particular, Neander (2017) and Shea (2018) have recently proposed (competing) theories of content for the type of subpersonal neural representations that are invoked in cognitive neuroscience, and both offer a teleosemantic account that is partly defined by history.

Philosophers usually appeal to *evolutionary* history in defining the function of a representation (e.g. Millikan 1984, 1989, 2004, Neander 1995, 2017), or to *developmental* history (Dretske 1988, 1995). Shea (2018) has also added the possibility that the relevant notion of function can be defined in virtue of "contribution to an organism's persistence", which is determined by merely a brief pattern of successful behavior. But this brief pattern of behavior is still a (brief) historical process which *had to have occurred* in order to define the relevant function. In this paper, I will argue against any theory of content that defines neural representations by appealing to the occurrence of *any type* of meaningful historical process (be it evolutionary, developmental, or other). Since, as far as I can tell, all the

relevant theories that appeal to history are teleosemantic, I will often regard this as an argument against historical of forms of teleosemantics.[1]

My argument will relate the two different debates that were mentioned thus far. The first centered on the role of representations in neuroscience and whether or not computational explanations are truly representational. The second debate revolves around which theory can account for neural representations and their contents. I will claim that those that opt for a historical account of representations with regards to the second debate, will necessarily find themselves siding with the eliminativists with regards to the first one. I argue that a historically defined notion of representation cannot possibly have an explanatory role in the computational, information-processing, explanations we see in cognitive neuroscience. Thus, to paraphrase Ramsey, proponents of historical forms of teleosemantics are secretly (and non-deliberately) moving in a direction that is abandoning representationalism.

2. Framework

My argument rests on the claim that many explanations in neuroscience, *including* the computational information-processing accounts that seemingly invoke the use of neural representations, are entirely independent of *any* historical considerations. That is to say that there is no historical process, whose occurrence (or

[1] I will sometimes use the term 'historical teleosemantics' to refer to such theories, and there are a few cases where I talk of 'teleosemantics', which should also be understood as regarding only historical accounts. In general, most teleosemantic theories, and certainly the most prominent ones, do appeal to history, but there are also non-historical teleosemantic accounts (e.g. Nanay 2010, 2014).

non-occurrence) is in *any* way essential to these neuroscientific explanations. I will spend the bulk of this paper (namely sections 3,4, and 5) defending this claim.[2]

It is important to stress that this is *not* a claim about representations at all, and accordingly I will not be assuming *anything* regarding the nature of neural representations or their explanatory role. It is only a claim about neuroscientific explanations. And I intend to make this claim by focusing on one paradigmatic example of a neuroscientific explanation- that of the vestibulo-ocular reflex (VOR), which I introduce in section 3. I wish to show that the occurrence or non-occurrence of *any type* of meaningful historical process is irrelevant to the explanation of VOR. Thus, in section 4, we will consider this explanation within a scenario which *lacks* any meaningful history. This will amount to a new version of the famous Swampperson thought experiment.

It is common practice to appeal to swampperson thought experiments when arguing against historical notions of representation.[3] The term 'Swampman' was originally coined by Davidson (1987), when he imagined a molecule-for-molecule physical replica of himself appearing miraculously in a swamp, without any process of design, and lacking any history. The classic swampperson argument against

[2] Perhaps one might wonder whether a simple distinction between mechanistic and etiological explanations will be good enough to make this claim (Craver 2007, 2013). There are a couple of reasons why I do not take this route. First, it will demand some assumptions on the nature of mechanistic explanations and some commitment to the characterization of the relevant neuroscientific explanations as mechanistic. But more importantly, I do not find it to be trivially true that all mechanistic explanations are *entirely* independent of *any* historical considerations. For example, it is possible that Shea's objection, discussed in section 5, can be construed as the claim that *some* mechanistic explanations (in particular, those that invoke representations) are indeed dependent on history. The line of argument in this paper is meant to deal with such objections and enable the strongest possible claim regarding the irrelevance of history, without assuming any characterization of neuroscientific explanations.

[3] See short discussion in (Neander 2012), section 4.2.

teleosemantics is based on the intuition that the swampperson must have *some* intentional states (i.e. beliefs, desires) if he shares precisely the same physical states as Davidson. But according to teleosemantics, since a swampperson lacks any history he cannot possess intentional states or structures. Therefore, teleosemantics must be wrong. Or so the argument goes. Proponents of teleosemantics have responded to such claims in various ways (e.g. Millikan 1996, Neander 1996, Dretske 1996, Papineau 2001)[4]. And, perhaps more importantly, it is doubtful whether this classic argument can apply to the type of subpersonal non-conscious neural representations that this paper focuses on, which are arguably impossible to intuitively characterize.

The Swampperson argument in section 4 is significantly different from the classic version, mostly because it does not deal with representations at all. As stated above, this is an argument about neuroscientific explanations. In fact, it is an argument about one specific neuroscientific explanation- the explanation of VOR, which, I will claim, can be applied in a history-lacking swampperson scenario. This argument is not based on our understanding and intuitions about the nature of intentional kinds, but rather on our understanding and intuitions about the nature of neuroscientific explanations. Shea (2018) considers a similar Swampperson argument and brings up an important possible objection. Section 5 will therefore include a lengthy consideration, and rebuttal, of Shea's claims. Taken together, sections 3,4, and 5 should ultimately establish that history is indeed irrelevant to the explanation of VOR.

[4] It is also worth noting that Porter (2020) has recently offered a different type of Swampperson argument against teleosemantics, which aims to show the existence of real-world cases of Swampperson-like representations. I thank an anonymous reviewer for bringing this to light.

Section 6 then shows that assuming a historical notion of representation means that representations are also irrelevant to the explanation of VOR. Thus, proponents of such historical accounts must conclude that the explanation of VOR is *not* representational. And that is despite the fact that it seems like a classic example of a computational explanation given in information-processing terms. I claim that this argument generalizes to a significant class of neuroscientific explanations, illuminating the anti-representational consequences of accepting historical views of representation. Section 7 concludes the paper and looks at the options we have moving forward.

3. The Explanation of VOR

Our eyes have the ability to maintain a fixed gaze. This is a feature of all types of eye movement- the eyes hold a steady gaze at the end of saccade, they maintain a steady gaze during smooth pursuit, and they can maintain a fixed gaze while the head in which they are situated moves. This last case is known as the Vestibulo-Ocular Reflex, or VOR. The phenomenon of VOR has been studied for a long time (cf. Magnus 1924, Lorente de Nó 1931, 1933), with significant success. Studies have pointed to 'the three-neuron arc', composed of primary vestibular neurons, secondary vestibular neurons, and oculomotor neurons, as the major physical pathway that enables this phenomenon. D. A. Robinson (1968, 1970, 1989) has highlighted the

importance of a second pathway that is essential for our gaze holding ability. Figure 1 features an explanation of the VOR phenomenon, adapted from (Robinson 1989)[5].

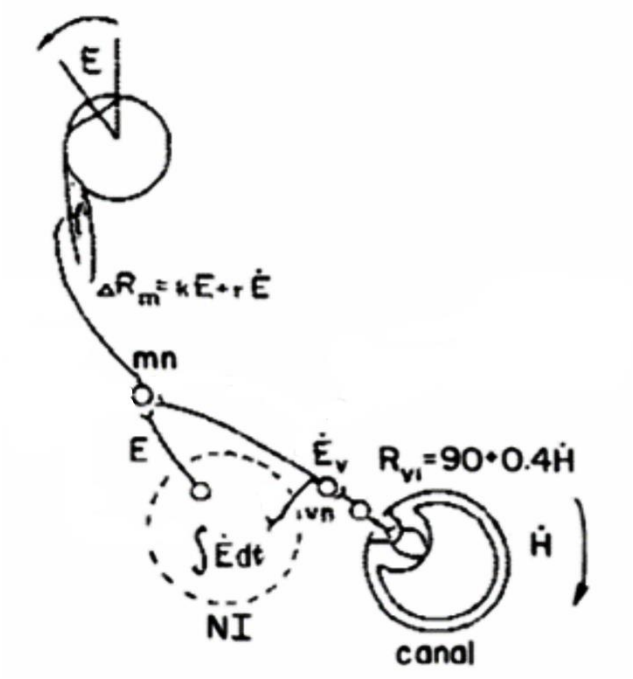

Figure 1. The explanation of VOR.
The semi-circular canals in the ear (labeled "canal") produce a 'head velocity' signal, $\dot{H}$. This signal is transferred to the vestibular nucleus (vn) by the primary vestibular afferents, whose discharge rate (R_{V1}) correlates with head velocity ($R_{V1} = 90 + 0.4\dot{H}$). From the vestibular nucleus, the signal becomes an eye velocity ($\dot{E}$) command and continues in two pathways. The first is the direct pathway to the ocular motoneurons (mn). This is the well-established three neuron arc. A second pathway passes through another set of neurons in the brainstem called the Neural Integrator (NI), since they convert the eye *velocity* signal to an eye *position* signal (E), which is then also sent to the ocular motoneurons (mn). Thus, the motoneurons, modulating by ΔR_m, get both the eye velocity signal (via the direct pathway) and the eye position signal (via the neural integrator) that are necessary for the vestibulo-ocular reflex.

[5] This explanation has already been featured in the philosophical literature numerous times (e.g. Bechtel & Shagrir 2015, Shagrir 2018, Hacohen *forthcoming*), in part thanks to its simplicity.

The explanation in figure 1 accounts for the phenomenon of VOR by describing an internal *information-processing* mechanism.[6] As stated, this explanation is taken from (Robinson 1989), and is, of course, in line with Robinson's original phrasing:

> "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)

We see that this explanation regards internal components of the brain, like the ocular motoneurons or neurons within the vestibular nuclei, as vehicles of content or information, encoding signals of eye position and velocity. Representationalists want to treat such explanations as evidence for the significance of representations in neuroscience. And what they need is a theory of content that can properly define what makes these internal components actual *representations* of their specific contents. At the same time, there are various eliminativist or instrumentalist views which might claim that "content carrying" states of the type described in this explanation *aren't* actually representational (Ramsey 2007)[7], or that the appeal to representational

[6] Notably, this (Robinson 1989) explanation of VOR is consistent with the contemporary textbook explanation (see Leigh & Zee 2015, chapters 3 and 6).

[7] Such states are consistent with Ramsey's (2007, chapter 4) 'receptor notion' which he claims does not do any explanatory work *as a representation*.

vocabulary is merely an informal presentation (Chomsky 1995)[8]. But as stated in section 2, none of that concerns us at the moment. We will not debate the representational status of the internal components that are described in the explanation of VOR. We are currently only interested in one question- is the explanation of VOR dependent on the occurrence of *any* historical process in *any* way?

At first glance, I think it does seem rather obvious that the explanation of VOR is not dependent on history. As it is described above, this explanation is a description of the physical mechanism that enables the phenomenon of VOR. It tells us what *is happening* in the brain when this phenomenon occurs. This is a description of a synchronous process, and not a historical one. Relatedly, when looking at the explicit scientific reasoning behind the use of representational vocabulary in the explanation of VOR, a similar conclusion seems to arise. Scientists do not motivate their use of such intentional terms (e.g. 'coding', 'signal', 'representing') by appealing to any type of historical process. Instead, they mostly point to empirical studies that show correlations between internal components and distal properties. Other times they might also claim that the task performed by the brain necessitates the existence of specific internal signals. (For a nice example of these reasons as given in a neuroanatomy textbook see (Leigh & Zee 2015, pp. 362-363).) Either way these types of reasons don't seem to demand the occurrence of any meaningful historical process. The correlations that scientists measure capture causal relations that occur as the phenomenon of VOR takes place. And if the task itself necessitates the use of internal

[8] As Chomsky (1995, p. 55) says of Marr's (1982) theory of vision: "The theory itself has no place for the [intentional] concepts that enter into the informal presentation, intended for general motivation."

representations, then that is clearly dependent on the state of affairs *as the task is being performed* and not on anything that might have happened in the past.

It thus seems that scientific practice confirms our initial judgement that the explanation of VOR is not dependent on the occurrence of any prior process. But this type of discussion, though not insignificant, would never be enough to convince a proponent of teleosemantics. And rightly so. History can still end up being essential to the explanation of VOR, regardless of its initial appearance and any explicit scientific reasoning. For example, even if scientists never mention natural selection in relation to the explanation of VOR, they might still implicitly assume it applies to all relevant subjects, and that assumption might end up being essential to the explanation. To truly show that the occurrence of a process of evolution is *not* essential to this explanation, we have to see if the explanation can hold in a scenario where evolution is absent. To show that the same is true for a process of learning, and that in fact- there isn't *any* possible historical process that is somehow necessary, we need to see if the explanation holds in a scenario that rules out *any* meaningful historical process. Which is why Swampperson must reenter the picture.

4. Swampperson 2.0

Suppose that, once again, Swampperson is a molecule-for-molecule physical duplicate of Donald Davidson, created by some accidental clashing of particles in a swamp. The first claim I wish to make is that Swampperson will exhibit the phenomenon of VOR. That means that Swampperson, like Davidson, will be able to maintain a steady eye gaze during head movement. Now, perhaps some would immediately state that Swampperson doesn't even have eyes or a head in the same

sense that Davidson does, only replica-eyes and a replica-head. I do not oppose this distinction. Obviously Swampperson is not a real-world person but a replica of one, and the same would therefore be true for all of Swampperson's internal components. But referring to Swampperson's internal components as replicas will have no effect on the argument I wish to make. I will be arguing that the explanation of VOR can apply in Swampperson scenarios, precisely as it applies in the real world. In other words, I want to show that the explanation of VOR works just the same even when we switch out all the real-world physical entities for their Swampperson counterparts. Insisting on calling these counterparts "replicas" is, while justified, irrelevant to this claim. If the *only* difference between the explanation of VOR in the real world and the explanation of VOR in a Swampperson scenario, is that the latter demands reference to "replica-eyes", "replica-neurons", and so forth, then that just goes to show that the distinction between these "replica-entities" and the real-world entities is irrelevant to the explanation of VOR. Which is precisely what I'm getting at. Hence, I think we can allow talk of the Swampperson's eyes, head, and other physical components, assuming it is clear that these are only physical replicas of real-world entities.

So again, I claim that Swampperson will be able to maintain steady eye gaze during head movement. An anonymous reviewer has noted that some philosophers might reject this claim, if the *ability* to maintain steady gaze is understood as carrying some (teleological) functional commitment. Hence let me clarify that I do not intend to imply anything of this sort. Swampperson's "ability" to maintain steady gaze is simply *something it can do*. I figure it is safe to assume that there will be *some* things that Swampperson will be able to do, and others that he won't. Swampperson will be able to raise his hands over his head, but he won't be able touch his nose to his back. Swampperson's heart will be able to make thumping noises. I assume we can agree on

such statements. I also assume we can agree that besides making thumping noises, Swampperson's heart will also be able to circulate blood. It might not have the *function* to circulate blood, according to etiological theories of function, but it is still *something it can do*. Hence, I hope it is relatively uncontroversial to claim that Swampperson will also be able to maintain eye gaze during head movement, and exhibit the phenomenon of VOR.[9]

Now let's imagine two scenarios. In scenario A (the "real-world scenario"), a scientist meets Davidson, and notices the fact that he can maintain a steady gaze while his head moves. We ask the scientist to explain this phenomenon and she gives us the explanation of VOR above. In scenario B (the "swampperson scenario"), the scientist meets Swampperson, and notices the fact that he can maintain a steady gaze while his head moves. We ask the scientist to explain this phenomenon and she gives us the explanation of VOR above. The question now is- is the explanation in scenario B somehow worse than the explanation in scenario A?

I believe we should maintain that the explanation of VOR is equally successful at accounting for the phenomenon in both scenarios. In both cases the explanation will enable the same predictions and the same understanding of the phenomenon and the underlying physical mechanism that enables it. There is therefore no sense in which the explanation in scenario A is essentially better or more suitable than the explanation in scenario B.

But let's expand the thought experiment a little further. Suppose that in scenario B, the scientist initially assumed she was dealing with an actual human when

[9] In the next section I consider an objection from Shea (2018) that claims that the phenomenon of VOR, understood in this manner, is not enough to define the relevant explanandum.

she proposed the explanation of VOR. We then tell the scientist that she *actually* met Swampperson - an *exact* physical replica of Davidson - that, like Davidson, has the ability to maintain a fixed eye gaze while his head moves. We ask the scientist again to explain this phenomenon exhibited by Swampperson. Should she change her explanation in any significant way?

To see why I think the scientist should *not* significantly change her explanation, it is perhaps useful to contrast this with one last imaginary scenario. Consider scenario C, in which the scientist again meets what she believes is an actual human and offers the explanation of VOR. But this time, after she offers the explanation, we tell her that she *actually* met Robotperson- a robot that was designed to look and act precisely like Davidson. No further details are given of Robotperson's internal design and how he achieves this perfect imitation of Davidson. Yet Robotperson is able to maintain a fixed eye gaze while his head moves. We now ask the scientist again to explain this phenomenon exhibited by Robotperson. Should she significantly change her explanation?

In this scenario C, the obvious answer is yes. While the scientist might still maintain that her account could provide a how-*possibly* explanation of the phenomenon exhibited by Robotperson, she would never claim that it provides a how-*actually* explanation of this phenomenon. This is because she has no knowledge of Robotperson's inner workings, and how he *actually* achieves steady gaze. This should starkly contrast with scenario B, where despite learning that she is not really dealing with Davidson, the scientist can remain confident of her understanding of the actual phenomenon at hand. Since Swampperson is physically identical to Davidson, the scientist still understands the mechanism that enables VOR in scenario B. Her

explanation can remain essentially the same, as it describes this same mechanism. This reinforces my claim that the explanation of VOR can apply to Swampperson.

It is important to emphasize again that this thought experiment does not presuppose *any* view of representations. The applicability or success of neuroscientific explanations is constantly evaluated, never waiting for a resolution to the problem of content. We can, and we *do*, assert that the explanation of VOR is a good explanation *independently* of any specific theory of content. Thanks to empirical testing, that has validated various predictions, we accept that this explanation points to the true physical mechanism that underlies the phenomenon of VOR, and we assert that it offers sufficient, and correct, understanding of the phenomenon. I claim that for precisely these same reasons we should assert that it is a good explanation in the swampperson scenarios.

Thus, given that the explanation of VOR applies in swampperson scenarios just as it does in the real world, we can conclude that it cannot possibly be dependent on the occurrence of *any* meaningful historical process. It is essentially independent of any historical considerations.

There is however one last point worth stressing. When I claim that the explanation of VOR is independent of any historical considerations, I am referring to a *specific* explanation of VOR – described in section 3 in accordance with (Robinson 1989). I am *not* claiming that history cannot have any role in explaining the phenomenon of VOR.[10] The conclusion we draw from the swampperson argument is

[10] In general, it is highly likely that considering how something came to be, can help us understand how it currently functions. Miłkowski (2016) discusses the ways history can figure into explanations of How-questions. I agree that history can play a role in explaining how-questions, but I think that in many existing computational explanations- it doesn't. And, as I

not about the *phenomenon* of VOR, or about VOR as an *explanandum*. Rather it is about an actual neuroscientific explanation. What we saw is that history has no role in Robinson's (1989) explanation, which was described in figure 1. But now let's consider an important objection to this claim.

5. Shea's 'No Explanandum' Argument

Though there has been much discussion of Swampperson in the teleosemantic literature, it has largely been focused on the *classic* Swampperson argument (described in section 2) and as such seems less relevant to the current argument.[11] One significant exception can be found in Shea's (2018) recent book:

> "At first pass, representational explanation does not seem to depend on evolutionary history at all. By recognizing that behaviour was driven by a representation of the location of an object, say, it seems that the cognitive scientist is picking out a synchronic property of the organism. It also seems that the representational vehicle is a synchronic cause of the behaviour. How internal processing unfolds, and hence how the organism will make bodily movements, is caused moment-by-moment by the intrinsic properties of

argue later on, the extent and significance of these non-historical explanations, should give pause to any proponent of historical teleosemantics.

[11] For example, Millikan (1996) makes the claim that Swampperson would not share the same 'real' kind as other humans (besides Davidson), and therefore we cannot deduce anything about human intentional states by thinking about Swampperson. But the current swampperson argument *doesn't* deduce anything about human intentional states. We only deduce something about the explanation of VOR. And even if Swampperson doesn't share a real kind with humans, we can still claim that the explanation of VOR applies to Swampperson, since the same explanation can apply to different kinds. Shea (2018, p. 170) also offers some criticism of Millikan's reply.

representational vehicles (e.g. activity of neurons). It follows that, if we take an organism that has evolved by natural selection and had a lifetime of interacting with its environment, make an intrinsic duplicate, and place the duplicate in just the same environment, we will be able to make the same predictions about how it will behave." (Shea 2018, pp. 21-22)

Shea is therefore well aware that many explanations in neuroscience do not seem to be dependent on history in any way, and he understands the problem this could cause for historical theories of representation. But he is quick to dispose of this problem with what he refers to as the 'no explanandum' argument:

"We point to representations to explain how organisms and other systems manage to interact with their environment in useful and intelligent ways. The explanandum is a pattern of successful behaviour of a system in its environment. That explanandum is absent at the moment swampman is created. It's not just that swampman has not yet performed any behaviour. (He already has dispositions to behave in certain ways.) It is that it's quite unclear that some behaviours should count as successful and others not. So the creature with no history has no contents but that is fine because it has nothing which contents are called on to explain." (ibid.)

Let us adapt this argument to our current discussion. Shea might accept that the explanation of VOR could account for cases where a swampperson maintains a steady gaze as his head moves. Shea might *also* accept that this explanation can account for cases where a swampperson does *not* maintain a steady gaze (for some

reason).[12] The problem, according to Shea, is that in swampperson scenarios, we have no reason to consider the former cases (where a steady gaze is maintained) as a "success" and the latter as a "failure". This distinction, Shea claims, is an essential part of the relevant explanandum, and it can only be accounted for by some *historical* "pattern of successful behaviour". Shea goes on to define his notion of "success" as determined by "natural selection, learning, or contribution to persistence" (p. 62), all of which are absent in any swampperson scenario (at least initially)[13].

I wish to reject this argument by claiming that Shea's notion of "success" is in fact *not* a part of the relevant explanandum, and that the explanandum that *is* relevant to the explanation of VOR exists in swampperson scenarios to the same extent it exists in normal real-world scenarios.

The explanandum is the phenomenon to be explained. For our current discussion the relevant phenomenon, or so it seems, is the vestibulo-ocular reflex, i.e. the phenomenon of maintaining a steady gaze while the head moves. This phenomenon *would* appear in swampperson scenarios, even before any pattern of "successful" behavior can emerge. Shea himself would agree that even at the moment of creation, Swampperson's eyes would have the disposition to robustly maintain a

[12] Such a scenario was not considered in section 4, but it should not be difficult to imagine. For example, suppose someone were placed in a lab with electrodes that artificially alter the discharge rate of her primary vestibular afferents. The explanation of VOR that was given in section 3 will allow us to successfully predict that this person would not be able to maintain her steady gaze. We can also explain why not. Now, the same would be true if we imagine the physical duplicate of this person in an identical scenario. Just as before, we would be able to predict that this swampperson replica would not maintain steady gaze, and we would offer exactly the same explanation as to why not.

[13] After Swampperson has had some time to interact with its environment, Shea does think a "successful" pattern of behavior could be established, which in turn would allow Swampperson to have representational states.

steady gaze as the head moves. If Swampperson only exists for 10 seconds, and exhibits VOR only a couple of times in that span, that is *still* a phenomenon that can be explained by the explanation of VOR that we have been discussing.

And, just in case anyone wants to claim that exhibiting only a couple instances of VOR does not amount to a robust phenomenon, we can also modify our thought experiment to imagine a whole population of swamppersons. Suppose there is a swamp where thousands of physical duplicates of humans are miraculously created, but all of them disappear within 10 seconds. Yet, for those 10 seconds, they still all exhibit the ability to maintain a steady gaze as their head moves. That is, quite clearly I believe, a phenomenon that can be explained. And it *is* explained, by the explanation of VOR we saw in section 3. And yet, Shea argues that this type of phenomenon, which clearly *can* be observed in swampperson scenarios, is not enough to define the relevant explanandum. He claims that in the real world, it is not merely the phenomenon, but rather the phenomenon *as a success* which is being explained.

I do of course agree with Shea that there is a sense in which the phenomenon of VOR is a "success" only in the real world, and not in the swampperson scenarios discussed above. Clearly, in the real world, the phenomenon of VOR does contribute to organisms' persistence, while in the swampperson examples it does not. We can also assume that this trait (exhibiting VOR) evolved through natural selection. Let us therefore accept it as fact that, in the real world, the phenomenon of VOR is indeed a "success" in some historical sense. What I want to claim is that this fact about the phenomenon of VOR is not a part of the explanandum. It is irrelevant to what is actually *being explained*.

There are many facts about the phenomenon of VOR which are not explained by Robinson's (1989) explanation, as discussed in section 3. There is, for example, some fact of the matter regarding the amount of heat that these neurons emit in enabling VOR, or the amount of energy that this reflex demands. It is also a fact that exhibiting VOR during free-head pursuit (where we follow a moving visual target with both the eyes and the head) is actually *detrimental* to the organism.[14] But none of these facts about the phenomenon of VOR are part of what is actually *being explained* by Robinson (1989) as described in section 3. If we lived in a world where the phenomenon of VOR emitted a significantly different amount of heat, it wouldn't make any difference to Robinson's (1989) explanation. Nor would it matter if we lived in a world where, for some reason, exhibiting VOR during free-head pursuit is actually useful. Such considerations are entirely irrelevant to the explanation of VOR which we have been discussing. The only thing Robinson (1989) explains is how the brain can maintain a fixed eye-gaze while the head moves. The amount of heat this process generates, or whether or not this process is useful to the organism during free-head pursuit, is *not* relevant to the actual explanandum.

In much the same way, Shea's notion of "success" is also not relevant to the actual explanandum. As stated, it is a fact that the ability to exhibit VOR can contribute to organisms' persistence (and might therefore be regarded as a "success"), but that fact is just as irrelevant to Robinson's (1989) explanation as were the facts in the previous paragraph. Again, what is actually *being explained* in this particular explanation is the brain's ability to maintain a fixed gaze during head movements.

[14] That is why this reflex is normally suppressed during free-head pursuit (Cullen 2012, Ackerly & Barnes 2011).

And that is *regardless* of whether or not this process contributes to the organism's persistence, or is the result of natural selection.

This line of thought doesn't change when we also consider Shea's characterization of cognitive explananda in general. It's certainly logical that scientists focus on the phenomena which has contributed to the well-being of the organism in one way or another. And so, just as it can be accepted as a fact about the explanation of VOR, perhaps it could also be accepted as a general fact about cognitive explananda- that the phenomena that is being explained is indeed a "successful" pattern in some historical sense. And still I would insist that this fact about cognitive explananda (or, to be more precise, this fact about the phenomena which constitute cognitive explananda), isn't *itself* a part of the explananda; it is not what is actually *being explained* by real neuroscientific explanations. As illustrated in our discussion of VOR, the actual explanatory work amounts to an account of an actual robust phenomenon. Whether or not this phenomenon is "successful" (because it contributes to persistence, or because it resulted from natural selection) goes beyond the scope of explanation.

But now let's focus a little more on normative vocabulary. For, as some might point out, scientists regularly use normative terms in describing explanations of this kind. Scientists will talk of "successfully achieving" VOR, and cases where VOR is not exhibited will often be regarded as "failures" or "deficiencies". This seems to point towards a normative aspect of this explanation and, some might claim, provide evidence to Shea's claim that the explanandum is indeed defined as a "success". Before I reply, allow me to point out that even if the explanandum *is* treated as a "success" that doesn't mean we must understand this notion of "success" historically, as Shea does. In general, evidence towards the significance of normative distinctions

should *not* be automatically regarded as evidence towards the significance of history. Proponents of historical theories of representation, or historical theories of function,[15] have often argued that only historical considerations can allow for normative distinctions (e.g. Neander 1991, 1996, Dretske 1996, Garson 2019). But there are also many alternative views of representation (e.g. Fodor 1987, 1990, Cummins 1989, Egan 2014, 2018) or function (e.g. Cummins & Roth 2010, Craver 2001, Davies 2001, Hardcastle 2002, Bigelow & Pargetter 1987, Maley & Piccinini 2017) which claim to account for normativity *without* appealing to history.

And yet, that is not the point I wish to focus on here. In fact, let us accept that at least some use of normative terms *does* imply a commitment to Shea's historical notion of "success". Suppose, as Shea claims, that in the real-world scientists can regard cases where VOR isn't exhibited as a "failure", while in the history-lacking "swamp-world" they can't. Does this difference mean that the explanation of VOR cannot be applied in swampperson scenarios in the same manner it is applied in the real world? Consider such a case where, for some identical reason, a person in the real-world and their physical duplicate in the swamp-world, do *not* maintain a fixed gaze while their head moves (see footnote 12). As stated before, Shea could probably agree that the explanation of VOR can account for such a case in both scenarios- it allows for the same predictions and the same understanding of the underlying mechanism. The *only* difference, therefore, is that in the real world the scientists can say that the person *failed* to exhibit VOR, while in the swamp-world, they cannot. Instead they can say that the swampperson *didn't* exhibit VOR.

[15] Historical notions of function have also had to face swampperson type objections (e.g. Boorse 1976).

So, again, should this difference, on its own, be taken as evidence for the significance of history in the explanation of VOR? It is clear, I think, that the word "failed" doesn't do any actual explanatory work here. It only implies the assumption that, in the real world, the phenomenon of VOR can be regarded as a "success" in some historical sense. We have already accepted this assumption as true. It *is* an actual difference between the real world and the "swamp-world" that only in the former can the phenomenon of VOR be considered a "success" (in Shea's sense). But as stated above, this actual difference is not explanatorily relevant to Robinson's (1989) explanation. And that remains true even if, in the real world, scientists sometimes use vocabulary which refers to VOR as a "success" in Shea's sense. Since VOR *is* a success, scientists can choose to regard as a such. That does *not* mean that this is part of the explanandum.

There are obvious (and less obvious) differences between the real world and the "swamp-world", which *necessarily* affect the vocabulary we can use in the different scenarios. To give the simplest example, in the real-world scientists can talk of an *evolved* mechanism that enables VOR, while in the swamp-world they clearly cannot. But unless regarding the mechanism as "evolved" can be shown to be explanatorily relevant, then this possible difference in vocabulary doesn't mean much. Similarly, scientists can talk of a *failure* (in Shea's sense) to exhibit VOR only in the real world, and not in the swamp-world. But that is just another possible difference in vocabulary that is irrelevant to what is actually *being explained*.

To summarize the discussion in this section, I believe we have no reason to complicate the obvious. The explanation described in section 3, and taken from (Robinson 1989), aims to explain the phenomenon of VOR. That is, it aims to account for the robust phenomenon of maintaining steady eye gaze while the head moves.

Nothing more. This straightforward explanandum exists in swampperson scenarios precisely to the same extent it does in the real world. Thus, we can maintain the conclusions of section 4. The explanation of VOR applies in swampperson scenarios just as it would in the real world, and is therefore entirely independent of any historical considerations.

6. Consequences for Historical Accounts of Representation

We have just seen that history is irrelevant to the explanation of VOR. Let us now assume that some historical form of teleosemantics is correct, and that neural representations are indeed defined, in part, by the occurrence of some historical process. That would mean that the established irrelevance of history for the explanation of VOR also implies an irrelevance of neural representations.

For example, the explanation of VOR seems to regard the ocular motoneurons as carriers of a specific signal (encoding both eye velocity and head velocity). Now, according to historical teleosemantic theories of content, these ocular motoneurons are only neural *representations* if some historical process defined them as such. But it turns out that whether or not such a process actually occurred is irrelevant to the explanation of VOR. That is, regardless of whether or not this historical process (that defines the ocular motoneurons as representations) actually occurs, the explanation works just the same. Assuming teleosemantics, that means that regardless of whether or not the ocular motoneurons are representations (of eye and head velocity), the explanation works just the same. These motoneurons' status as *neural representations* turns out to be inconsequential to this explanation of VOR.

Just to be clear, the ocular motoneurons themselves would obviously still be essential to the explanation, but not *as representations*. One could still choose to insist that the ocular motoneurons *are* neural representations, but doing so would have no bearing on the explanation of VOR. As long as the representational status of these neurons is defined, in part, by some historical process, we must conclude that the question of whether or not these neurons are representations is entirely insignificant for this explanation. And the same would obviously also hold for any other potential representation in the explanation of VOR. Given the irrelevance of history, it would be impossible for historical versions of teleosemantics to maintain the claim that *any* representation is essential to the explanation of VOR. That means that proponents of such theories must accept that the explanation of VOR is *not* a representational explanation.

As mentioned before, the explanation of VOR certainly does *seem* like a representational explanation. It specifies an *information processing* mechanism, and regards internal brain components as carriers of signals/information/content. It is the type of explanation representationalists like to point to in arguing for the essential role of representations in cognitive neuroscience. And yet, proponents of historical accounts would have no choice but to accept that representations actually have no explanatory role in the explanation of VOR. As mentioned in section 3 (footnote 7), I presume that is precisely what Ramsey (2007) would say regarding the explanation of VOR (albeit for different reasons).

And of course, the explanation of VOR is not unique in this regard. Consider Marr's (1982) theory of vision, which has been "the principal battlefield" (Piccinini 2008, p. 208) for philosophical discussions of representations. Is Marr's explanation of how the brain filters the retinal image to detect possible edges dependent on history

in a manner in which the explanation of VOR was not? I believe we can apply the same swampperson argument that was discussed in section 4 to Marr's explanations and reach the same conclusion. Namely, that history is irrelevant. That would mean, for the same logic discussed above, that proponents of historical teleosemantic views would have to accept that Marr's explanations are in fact *not* representational.

And obviously, this issue is also not limited to any specific cognitive domain. Shadmehr and Wise (2005), for example, offer a computational theory of motor control, which has also been featured in the philosophical discourse (Shagrir 2006, Egan 2012). When they explain how the brain directs hand movement towards a target object, are they somehow assuming the previous occurrence of some type of historical process? In general, I think we can safely say that there is a wide class of computational, information-processing explanations in neuroscience that can be shown to be independent of history, just like the explanation of VOR. And for all such explanations, assuming a historical account of representation will entail the explanatory irrelevance of neural representations. Proponents of historical teleosemantics will need to accept that many computational explanations in neuroscience are in fact *not* representational explanations.

Accepting that many computational and information-processing explanations are not representational does not amount to an eliminativist view, but it's certainly consistent with eliminativist views, and it substantially deflates the representationalist stance. I do not know of any proponent of teleosemantics that has expressed a willingness to accept such consequences. Shea (2018) and Neander (2017), for example, explicitly intend for their theories of content to account for the type of "information-carrying" we find in these computational explanations, like the explanation of VOR. They also argue against Ramsey's (2007) claims that such

notions are not explanatorily significant. But what the argument in this paper shows is that there are really only two options for any representationalist. Either accept that representations are not as significant to neuroscience as they appear to be, or renounce historical accounts of representation.

7. Conclusion

Many computational explanations in neuroscience appeal to information processing mechanisms and content-carrying states. Such explanations are regularly regarded as *representational*, and often motivate a representationalist view of cognitive neuroscience. In this paper, I claimed that many of these explanations are entirely independent of any historical considerations, and thus cannot be regarded as representational by any proponent of historical teleosemantics. This was illustrated on Robinson's (1989) explanation of VOR, partly by appealing to an adapted version of the Swampperson thought experiment.

Where we go from here depends mostly on whether or not one is willing to accept that many computational explanations in cognitive neuroscience are *not* representational explanations. If we accept that, we are left with two options. The first is to follow the eliminativist approach and claim that cognitive neuroscience really isn't about representations at all. The second is to reject eliminativism and try to carve out a somewhat more modest role for representations in neuroscience. For example, neuroscientific "Why questions" (Tinbergen 1963) lend themselves quite naturally to historical analysis, so perhaps that would also be where historically defined representations eventually find their main explanatory role.

In my opinion though, we should choose neither of these paths, since we should *not* be willing to accept that the explanation of VOR and other computational explanations are not representational. I think a large part of this paper can be described as defending the naïve claim that 'the explanations that *seem* like they're independent of history, really *are* independent of history'. I tend to hold a similar naïve claim with regards to representations. I think the explanation of VOR, and others like it, which *seem* representational, really *are* representational. It's just that the notion of representation they appeal to, cannot be defined historically. There are other, non-historical, theories of representation that deserve consideration,[16] though each has its own challenges to deal with. I propose we start facing these challenges, and let go of historical accounts of representation.

References

Ackerley, R., & Barnes, G. R. (2011). The interaction of visual, vestibular and extra-retinal mechanisms in the control of head and gaze during head-free pursuit. *The Journal of physiology*, *589*(7), 1627-1642.

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.

Bigelow, J., & Pargetter, R. (1987). Function. *Journal of Philosophy*, *84*(4), 181-196.

Boorse, C. (1976). Wright on functions. *The Philosophical Review*, *85*(1), 70-86.

[16] As mentioned earlier, there is a wide variety of representational views that are (arguably) non-historical (e.g. Fodor 1987, 1990, Cummins 1989, 1996, Egan 2012, 2014, 2018, Sprevak 2013, Hacohen *forthcoming*, Nanay 2010, 2014).

Burge, T. (1986). Individualism and Psychology. *The Philosophical review,95*(1), 3-45.

Chemero, A. (2009). *Radical Embodied Cognitive Science*, Cambridge, MA: MIT Press.

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

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

Craver, C. F. (2007). *Explaining the Brain.* Oxford: Oxford University Press.

Craver, C. F. (2013). Functions and mechanisms: A perspectivalist view. In *Functions: Selection and mechanisms* (pp. 133-158). Springer, Dordrecht.

Cullen, K. E. (2012). The vestibular system: multimodal integration and encoding of self-motion for motor control. *Trends in neurosciences*, *35*(3), 185-196.

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.

Davidson, D. (1987). Knowing One's Own Mind, in *Proceedings and Addresses of the American Philosophical Association*, 60: 441–58.

Davies, P. S. (2001). *Norms of nature: Naturalism and the nature of functions*. 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.

Dretske, F. (1996). Absent Qualia. *Mind and Language*, 11 (1): 70–130.

Egan, F. (2012). Representationalism. In E. Margolis, R. Samuels, and S. Stich (Eds.), *The Oxford Handbook of Philosophy of Cognitive Science* (pp.250-72). 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.

Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. Cambridge MA: MIT Press.

Fodor, J.A. (1990). *A Theory of Content and Other Essays*, Cambridge, MA: MIT Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*(1), 3-71.

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

Hacohen, O. *forthcoming*. What Are Neural Representations? A Cummins Functions Approach. *Philosophy of Science*.

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, 144–56. Oxford: Oxford University Press.

Hutto, D., & Myin, E. (2013). *Radicalizing enactivism: Basic minds without content*. Cambridge, MA: MIT Press.

Leigh, R. J., & Zee, D. S. (2015). *The neurology of eye movements (Fifth Edition)*. Oxford University Press, USA.

Lorente de Nó, R. (1931). Ausgewalte Kapitel aus der vergleichenden Physiologie des Labyrinthes. Die Augenmuskel-refiexe beim Kaninchen und ihre Grundlagen. Ergeb. PhysioL Bioi. Chern. Exp. PharmakoL 32:73-242.

Lorente de Nó, R. (1933). Vestibulo-ocular reflex arc. *Arch. Neurol. Psychiatr. Chicago* 30:245-91.

Magnus, R. (1924). Kiirperstellung. Berlin: Springer. 740 pp.

Maley, C. J., & Piccinini, G. (2017). A Unified Mechanistic Account of Teleological Functions for Psychology and Neuroscience. In D. M. Kaplan (ed.), *Integrating Mind and Brain Science: Mechanistic Perspectives and Beyond*. Oxford University Press, 236-256.

Marr, D. (1982). *Vision*. New York: W.H. Freeman.

Miłkowski, M. (2016). Function and causal relevance of content. *New Ideas in Psychology*, *40*, 94-102.

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

Millikan, R. (1989). Biosemantics. *Journal of Philosophy*, 86: 281–297.

Millikan, R. (1996). On Swampkinds. *Mind & Language*, 11: 103-117.

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

Morgan, A. (2014). Representations gone mental. Synthese, 191(2), 213–244.

Nanay, B. (2010). A modal theory of function. *The Journal of Philosophy*, *107*(8), 412–431.

Nanay, B. (2014). Teleosemantics without etiology. *Philosophy of Science*, *81*(5), 798-810.

Neander, K. (1991). Functions as selected effects: The conceptual analyst's defense. *Philosophy of science*, *58*(2), 168-184.

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

Neander, K. (1996). Swampman meets swampcow. *Mind & Language*, 11(1), 118-129.

Neander, K. (2012). "Teleological Theories of Mental Content", *The Stanford Encyclopedia of Philosophy* (Spring 2012 Edition). Retrieved from <http://plato.stanford.edu/archives/spr2012/entries/content-teleological/>.

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

Papineau, D. (2001). The status of teleosemantics, or how to stop worrying about swampman. *Australasian Journal of Philosophy*, *79*(2), 279-289.

Piccinini, G. (2008). Computation without Representation. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *137*(2), 205–241.

Porter, B. (2020). Teleosemantics and tetrachromacy. *Biology & Philosophy*, *35*(1), 1-22.

Ramsey, W. (2007). *Representation reconsidered.* Cambridge: Cambridge University Press.

Robinson, D. A. (1968). Eye movement control in primates. *Science* 161: 1219-24

Robinson, D. A. (1970). Oculomotor unit behavior in the monkey. *J. Neurophysiol*. 33: 393-404

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

Shadmehr, R., & Wise, S. P. (2005). *The computational neurobiology of reaching and pointing: A foundation for motor learning*. Cambridge, MA: MIT Press.

Shagrir, O. (2006). Why We View the Brain as a Computer. *Synthese, 153*(3), 393-416.

Shagrir, O. (2012). Structural representations and the brain. *The British Journal for the Philosophy of Science*, 63(3), 519–545.

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. (2011). Review of Representation Reconsidered by William Ramsey. *British Journal for the Philosophy of Science, 62*, 669-675.

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

Tinbergen, N. (1963). On aims and methods of ethology. Zeitschrift für Tierpsychologie, 20(4), 410-433.