Structural Representations do not meet the job description challenge

Abstract:

Structural representations are increasingly popular in philosophy of cognitive science. A key virtue they seemingly boast is that of meeting Ramsey's job description challenge. For this reason, structural representations appear tailored to play a clear representational role within cognitive architectures. Here, however, I claim that structural representations do not meet the job description challenge. This is because even our most demanding account of their functional profile is satisfied by at least some receptors, which paradigmatically fail the job description challenge. Hence, the functional profile typically associated with structural representations does not identify representational posits. After a brief introduction, I present, in the second section of the paper, the job description challenge. I clarify why receptors fail to meet it and highlight why, as a result, they should not be considered representations. In the third section I introduce what I take to be the most demanding account of structural representations at our disposal, namely Gładziejewski's account. Provided the necessary background, I turn from exposition to criticism. In the first half of the fourth section, I equate the functional profile of structural representations and receptors. To do so, I show that some receptors boast, as a matter of fact, all the functional features associated with structural representations. Since receptors function merely as causal mediators, I conclude structural representations are mere causal mediators too. In the second half of the fourth section I make this conclusion intuitive with a toy example. I then conclude the paper, anticipating some objections my argument invites.

Keywords: Sub-personal representations; Structural representations; Feature detectors; Eliminativism; Job description challenge

1 - Introduction

Philosophers of cognitive science are increasingly attracted by structural representations: vehicles of content which represent their targets in virtue of the structural similarity holding between them (e.g. Churchland 2012, Shea 2018, Williams 2018, Ramsey 2019). Structural representations (SRs henceforth) are popular for a variety of reasons.

One is empirical adequacy. Cognitive scientists often conceive representations as *maps*, which are a prime example of SRs (e.g. O'Keefe and Nadel 1978; Moser, Kropff and Moser 2008). Even when they do not mention maps, cognitive scientists are nevertheless inclined to think of representations as having the same *structure* of their target domain (e.g. Shepard and Chipman

1970; Gallistell and King 2010). The increasingly popular framework of predictive processing relies on SRs too (Kiefer and Hohwy 2018). Since predictive processing is commonly taken to score the victory of the representationalist front in the "representation wars", SRs appear to be substantially immune to anti-representationalist philosophical arguments (Williams 2017).

SRs seems also naturally suited to meet the strict theoretical demands of the *job description challenge* (Ramsey 2007). Their vehicles appear to be nicely tailored to play a genuine representational role within the functional economy of cognitive systems. As a consequence, philosophical accounts of SRs seem to fulfill their foundational role, providing a sharp metaphysical vindication of the representational lexicon cognitive science deploys.

Here, I claim that the latter point is mistaken. I will argue that even our *most demanding* account of SRs (Gładziejewski 2015b, 2016) is satisfied by at least some receptors. Given that receptors paradigmatically *fail* the job description challenge (e.g. Ramsey 2003, 2007; Orlandi 2014; Downey 2018; Anderson and Chemero 2019), I will conclude that SRs, at least thus characterized, fail it too. Even the most demanding account of SRs at our disposal fails to specify a sufficiently robust representational functional profile.

The essay is structured as follows. In section 2 I briefly introduce the job description challenge, sketching the reasons as to why receptors are supposed to fail it. In section 3 I present Gładziejewski's (2015b, 2016) account of SRs, which strikes me as the *most demanding* account of SRs currently on offer. In section 4 I turn from exposition to criticism. In the first half of the section, I show that at least some receptors can satisfy all the theoretical demands Gładziejewski's account places on SRs, thereby showing that their functional profile is not substantially different. In the second half of the section I propose a toy example to show, by analogy, that SRs (as Gładziejewski defines them) fail the job description challenge for the same reasons receptors fail it. Concisely, our most demanding account of SRs identifies as representations structures that do not play any *recognizable representational* role within the systems in which they function. Section 5 closes the paper defending the argument from a number of objections.

2 - The job description challenge, and receptors

The job description challenge has a simple premise: representations belong to both an intentional kind and a functional kind (Ramsey 2007, pp. 1-36). Representations have intentional properties: they are things (vehicles) *about* other things (targets), which they represent. Representations also have a specific *functional profile.* Pictures, sentences, engravings and utterances form a kind not in virtue of a common physical property, but because they all function as representations for us. So, if the representationalist commitments of cognitive science are justified, the *explanantia* cognitive science posits must meet two distinct demands (Ramsey 2015): (i) a demand for content (i.e. they must possess intentionality) and (ii) a demand for function (i.e. they must be deployed *as representations* within cognitive systems).

Notice that both (i) and (ii) are ordinary membership conditions for the corresponding kinds. An item belongs to an intentional kind only if it has content, as (i) demands. Similarly, an item belongs to a functional kind only if it performs some relevant function, as (ii) demands. Therefore, if representations belong to both kinds and cognitive science *really* posits representations, these posits must satisfy (i) and (ii) jointly. When it comes to mental or cognitive representations, naturalistic theories of content (e.g. Millikan 1984; Fodor 1990) provide a principled way to satisfy (i).[1] However, we lack a principled way to satisfy (ii). Indeed, aside from quick remarks about "standing

[1] Importantly, these theories try to provide an account of original (non-derived or intrinsic) content. Roughly put, content is original when it is not grounded in some already contentful state, item or process. Notice further that the distinction between mental and public representations is *orthogonal* to the distinction between original and derived content according to at least some naturalistic accounts of content. For instance, according to Millikan's teleosemantics, bee dances have original content, even if they are not mental representations (see Millikan 1984; see also Lyre 2016; Vold and Schlimm 2020 for other examples). There might even be mental representations whose content is not original (see Clark 2010 for a possible case). At any rate, in the following I will use "intentionality" and "content" as meaning "original intentionality" and "original content", unless stated otherwise. I will also use the term "representation" as a shorthand for "representation with original content", whether public or mental.

in" (Haugeland 1991; Grush 1997; Clark 1997), no account of the *functional* profile of representations was proposed. The job description challenge is the challenge of providing such a profile. The challenge is thus that of specifying a set of functional features the possession of which is *sufficient* for an item to function as a representation within some system.

This challenge can be met by analogy (Ramsey 2007, p. 10; see also Gładziejewski 2015b pp. 82-84; 2016 pp. 564-566, Smortchkova *et al*. 2020, pp. 10-11). If the functional profile of a purportedly representational posit is nontrivially similar to the functional profile of an entity that we would *pretheoretically* categorize as a representation, we can leverage their similarity to satisfy (ii). For instance, if neural states have a functional profile nontrivially similar to that of models, we can say neural states function as representations *by functioning as models* (e.g. Ramsey 2007, pp. 67-118; Williams 2018).[2] Notice, however, that the same procedure can deny that (ii) is satisfied. If, for instance, an alleged representational posit has a functional profile nontrivially similar to that of a battery, we clearly cannot say that it functions as a representation *by functioning as a battery*. Indeed, if the structure under scrutiny is *correctly* characterized as a battery, describing it as a representation (e.g. by saying it represents how much longer a process can still run) is explanatorily redundant, and might put at risk future research (e.g. by leading us to wonder how content is encoded rather than how energy is stored).

Receptors are typically offered as a kind of posit that paradigmatically fails the job description challenge (e.g. Ramsey 2003; 2007, pp. 118-151; Orlandi 2014; Williams and Colling 2017, p. 1949; Downey 2018).[3] Painted with a broad brush, the core idea behind the receptor notion of representation is that if an internal state V of some system reliably co-occur with some distal event T, then V is a representation of T.

[2] It should be noted, however, that such a similarity, albeit *sufficient* to meet the challenge, is not *necessary* to meet it. In fact, Ramsey seems to allow that certain posits actually qualify as genuinely representational mostly because of their explanatory role within a theory. Arguments by analogy, however, are by far the most popular way to confront the challenge, and therefore they will be the focus of the present treatment.

[3] See also (Artiga and Sebastián 2018) for an argument to the same effect which is largely independent from Ramsey's (2007) framework.

This notion is often further elucidated referring to Dretske's (1981; 1988) account of representation, which is often taken to underpin receptors (e.g. Ramsey 2003; 2007; Morgan 2014; Nirshberg and Shapiro 2020). At the core of Dretske's account of representation lies the notion of indication. Succinctly, V *indicates* T if, and only if, $P(T|V) = 1$. Put this way, however, the notion of indication is extremely susceptible to the disjunction problem (e.g. Fodor 1989). Crudely put, it is too often the case that $P(T|V) < 1$ and $P(T^*|V) < 1$, but $P(T \text{ v } T^*|V) = 1$. In such a case, given the notion of indication previously proposed, we should conclude V indicates (and thus, represents) T *or* T*. To avoid this problem, Dretske (1988; 1994) revises the definition of indication and adds a teleological component to it. According to the revised definition, V indicates T if, and only if, $P(T|V) > P(T)$; where T is a subpart of the (possibly disjunctive set of) distal states of affairs T* such that $P(T^*|V) = 1$[4] (see also Rupert 2018, pp. 207-209). The teleological component, instead, requires V to be "supposed to" indicate T, where the "supposed to" part gets unpacked by saying that V is supposed to indicate T just in case V has been recruited within some system in virtue of the fact that it indicates T (according to revised definition of indication). The recruitment procedure might vary: Dretske (1988) extensively relies on reinforcement learning, but natural selection and intentional design are typically held to be sufficient recruitment procedures too (e.g. Shea 2018, Ch. 3).

Several structures[5] qualify as receptors according to this picture. Single neurons, for instance, are often said to represent whichever distal variable (object or state of affairs) triggers their suprathreshold firing the most (e.g. Levittin, Maturana, McCulloch and Pitts 1959; Hubel and

[4] Notice that according to *both* definitions, indication is *not* a causal notion. The fact that V indicates T *might, but need not*, obtain in virtue of a causal relation holding among V and T (Dretske 1981, pp. 26-39). Nor does "Shannon information" (around which the original notion of indication was modeled) necessarily depend on any straightforwardly causal link (Shannon and Weaver 1949). In fact, textbooks on information theory are silent on causality (e.g. Cover and Thomas 2006). All that matters seems to be uncertainty reduction.

[5] Importantly, as a reviewer noticed, taking entire structures as representations is a deviation from Dretske's framework. In Dretske's view, it is not correct to say that, for instance, a barometer represents the pressure. Rather, we should say that the barometer being in state *s* represents the fact that the pressure is *n* Pascals. However, this loose usage is not just prominent in the literature (e.g. Morgan 2014, pp. 231-232; Williams and Colling 2017, p. 1947), it also strikes me as entirely unproblematic. To continue with the previous example, the claim that a barometer represents the pressure is entirely intelligible and easily unpacked by saying that the barometer represents the pressure of a given environment by occupying, at any moment, the state that indicates the pressure at that moment.

Wiesel 1962, 1968; Nieder, Diester and Tudusciuc 2006). In this view, their increased firing rate indicates the presence of some specific target in the animal's visual field (see Eliasmith 2005 for an updated discussion). In a similar spirit, the nodes in the hidden layers of connectionist architectures are often said to represent the input patterns with which their activity correlates the most. Furthermore, each individual node is said to represent the microfeature driving the node's activity the most (e.g. Gorman and Sejnowski 1988; Gosche and Koppelberg 1991; Clark 1993).

Receptors surely meet condition (i) of the job description challenge. Indeed, Dretske's (1988) theory of content can be leveraged to assign content to these structures. Yet, they seem unable to meet condition (ii). Indication is surely not sufficient for representation (the sea level indicates the position of the moon, but surely the sea does not represent the moon[6]). Having the function of indicating does not seem sufficient either. In fact, all sorts of things are recruited within systems in virtue of their indicator properties, without thereby becoming representations of what they indicate. Bi-metallic strips of thermostats and photosensitive cells of optical smoke detectors all have the *function* (by purposeful design) of indicating some distal target; yet they are not, *prima facie*, representations. In fact, within these mechanisms, both receptors act just like reliable causal mediators, allowing the system to robustly produce a certain output (for instance, turning off a furnace) when a given environmental condition obtains. The same holds, for instance, for the firing pin of a gun. The state of the firing pin indicates the position of the trigger: if the firing pin is in contact with the bullet, then the trigger has (typically) been pulled. Hence P(trigger pulled|firing pin in contact with the bullet) > P(trigger pulled). Moreover, firing pins are included in guns *because* of this relation: it is the fact that their position indicates whether the trigger has been pulled that enables us to control when to shoot. But surely guns are not representational systems. Thus, when it comes to the functional profile of receptors, they behave as mere causal mediators (such as firing pins); and, for this reason, they shouldn't be considered representations. Indeed, many believe that

[6] In order to justify this claim, it is sufficient to notice that the level of the sea cannot misrepresent the position of the moon. But something can count as a representation only if it can misrepresent in at least some cases.

considering receptors as representations has nasty consequences.

Panrepresentationalism is the first. Considering receptors as representations entails that they satisfy (i) and (ii) jointly. But then it is almost impossible to deny bi-metallic strips (or firing pins) also satisfy them. Given the shared functional profile, if receptors satisfy (ii) then bi-metallic strips (and the like) satisfy it too. And we can apply Dretske's (1988) account of content to allow them to satisfy (i). After all, they have, by design, the *function* of indicating something within the systems deploying them. Thus, accepting that receptors are representations entails panrepresentationalism: the (clearly mistaken) view that whichever entity reliably coordinates with environmental contingencies is *representing* these contingencies.[7] But any account of representations entailing panrepresentationalism is surely metaphysically flawed, as it fails to establish a substantial distinction between representational and non-representational states (Ramsey 2003; 2007, pp. 125-127).

The empirical adequacy of the relevant notion of representation is also under threat. If philosophical theories of cognitive representation aim at capturing the notion of representation cognitive science deploys, they must provide a notion of representation which is distinctively psychological or cognitive. But a notion of representation that applies to thermostats or firing pins seems to lack any distinctively psychological or cognitive connotation (Orlandi 2014 pp. 107-110; Ramsey 2017).[8]

[7] Notice here that panrepresentationalism is a problem only because I'm assuming that the content at play here is *original*. There is, I believe, no problem of panrepresentationalism related to *non-original* (or derived) content, for each and every thing can, in principle, be assigned some derived content. We could surely stipulate, for instance, that a mug represents Napoleon, or that a pair of shoes represents Castor and Pollux. This seems also the reason why semioticians (who are interested in representations with both original and derived content) have no problem in saying, for instance, that a cigarette butt found on a crime scene represents the fact that the murder is a smoker, or that finding my fingerprints on a surface signals the fact that I touched that surface. In all these cases, the relevant signs (or representations) are tied to their targets only by a loose causal connection. However, this does not generate any problem with panrepresentationalism because their content is derived, as it depends on the interpretation of some clever detective (or some other interpreter).

[8] In the original formulation, I employed the term "(neuro)psychological" instead of "psychological or cognitive". An anonymous reviewer noticed that the original formulation was too strong: what if intentionality were to be naturalized in purely causal-computational terms? I agree with the reviewer, thus I resort to the more neutral (and I fear more vague) "psychological or cognitive" formulation. Yet, I wish to highlight that the same problem remains, regardless which sort of account will finally succeed in fully naturalizing intentionality. For, even if intentionality were to be fully naturalized in causal-computational terms, *at least some* causal-computational goings on should turn out to be

Accepting that receptors are representations also reduces the *explanatory power* of the notion of representation invoked. Since treating bi-metallic strips (and the like) as representations add nothing to our non-semantic comprehension of these devices, the notion of representation appears to be merely a semantic *gloss* glued to an ultimately non-semantic understanding. This explanatorily inert notion of representation is at odds with the representationalism of cognitive science – at least as long as we regard it as a *substantial* empirical hypothesis (Ramsey 2017).

Many found that these problems are collectively sufficient to reject the receptor notion of representation (e.g. Ramsey 2003; 2007; Orlandi 2014; Downey 2018; Anderson and Chemero 2013; 2019). And even when the notion is not *explicitly* rejected, more than a shadow of doubt is cast over its explanatory potential (e.g. Williams and Colling 2017 p. 1949).

This concludes the presentation of the job description challenge. In the next section, I introduce a strong account of SRs and highlight why they are supposed to meet the job description challenge.

3 - Gładziejewski's account of structural representations

Accounts of SRs abound in the philosophical literature (see references in section 1). Here, I focus only on Gładziejewski's (2015a; 2015b; 2016) account, as it is the ideal target of the present discussion. This is because of three distinct reasons.

To start, Gładziejewski's account aims at spelling out the *functional profile* of SRs (Gładziejewski 2015b; 2016), trying to distinguish them from receptors (Gładziejewski and Miłkowski 2017). For this reason, Gładziejewski's account is *demanding*, as it is designed to avoid trivializing counterexamples. Hence, it should be particularly *resistant* to my argumentative strategy, guaranteeing I'm not attacking a strawman.

non-intentional; otherwise, the empirical adequacy of the account would be seriously threatened (minimally, because pan-intentionalism is not a *desideratum* of a naturalistic theory of intentionality).

Secondly, and relatedly, Gładziejewski's account can leverage two theories of content to meet demand (i). One is varitel semantics (Shea 2018, pp. 111-144). The other is a special purpose theory of content tailored to Gładziejewski's account (Lee 2018). This allows me to simply *assume* the account meets (i), so to focus exclusively on (ii); *pace* Segundo-Ortin and Hutto (2019).

Lastly, Gładziejewski's account of SRs should boast a significant theoretical strength. In fact, it provides the standard understanding of SRs in predictive processing (e.g. Keifer and Hohwy 2018; Ramstead, Kirchhoff and Friston 2019). If predictive processing really is safe from anti-representationalist attacks (Williams 2017), then Gładziejewski's account should be extremely strong.

Now, the account. According to Gładziejewski (2015b; 2016), a vehicle V is a SR of a target T only if:

(1) V and T are structurally similar, &

(2) A system S exploits V's structural similarity with T to guide its actions regarding T, &

(3) V is decouplable from T, &

(4) S can detect the representational errors of V

Concise unpacking is needed. The relevant notion of structural similarity leveraged in (1) is *second order structural resemblance* (O'Brien and Opie 2004). Hence, V is structurally similar to T *if and only if*: (a) at least some features of V map one-to-one onto at least some features of T; (b) at least some relations defined over the features of both V map one-to-one onto at least some relations defined over the features of T; and (c) for each pair of features of V in a given relation, the corresponding features of T are in the corresponding relation.[9] More intuitively still, V is

[9] Here I'm trading precision for clarity: in particular, I'm suppressing the set-theoretic lexicon of the original formulation in favor of intuitiveness and ease of exposition.

structurally similar to T just in case the same inner *abstract pattern* of relations holds among the features of both V and T.

As representations in cognitive science are supposed to explain intelligent behavior, (2) is a natural requisite. Notice, however, (2) requires the structural similarity to be *exploited* by a system S. Therefore, S’s behavior must be sensitive to the relations among the features of V; which must map onto features of T relevant to the functioning of S. (Shea 2014; 2018, p. 120). Furthermore, (2) makes action-guidance *constitutive* of the representational status of V. This brings about hefty theoretical advantages. Since (2) is defined over a system S, it makes the representational relation triadic, solving the problems of reflexivity, symmetry and content underdetermination (see Goodman 1969). Content underdetermination and reflexivity are solved because the fact that V guides S's actions about T *makes* V a representation *of T* rather than any other thing V might structurally resemble (V itself included). Similarly, the problem of symmetry is dealt with by noting V guides S's actions about T, but not *vice versa* (Williams and Colling 2017).

Representations are often defined as stand-ins for absent targets (e.g. Haugeland 1991), so (3) is an obvious component of any theory of representations. Gładziejewski's (2015b, p. 77) definition of decouplability is twofold. V is *weakly* decoupled from T only if no causal chain runs from T to *both* V and V's consumer within S. V is *strongly* decoupled from T only if no causal contact obtains between S and T.

Lastly, (4) requires that V might generate some error S can detect. Thus, if V is a SR in S, S must be capable of assessing the accuracy of V through some monitoring device. Condition (4) is not often required, and its dispensability has been suggested[10] (Lee 2018, p. 4). Yet I see no reason to dispense it. It nicely fits the theoretical apparatus of predictive processing, whose representations Gładziejewski’s (2016) account is trying to capture. Moreover, it protects the account from

[10] Notice that some would also suggest that decouplability (i.e. point (3)) is dispensable (e.g. Miłkowski 2017). See also (Chemero 2009, pp. 50-55).

trivializing counterexamples (see Miłkowski 2013, p. 161 for a brief, but insightful, case).

According to Gładziejewski (2015b, pp. 69; 2016), (1) to (4) are *sufficient* to identify vehicles with a functional profile satisfying (ii) because they are the functional features of *cartographic maps*. As (1) requires, maps are structurally similar to the terrain they map, as (a) each point on the map corresponds to an environmental landmark; (b) the spatial relations among points correspond to the spatial relations between landmarks; and (c) for each two points the map displays in a certain spatial relation, the corresponding relation holds for the corresponding landmarks. This structural similarity guides our actions, as (2) requires: we are sensitive to the spatial relations a map displays (e.g. we rely on them to find the *shortest* path from A to B); and the layout of a map displays the features of the environment relevant to our navigation of that terrain. Maps are clearly decouplable from their targets, as (3) requires: we can use a map of an arbitrary city to plan ahead our trip there, without any causal link connecting us to that city. Lastly, as (4) requires, we can detect the representational error of maps. For instance, we would deem inaccurate a map that reliably gets us lost. Hence, according to Gładziejewski, SRs function as representations *by functioning as maps*, meeting the job description challenge head on.

Summarizing, Gładziejewski takes (1) to (4) to be *jointly sufficient* to spell out a robustly representational functional profile for SRs (e.g. Gładziejewski 2015b, p. 69). Moreover, Gładziejewski takes (1) to (4) to be sufficient *because* of the analogy with cartographic maps. By so doing, he directly answers Ramsey's job description challenge through an argument by analogy. In the next section, I will claim that Gładziejewski's answer is incorrect. I will show that at least some receptors (which, it is assumed, do not meet the job description challenge) can meet (1) to (4) jointly. Moreover, I will show, using a toy example, that doubtlessly non-representational structures can meet (1) to (4) too. If this is correct, (1) to (4) do not spell out a robustly *representational* functional profile. Thus, satisfying them cannot be sufficient for a posit to meet the job description challenge. As a consequence, if SRs are defined in terms of conditions (1) to (4), then they do not meet the job description challenge.

4 - Structural representations fail the job description challenge

I divide this section in two blocks. In the first, I show that at least some receptors can meet (1) to (4) in conjunction. In the second, I propose a toy example to show that (1) to (4) can be jointly satisfied by non-representational structures.

4.1 - Receptors can meet (1) to (4)

According to (1), the vehicles of SRs are structurally similar to the targets they represent. But every receptor is structurally similar to its target, as indication is *sufficient* to establish a structural similarity (see Morgan 2014; Nirshberg and Shapiro 2020). This should not be puzzling: structural similarities are fairly cheap.

Thus, consider a paradigmatic receptor such as the bimetallic strip of a thermostat. It surely indicates the temperature: finding the strip occupying a given state raises the probability that the temperature in the room is in the corresponding state. Moreover, the strip has the function of indicating the temperature. In fact, bi-metallic strips are included in thermostats (by human design) precisely because of their properties as indicators.

It is fairly easy to show to show that such a receptor is structurally similar to the environmental temperature (its target). Let the various states of the strip be defined as elements v_x belonging to a set V, and let the range of temperatures indicated by the strip be defined as elements t_x belonging to a set T. By definition, V and T have the same cardinality. Moreover, since each element of V indicates one and only one element of T, the one-to-one mapping from V onto T required by (a) obtains. Let now two relations be defined, one (*longer than*) over V, and one (*hotter than*) over T. Both relations impose a strict total order among the elements of the respective sets. Hence they have the same mathematical structure and *non gratuitously* map onto each other, just as (b) requires.[11] Notice also both relations here defined are far from arbitrary: indeed, these relations are *essential* to the functioning of a thermostat. Lastly, for each arbitrary pair of elements (v_a,v_b) ordered by *longer than*, there exists a pair (t_a,t_b) ordered by *hotter than* such that v_a maps onto t_a and v_b maps onto t_b as (c) requires. This is just the way the bi-metallic strip works: it gets longer as the temperature rises. Hence, the relation of indication making the bi-metallic strip a receptor of the environmental temperature is *per se sufficient* for a structural similarity to obtain between the two.

This point easily generalizes. Given any arbitrary receptor, its states will always map one-to-one onto the states of the environment they indicate, providing the mapping in (a). The states of the receptor and the states of the environment will also always bear some receptor specific[12] reciprocal relations, providing what (b) requires. Lastly, each arbitrary pair (or other polyadicity) of receptor states in a given relation will map one-to-one onto the corresponding states of the environment in the corresponding relation, satisfying (c). This is just how receptors work. Thus (1) obtains for all receptors.

As far as I know, none has ever denied receptors are causal mediators performing action-guiding duties. The artificial agents produced by behavior-based robotics (e.g. Brooks 1999) nicely exemplify how receptors can guide the seemingly cognitive behavior of a system. Yet, this is insufficient to claim receptors meet (2). Indeed, even when receptors are rings in the causal chain that leads a system to the production of a given behavior, the system might not be *exploiting* any receptor-target structural similarity (Gładziejewski and Miłkowski 2017).

[11] I owe the phrasing of this point to my colleague Silvia Bianchi.

[12] Some examples servicing intuitive clarity: the hair in a hair hygrometer gets *longer* as the humidity *rises*; the floating unit of a fuel gauge gets *lower* as the tank gets *emptier*; the return signal of a proximity sensor is *faster* as the target gets *closer*, and so on.

The argument can be summarized as follows. Consider again the bi-metallic strip of the thermostat. Let it be sensitive to three environmental temperatures, ordered by *hotter than* in the triplet (t_a,t_b,t_c). Let v_a, v_b and v_c be the corresponding states of the bi-metallic strip. Suppose now that *longer than* orders these states in the triplet (v_b,v_c,v_a), preventing the relevant strip-temperature structural similarity from obtaining. Yet the strip can still successfully orchestrate the behavior of the thermostat, at least as long as it enters in each state when the environment is in the corresponding temperature (i.e. as long it correctly indicates) and each state leads the system to behave as it has been designed to behave. So the relations among the features of the vehicle are *irrelevant* to the functioning of the system. As a consequence, the structural similarity is not exploited, as a structural similarity is exploited *only if* a system is sensitive to the relations among the features of the vehicle (Shea 2014; 2018 p. 120).[13] Receptors might be structurally similar to their targets (and as a matter of fact they are). Yet, this similarity does nothing for the system and deserves to be called an epiphenomenon (Gładziejewski and Miłkowski 2017).

However, there exists a target-receptor structural similarity which every receptor must instantiate and that cannot be epiphenomenal in the sense just seen. To see why, consider again the triplet (t_a,t_b,t_c), this time letting the three temperatures be ordered by their temporal relations (i.e. t_x *is followed after an amount of time x by* t_y). Again, let v_a, v_b and v_c be the corresponding states of the strip. Let them be ordered again in the triplet (v_a,v_c,v_b), this time by their temporal relations[14] (i.e. v_x *is followed after an amount of time x by* v_y). *Ex hypothesis*, the structural similarity is again absent. Yet, in this case, the system will malfunction. The reason is simple: if v_a *is followed after an amount of time x by* v_c and t_a *is followed after an amount of time x by* t_b, then the strip will occupy state v_c when the temperature is t_b. But the state of the strip indicating t_b is v_b, not v_c. Therefore, the

[13] To be precise, Shea's definition of exploitability also imposes that the features of the target and their relations must be of significance to the system, where "significance" is at least partially determined by the system's functions. Given that Gładziejewski and Miłkowski (2017) do *not* discuss this aspect of exploitability and simply assume it obtains, I will assume it too.

[14] Notice having the same kind of relations on both sides of the similarity is perfectly legitimate. Indeed, maps do represent spatial relations through spatial relations.

receptor mis-indicates. As a consequence, the system will malfunction: its inner state will bring about the behavioral outcome appropriate to t_c instead of the one appropriate to t_b. Therefore the system is sensitive to (at least) the temporal relations holding among the features of V; and the obtaining of such a time-dependent structural similarity between V and T determines the appropriate functioning of the system. At least this time-dependent structural similarity is not epiphenomenal.

Notice that this structural similarity too obtains purely in virtue of indication, as indication is time-dependent. In fact, each receptor must instantiate at least this "special" structural similarity, as an item failing to instantiate it cannot be a receptor. This can be shown by *reductio*.

Thus, suppose an item V is a receptor of a target T. Suppose further no relation (not even temporal ones) can be found such that (c) obtains. *Ex hypothesis*, V and T are not structurally similar. This implies that *when* the receptor is in a state v_a, the target can be in any arbitrary state t_x. To see why, consider the following scenario. Suppose that, at time *t*, the receptor is in a state v_a and the target is in a state t_a. Now, at time t^*, the receptor and the target change state: the receptor goes in state v_b and the target goes through a sequence of state changes t_b...t_n.[15] Suppose further that, at time t^{**}, the receptor returns in state v_a. Let us call *x* the amount of time lapsed between *t* and t^{**}. It is thus correct to say that v_a was followed v_a after an amount of time *x*. Now, it is fairly easy to show that, *ex hypothesis*, at time t^{**} the target must be in any other arbitrary state t_x different from t_a. For, if it were in state t_a, it would be correct to say that t_a was followed by t_a after an amount of time *x*, which is enough to make the receptor and the target structurally similar.[16] But, by stipulation, receptor and target are not structurally similar. Notice that this line of reasoning is perfectly general, as it holds for all time-spans, receptor states and target states (e.g. if at t^{**} the receptor is in state v_k and the target is in state t_b, and then at a further time t^{***} the receptor is again in state v_k and the

[15] This sequence might also include t_a. The point I'd like to make would not be challenged by its inclusion.

[16] To be sure, that would be a very *thin* structural similarity. Yet notice that the relevant definition of structural similarity Gładziejewski endorses quantifies only on "at least some", and it thus seems satisfied by what it is shown in my example.

target is again in state t_b, there *would* be a time-dependent receptor-target structural similarity).

Thus, if a receptor and its target are not structurally similar, *when* the receptor is in a given state v_a, the target can be in any arbitrary state t_x.[17] But if when the receptor occupies state v_a the target can be in any state t_x, then the probability of finding the target in any individual state given that the receptor is in state v_a equals the probability of that state itself. Hence, it would be false that v_a indicates any state t_x of the target, as $P(t_x|v_a) = P(t_x)$. Moreover, as this reasoning holds for all the states of the receptor, it would be false that V is a receptor of T. And this runs counter to the initial stipulations; namely, that V *is* a receptor of T.

In perhaps less convoluted terms, for any arbitrary receptor state v_a to indicate an arbitrary target state t_a it must be the case that, *when* the receptor occupies state v_a, it is more likely than otherwise that the target occupies state t_a. The same holds for all other receptor states v_b...v_n and the corresponding target states t_b...t_n. As a consequence, if v_a is followed after an amount of time *x* by v_b, then it must be likely that t_a is followed after the same amount of time by t_b. Thus, it seems that the relevant time-dependent structural similarity holds purely in virtue of indication.[18]

An anonymous reviewer greeted this passage with a counterexample and a challenge. Let me start with the counterexample. Consider litmus papers: stripes of chemically treated paper that change color when immersed in chemical substances, thereby indicating the pH of the substance. Suppose I use one such device to measure the pH of a substance at time *t*. At t^*, I extract the paper from the substance, which I then dilute with water. The substance's pH has changed, but the color of the paper has not. Yet, it is still correct to treat the paper as a receptor representing the substance's pH. Isn't this a proof that the time-dependent structural similarity discussed above does not hold universally for receptors?

[17] Notice that t_a is included, as it was (by stipulation) the state occupied by the target in the beginning of the example.

[18] Importantly, from this it follows that all systems relying on receptors to organize their behavior are exploiting at least this time-dependent structural resemblance, as it cannot be merely epiphenomenal.

I concede that the litmus paper at t^* is still indicating. In fact, I would add that it is mis-indicating[19], as its color does *not* match the substance's pH. Notice however, that such a mis-indication occurs *at time* t^*, and only because the litmus paper has not changed color as the relevant time-dependent structural similarity prescribes. As long as misindication occurs, the time-dependent structural similarity is broken. But suppose now that, at a further time t^{**}, the litmus paper is put in contact again with the substance. It *would* change color, and it *would* correctly indicate the substance pH. Let x be the amount of time lapsed between t and t^{**}. The substance pH at t is thus followed, after an amount of time x, by the substance pH at t^{**}. But the same relation holds for the states of the litmus paper: color at t is followed, after an amount of time x, by color at t^{**}. Hence the time-dependent structural similarity is restored. Of course, the time-dependent structural similarity instantiated by the litmus paper has, in this example, proven insensitive to the change of state of the substance at t^*. But similarity is a *graded* notion; and even uncontroversial cases of SRs are manifestly not *perfectly* structurally similar to their targets (Williams and Colling p. 1947; Gładziejewski and Miłkowski 2017). A map perfectly (e.g. millimeter by millimeter) similar to the depicted terrain would be useless.

Now, the challenge. Shea (2018, p. 119) illustrates a non-exploited structural similarity with the following example: suppose that a pack of vervets has three kinds of predators p_1, p_2 and p_3. Suppose that the vervets have three types of alarm calls c_1, c_2 and c_3, one for each predator. Suppose that p_1, is taller than p_2; which is in turn taller than p_3. Suppose further that the same ordering holds for the calls: c_1 has a higher pitch than c_2 which in turn has a higher pitch than c_3. The system of calls is thus structurally similar to the system of vervet's predators. However Shea argues that vervets are not sensitive to the relation "*higher pitch than*" holding between their calls. All vervets do, in Shea's view, is to respond separately to each individual call. Thus, Shea concludes that the

[19] The anonymous reviewer also suggested that in such a case the litmus paper would count as a decoupled receptor of pH-in-the-past. I think I disagree. It seems to me that litmus papers have, by design, the function of indicating the pH of substances *in the present*.

structural similarity holding between vervet calls and predators is not exploited. Importantly, Nishberg and Shapiro (2020, p. 16) concede Shea the point that, *taken as an array*, the system of calls is not a SR of the heights of predators.[20] The reviewer asks whether the time-dependent structural similarity I'm discussing contradicts Shea's verdict, showing that an exploitable structural similarity holds between the system calls and the predators.

I believe that, in this regard, it is important not to conflate two distinct issues. The first is whether it is *necessary* that a structural similarity holds between an array of receptors and the ensemble of their targets. To this question, I, together with Nirshberg and Shapiro (and presumably Shea), answer negatively. The hygrometer measuring the humidity of room A is structurally similar to the humidity in room A, and the thermometer measuring the temperature in room B is structurally similar to the temperature in room B. However, the *thermometer plus hygrometer* system need not be (albeit it might) structurally similar to anything. An array of SRs need not be a SR on its own. Notice that the same thing holds for *uncontroversial* instances of SR too. I can place my map of Sydney *north of* my map of Rome without thereby generating a new SR that misrepresents the relative positions of Sydney and Rome.

The second issue regards whether that system of calls actually is structurally similar to something (and whether such a structural similarity is exploited). And I believe the time-dependent structural similarity I introduced actually allows for a positive answer to both questions. For the alarm calls to be effective, these must be tokened in a way such that the temporal ordering between calls matches the one holding between the apparences of predators. Thus, if the three predators appear in the temporally ordered sequence (p_1, p_2, p_3), the alarm calls need to be uttered in the corresponding temporally ordered sequence (c_1, c_2, c_3). Changes in this sequence result, at least *prima facie*, in dead vervets. Hence, the system relying on these calls to orchestrate its behavior (i.e.

[20] Albeit they hold that each call is structurally similar to one predator (see Nirshberg and Shapiro 2020, p.16).

the pack of vervets) seems sensitive to at least these relations.[21]

I now return to the main argument. Can receptors be decoupled as (3) requires? The intuitive answer seems negative. Thermostats, hygrometers and the like can indicate only in virtue of the constant causal contact holding between them and their target. A thermometer somehow shielded from the causal touch of the surrounding mean kinetic energy *would simply stop indicating*. This is troublesome, given that decouplability is often taken as the hallmark of genuinely representational states (e.g. Clark and Toribio 1994; Clark 1997; Clark & Grush 1999; Clark 2013a, pp. 128-131; pp.151-156). The very notion of receptor apparently points at a significant functional difference separating them from (structural) representations.[22]

However, this is surely a misguided appearance, for some receptors routinely perform their action-guiding duties when decoupled. For an example in the cognitive domain, consider a simple Braitenberg vehicle displaying a light-following behavior (Braitenberg 1984). The control system of this robot is fairly rudimentary: it consists only in two laterally placed front-facing photoreceptors, contralaterally connected to a motor by an excitatory link. When this simple agent faces a light source, two beams of light will impinge onto its photoreceptors, coupling the two. The receptors will thus excite the two motors, causing the robot to beeline towards the light source. But if the light source is located on one *side* of the vehicle, only one receptor will be coupled to it by a light beam. Thus only one wheel will turn, causing the robot to spin in place, re-orienting it towards the light source. Notice that albeit in this case only one receptor is coupled, the behavior is orchestrated by *both* receptors. Indeed, it is only *because* one receptor is not coupled to the light source that one

[21] Notice also that, at least in this case, single calls afford the detection of representational error. It is in fact suggested that repeated mistokening of these calls might cause the “liar” vervet to be ignored by the pack (e.g. Cheney and Seyfarth 1985, p. 160).

[22] An anonymous reviewer suggested that Gładziejewski actually embraced the existence of decoupled receptors in his (2015a). Specifically, the reviewer argues that Gładziejewski (2015a) used to consider indications of interactive potentialities (see Bickhard 1993; 1999) as decoupled receptors (they are decoupled because they indicate *future* actions). I’m unsure whether this is the correct interpretation of (Gładziejewski 2015a). In fact, it seems to me that Gładziejewski understood (and maybe still understands) indications of interactive potentialities as *tacit* representations rather than receptors (see Gładziejewski 2015a, p. 19). However, as far as I can see, nothing in the present essay hinges over this point.

wheel does not turn, allowing the robot to spin in place. Even in this minimal case, a *decoupled* receptor is causally contributing to the behavior of a system.

Since in this case one receptor still is coupled to its target, it might be objected that (3) is not *really* met. However, a minimal increase of complexity allows for a *weak* decoupling to completely obtain. A nice example is provided by DidaBots (Maris and Schaad 1995; Maris and te Boekhorst 1996): simple robots tasked with clustering cubes in an arena. Their control architecture shares many features with the Braitenberg Vehicle examined before. It consists in four lateral proximity sensors connected to two lateral motors through both excitatory (ipsilateral) and inhibitory (contralateral) connections. Thus, when a receptor “sees" a cube, it speeds up the movements of the wheels on its side and slows down the speed of the wheels of the other side, causing the robot to turn away from the cube. Notice these robots are “blind” to the front, so if a robot and a cube are lined up, the robot will impact the cube, “picking it up” and pushing it along the way. When, while pushing a cube, the robot “sees" a cube on its side, it will turn away from it, “dropping” the cube it was pushing near the one it has sensed. This is how the robot cluster cubes. The important point to notice here is that the “picking up” and pushing of a cube is a behavior *governed by* decoupled receptors, as the robot can enact this behavioral routine as long as *all* its sensors are not coupled to any cube. Were one of them coupled to a cube, the robot would immediately turn away from it, dropping the cube it was pushing as a result. So the “picking up a cube” behavioral routine is, in Gładziejewski's terminology, orchestrated by weakly[23] decoupled receptors.

Apparently, however, this is still insufficient to vindicate (3), as receptors and SRs boast very different kinds of decouplability (Gładziejewski and Miłkowski 2017; see also Pezzulo 2008). Representations can account for the *proactive* behavior of a system, whereas receptors can, at best, *passively* coordinate a system's responses to environmental contingencies. Representations allow for behavioral coordination with absent targets, from which the whole system is *strongly* decoupled. In

[23] Notice strong decouplability fails to obtain: the whole robot is coupled to the cube it's pushing.

such cases, behavior is endogenously caused: the causal chain leading a system to produce a behavior starts within the system itself. In contrast, receptor-driven behavior is produced by a causal chain that has its starting point in the environment. This seems a principled distinction between the functional profile of (structural) representations and receptors.

But this distinction appears to be illusory too, for some receptors *can* be the endogenous causes of proactive behaviors directed to targets from which the whole system is strongly decoupled. The recurrent artificial neural networks Harvey and colleagues "evolved" as control systems for robotic agents provides a nice example (Harvey *et al.* 1997). One such agent was tasked with visually tracking a moving target (Harvey, Husbands and Cliff 1994). Since the target was moving and the robot was not placed in front of it at every trial, there were significant spans of time in which no robot-target coupling obtained, and thus significant spans of time in which the two were *strongly* decoupled. In such cases, the robot self-initiated an exploratory behavior (namely, spinning in place to detect the target). This behavior was produced by a generator unit of the net (Husbands, Harvey and Cliff 1995): an artificial neuron able to "recycle" its output at time *t* as input at time t_{+1} through a recurrent connection. Since the network was noisy, generator units were able, by constantly feeding themselves back their noisy output, to generate significant activity within the net in absence of any environmental input. In the case at hand, the generator unit was a tactile receptor selected (by genetic algorithms) to trigger the "look around" behavioral routine in absence of any relevant external input. Notice the "look around" routine is caused by the *intrinsic* (noisy) dynamics of one receptor in the net. In other terms, the causal starting point of that behavior is within one of the net's receptors, not in the environmental input or lack thereof. Hence, a simple receptor was able both to coordinate a system's behavior regarding a strongly decoupled target and to do so by endogenously initiating the causal chain leading to the relevant behavior of the system. Receptors can thus meet (3), even in its most demanding form.

Lastly, receptors can generate system-detectable error, as (4) requires. The control architecture for robotic agent Bovet (2007) engineered provides a clear example. The architecture consists in a series of homogeneously connected feedforward artificial neural networks, one for each sensory or motor modality of the robot.[24] Simplifying a bit (for reasons of space), each net consists of three identical populations simple neuron-like receptors. Two of these populations jointly form the input layer, and the other is the output layer. Each net works as follows.[25] The *current state population* receives input from the sensors of the modality controlled by the net, entering in the state corresponding to the incoming sensory barrage. The *desired state population* receives instead input from the nets of all other modalities, thus entering in the state the controlled modality *should* occupy, given the activity of the rest of the system. For instance, if the visual desired state population receives the signal that the robot is moving forward, it will enter in the state corresponding to an optic flow expansion, as moving forward typically correlates with optic flow expansion. Together, the current state population and the desired state population form the input layer. The output layer consists in the *desired state change* population, responding to the *difference* between the states of the two halves of the input layer, and spreading that difference to the rest of the system. So the receptors of the output layer respond to the *mismatch* between "desired" and received sensory input, which is a very simple form of *prediction error*.[26]

Notice these "error receptors" are as causally potent as any other receptor in the system. In fact, the activity of the motors is determined (through the motor desired state population) by the output layer of each modality, which spreads the *mismatch* between the two halves of the input layer. This means the motors are active *only if* there is at least one net spreading error. So error is what, causally speaking, drives the system around. Moreover, in a series of experiments (Bovet and

[24] Notice these nets lack both self-recurrent connections and hidden units: the typical resources that are considered representational vehicles in connectionist systems (e.g. Shea 2007; Shagrir 2012). Their activity is thus interpretable in a straightforwardly non-representational manner (Ramsey 1997).

[25] After the learning period, in which the net learns the robot's sensorimotor contingencies (see O'Regan & Noë 2001): the ways in which stimulation changes as a consequence of movement.

[26] Technically, the architecture behaves as if it were detecting the mismatch between the received inputs and the ones self-generated by a forward model (see Bovet 2007, pp. 79-106). This mismatch is ordinarily considered as prediction error in the predictive processing literature, and Gładziejewski (2015b; 2016) himself relies on this very same notion of error.

Pfeifer 2005a; 2005b, see also Pfeifer and Bongard 2007, pp. 295-333) the robot learned to solve a simple working memory task (i.e. finding the reward at one end of a T-maze) by learning to trust a tactile-motor correlation (it learned to "expect" to turn in the direction of the active tactile receptor sensing the cue) over a visuomotor one. This shows that the robot can, implicitly, assess which error is important to minimize and which error is irrelevant.

Importantly, in these experiments, the receptors of the net meet (1) to (4) *jointly*. If the arguments provided thus far are sound, (1) and (2) must obtain, as they obtain for every receptor, and the net is just a series of receptors systematically connected. (4) obtains, as the system has a specialized set of receptors in the task of detecting the error between "expected" and actually occupied sensory states. Lastly, (3) obtains too, as, at the onset of each trial, the robot was *strongly decoupled* from the reward it had to find. Indeed, at the onset of each trial the robot and the reward are in different "arms" of the T-maze, and no causal chain connects the two. Moreover, the robot exploration of the maze was self generated, as it was due to an inbuilt discrepancy in the two halves of the input layer for the "reward" modality (i.e. battery level).

So, in appropriately complex systems, receptors do really meet (1) to (4) jointly, and have the same functional profile of SRs. But given that receptors *paradigmatically* fail the job description challenge, it seems that SRs (as defined by Gładziejewski) fail it too. In other words, the conjunction of (1) to (4) does not seem sufficient to identify a *representational* functional profile.

Or does it? After all, one could simply object that all what I've shown is that there are receptors that *meet* the job description challenge, namely the receptors that jointly meet (1) to (4).[27] Perhaps one could say that receptors that do not satisfy (3) and (4) actually function merely as causal mediators, but those which *do* satisfy (3) and (4) are endowed of a genuine representational status. Or perhaps one could say that I've only shown that some structures that *prima facie* qualify as

[27] Here my gratitude goes to an anonymous reviewer, to which I owe both the objection and its brilliant framing.

receptors actually are, upon closer scrutiny, SRs and thus meet the job description challenge. This would be in line with the conclusions of (Morgan 2014; Nishberg and Shapiro 2020).

I wish to resist these conclusions. In the next block, I will put forth an argument by analogy to intuitively show that (1) to (4) do not spell out a representational functional profile, in the style of both Ramsey's (2003; 2007) original analysis or receptors and Gładziejewski's (2015b; 2016) treatment of SRs.

4.2 - The argument by analogy

Consider an optical smoke detector: a simple device tasked with ringing an alarm when it detects a fire. Fires generate smoke, and, as smoke fills the air, it fills the inner chamber of the detector, refracting a beam of light on a photosensitive surface. This, in turn, closes a switch supplying electric power to an alarm. This is a simple, receptor-based, non-representational device.

Suppose one such device operates in an environment in which the typical combustion generates also *heavy smokes*: toxic fumes that tend *not* to rise even when heated, and that linger in the environment even *after* the fire has been put off. Suppose we want to enable the device to signal us their presence. It has to keep the alarm ringing when heavy smokes linger in the environment, putting it off when the heavy smokes have been dispersed by the ventilation system. This poses a challenge: heavy smokes tend (being *heavy*) to linger on the *floor*. But the optic smoke detectors are mounted on *ceilings*: "normal" smoke *rises* when heated. So the system, as it stands, is incapable of indicating the presence of heavy smokes, as they will not deflect the light beam. Indeed, the two are in no obvious causal contact.

Placing a *capacitor* between the switch and the alarm enables the system to indicate the presence of heavy smokes. When the system detects a fire, it closes the switch feeding energy to the alarm. If a capacitor is placed between the two, it will store some energy when the circuit is closed, slowly releasing it when the circuit opens (i.e. when the fire has been put off). So it will keep the alarm ringing when there is no fire but heavy smokes still linger.

Strikingly, the capacitor will *function as a receptor* of heavy smokes. This is because the amount of energy stored by the capacitor depends upon the time the circuit has been closed, which, in turn, depends on the time the fire has been raging. But so does the amount of heavy smokes. The longer the fire, the more the material combusted, and the more the material combusted, the more the heavy smokes produced. Thus, due to a common cause[28], the states of the capacitor actually indicate the amount of heavy smokes present in the environment. Observing the capacitor having in store an amount of energy v_x rises the probability that a corresponding amount of heavy smokes t_x is present in the environment.

Notice also that the capacitor satisfies (1) to (3). If the arguments given above are correct, there is at least one non-epiphenomenal structural similarity holding between it and the heavy smokes, ensuring that (1) and (2) obtain. Namely, the chronologically ordered sequence of capacitor states (v_a,v_b,...v_n) must map onto the chronologically ordered amounts of heavy smokes (t_a,t_b,...t_n). Otherwise, the system malfunctions: it either shuts up the alarm too soon (failing to indicate the presence of heavy smokes) or too late (indicating the presence of non-existing heavy smokes). Moreover, the whole system is not in any causal contact with heavy smokes, so the capacitor is *strongly decoupled* from them. Indeed, this is the reason why the capacitor is needed.[29]

A slight modification of the system enables the capacitor to satisfy (4). Suppose a second switch is placed after the capacitor, and let it be *closed* by default. Suppose further the first switch also feeds energy to a mechanical timer running a countdown. When the countdown reaches 0, the timer

[28] Notice that this is just a "ghost channel" in the sense of Dretske (1981, pp. 38-39): a set of statistically salient dependency relations between the state of two systems that are not in causal contact.

[29] Importantly, if, as Lee (2018) suggests, condition (4) can be dispensed, the capacitor *already is* a structural representation. If, however, condition (4) cannot be dispensed (as surely Gładziejewski holds), then a fairly simple modification of the system is needed.

opens the second switch, putting off the alarm. Lastly, let the circuit supplying energy to the timer be controlled by a bi-metallic strip, whose expansion opens the circuit, stopping the countdown.[30] Collectively, these components will act as a *control mechanism* for the device. Their functioning principle is simple: if, in a set amount of time, no significant increase in temperature is detected (i.e. the bi-metallic strip does not expand), then there likely is *no fire*. So, the photosensitive cell misdetected a fire, leading the capacitor to "hallucinate" heavy smokes. The system corrects the error of its receptors opening the second switch, putting the alarm off. However, if a high temperature is detected (i.e. the bi-metallic strip expands), then there likely is a fire. So, the photosensitive surface and the capacitor are working properly and the timer is stopped to keep the alarm ringing.

In this modified system, the capacitor satisfies (1) to (4), and thus, according to Gładziejewski, has the functional profile of a SR. But capacitors surely are *mere* causal mediators, and even in this (fairly complex) toy system the capacitor functions simply *as a battery* to keep the alarm ringing. It thus seems that bearing features (1) to (4) is not *sufficient* for an item to function as a representation. Hence (1) to (4) do not spell out a robustly representational functional profile. As a consequence, if SRs are defined in terms of items bearing features (1) to (4), SRs do not meet the job description challenge. Indeed, it seems to me that the same sort of worries that motivated either the rejection of the receptor notion of representation (e.g. Ramsey 2007; Orlandi 2014) or a strong suspicion about its explanatory potential (Williams and Colling 2017) emerge again. If our *most demanding* account of SRs identifies simple capacitors as representations, how could panrepresentationalism be avoided? How does such a notion of representation capture a distinctive psychological or cognitive phenomenon? Is the proposed notion of representation doing valuable explanatory work? Surely my toy system's functioning can be entirely *and transparently* understood without invoking representations. If these are reasons to reject, or be skeptical of,

[30] Notice that in thermostats bi-metallic strips are used as switches in the same way.

receptors, they will also be reasons to reject, or be skeptical of, SRs. As Nishberg and Shapiro (2020, p.2) nicely put it, SRs and receptors have a common fate.

5 - Possible objections and conclusion

I have argued that, just as receptors, SRs (as defined by Gładziejewski) do not meet the job description challenge. Here, I consider some objections to my claim.

I begin by considering an objection raised by an anonymous reviewer (to which this essay owes much). The objection is that my treatment has simply *sidestepped* the job description challenge. This objection arises because of two worries. The first concerns the call to intuitions embedded in the argument by analogy. The second is that not enough care has been taken in discussing whether truth/accuracy conditions are causally relevant in accounting for a system's success. If they are, then the job description challenge is met (the reviewer also points out that this is the argumentative strategy of Gładziejewski and Miłkowski 2017).

Let me begin by addressing the first worry. As things stand, it seems to me that calls to intuition are licensed as valid moves to address the job description challenge (see Ramsey 2007, pp. 10-11). Indeed, one of the significant aspects of the challenge is that of checking whether the term "representation", as it is used by cognitive scientists, is sufficiently "in touch" with its everyday usage. Moreover, arguments by analogy seem *sufficient* to face the challenge. This is the case, for instance, for Ramsey (2007, pp. 83-89) and Gładziejewski (2015b; 2016). Hence, if these arguments by analogy are sufficient to face the job description challenge, mine should be too. Surely, one can deny that these arguments are sufficient to face the challenge, perhaps because they rely too much on intuition.[31] However, determining the role intuitions should play in philosophical theorizing lies

[31] Notice also that such a move would undermine the claim that SRs meet the job description challenge. In fact, to the best of my knowledge, that claim has only been supported by means of arguments by analogy.

significantly outside the scope of the present essay.

What, then, about the second worry? Is checking whether the truth or accuracy condition of a posit are causally relevant to a system's success *sufficient* to determine whether the posit meets the job description challenge? I doubt this is the case. To see why, consider the following two cases.

First, the firing pin of a gun. As highlighted above, it indicates the position of the trigger, and has (by design) the *function* of doing so (firing pins are included in guns precisely *because* their state indicates the state of the trigger). Under mild teleo-informational commitment, this is *sufficient* to yield accuracy conditions to the firing pin: the firing pin accurately represents the position of the trigger if, and only if, it occupies the position it *should* occupy, given the state of the trigger. It is now possible to follow Gładziejewski and Miłkowski (2017) and wonder whether intervening on the degree to which these accuracy conditions obtain causally influences the success of the gun. And this is surely the case. The less the position of the firing pin corresponds to the position of the trigger, the more *unreliable* the gun is. In fact, the less the positions of the trigger and the pin correspond, the more the gun will fire at random. So, the accuracy conditions of the firing pin are causally relevant to the successful functioning of the gun, but I (and, I think, many others) would be hard pressed to conclude *on this sole basis* that guns are representational systems.

Consider now false, but *useful*, beliefs.[32] The research on optimism bias, for instance: "Highlights the possibility that the mind has evolved learning mechanisms to mis-predict future occurrences, as in some cases they lead to better outcomes than do unbiased beliefs" (Sharot 2011, p. R495). It is also said that the lack of such an optimism bias negatively correlates with mental health (Taylor 1989; Sharot 2011). It thus seems that certain beliefs lead a system to its success *because* they are false or inaccurate. However, it is commonly assumed that only *correct* representations non-accidentally lead to a system's success (e.g. Shea 2018, p. 10). Thus, when

[32] See also (Wiese 2017) for a case of false but useful representations at the sub-personal level of explanation.

checking whether the conditions of satisfaction of a posit lead to a system's success, one checks whether *correct* representations lead to successful behavior. But this is not the case for optimistically biased beliefs. So our verdict, in this case, should be negative: these beliefs do not meet the job description challenge and thus are *not* representations. However, optimistically biased beliefs are *beliefs* (in the ordinary sense of the term), and thus surely qualify as representations. Hence, checking whether the conditions of satisfaction of one posit are causally relevant in explaining a system's success is not *sufficient* (albeit it surely is necessary) to meet the job description challenge.

The same reviewer also urged me to discuss Rupert's (2018) defense of receptors. I cannot, due to space limitations, fully discuss Rupert's nuanced position here. Thus, I will only briefly gesture towards what strikes me as the biggest shortcoming of Rupert's position. If I understand him correctly, Rupert suggests a new positive account of receptors, able to overcome the problems raised by the job description challenge. On his view, receptors qualify as representations because, in addition to the properties discussed above, they: (a) appear in architectures which produce the distinctive *explananda* of cognitive science (i.e. intelligent behavior); (b) their contribution to the functioning of these architectures rests on their representational capacities and (c) their playing such an explanatory role partially depends on the presence, within the architecture, of distinctively cognitive forms of processing (Rupert 2018, p. 205). Clearly, conditions (a) to (c) block Ramsey's (2003; 2007) arguments. Given that firing pins of guns do not meet (a) to (c), no *genuine* analogy holds between them and genuine receptors (Rupert 2018, p. 213).

However, it seems to me that condition (c) is too underspecified. I'm frankly unsure on what counts as a "distinctively cognitive" form of processing, and Rupert never unpacks the point further (he only provides a couple of examples). Rupert (2018, p.210) seems to suggest that only forms of processing found only in cognitive architectures count as distinctively cognitive, but this is surely too strict. Predictive coding, to give but an instance, originated as a form of data compression with no *essential* link to cognitive science (Shi and Sun 2008: ch. 3; Spratling 2015; 2017); but one would be hard pressed to conclude that predictive processing is *by definition* a non-representational theory of cognition. Thus, some further clarification on what makes certain forms of processing "distinctively cognitive" seems needed.

I now turn to more local objections to my claim. One possible way to defuse my conclusion might be that of changing the relevant notion of structural similarity in (1). Perhaps second order structural resemblance is *too* cheap, and structural similarities might be better understood in terms of isomorphism or homomorphism (see Swoyer 1991; Plebe and De la Cruz 2017; Shea 2018). As these are more *restrictive* than second order structural similarity, leveraging them might prevent receptors from meeting (1) or (2) or both. But this is not the case. In every example I proposed when discussing (1) and (2) an *isomorphism* obtained. Each and every relation (v_x,v_y) among the features of the vehicle corresponded to only one relation (t_x,t_y) among the features of the target *and vice versa.* So appealing to isomorphisms does not challenge my conclusion. As isomorphisms are a special class of homomorphism, appealing to them will not alter my claim either.

Perhaps a *fifth condition* could be added to Gładziejewski's account. That might be sufficient to differentiate SRs from receptors, and to block the argument here presented. But I see no obvious candidate for this role. Moreover, Gładziejewski's account is *already* demanding: in fact, only the posits of few cognitive theories satisfy it (see Gładziejewski 2015b p. 84-85). Adding a fifth condition might thus run the risk of delivering us an account of SRs which is *too* demanding to be satisfied by any structure investigated by cognitive science.

Another possible reply might be that albeit receptors are (non-epiphenomenally) structurally similar to their targets, what *explains* the behavior success they bring about is the fact that they *indicate* their targets, not the fact that the two are structurally similar (Shea 2018 p. 130 voices precisely this concern, if I understand him correctly). Yet, if my arguments thus far are sound, the distinction between indication and structural similarity that this objection leverages is illusory (see also Morgan 2014; Nirshberg and Shapiro 2020). If my treatment of point (2) is on the right track, indication *just is* a case of structural similarity.[33] So, unless my treatment is proven wrong, or some additional reason is provided to enforce a sharp distinction between indication and structural similarity, this objection seems to have very little bite.

Can indication and structural similarity be separated? An obvious difference is that maps (and other paradigmatic SRs) do not instantiate the structural similarity with their targets *through time*, whereas receptors necessarily do. This is intuitively appealing, and might be made part of condition (1) to counter my claim. But this stipulation is unattractive for two reasons. First, it seems *ad hoc*. What *independent* reasons support the claim that the only relevant structural similarities are not time-dependent? I know of no such reason. Secondly, this stipulation comes at a high price, as many SRs posited in cognitive science actually are time-dependent. Artificial neural networks, for instance, are said to embody a structural similarity with their targets not because of how they physically are, but because of how they *dynamically react* to the inputs they are provided (Churchland 2012; Shagrir 2012; Morgan 2014). So, albeit this stipulation *would* sharply separe SRs from receptors, it seems unmotivated, and its adoption would make the relevant notion of SRs *less empirically adequate*. As pointed out above, Gładziejewski's account is *already* demanding, and making it more demanding is not necessarily helpful.

Another intuitive difference is that whereas the relations between features of maps (and other paradigmatic SRs) *obviously* represent, it is much less clear that the temporal relations between the states of a receptor I invoked when dealing with (2) represent anything. But as intuitive as this difference is, it seems to me that it cannot be *a part* of how we spell out the relevant notion of SRs, as we would circularly spell a relevant notion of representation in terms of representations.[34] Of

[33] Perhaps indication is a *special* case of structural similarity, as not all structural similarities need to involve indication (see Shea 2018, p. 138 for one example). But special cases of structural similarities still are structural similarities.

[34] Strikingly, most of the time SRs are defined in terms of representations (see Swoyer 1991; Ramsey 2007 pp 77-92. See also the insightful discussion in Shea 2018 pp. 117-118).

course, we can require the system must be sensitive to the relations among the features of the vehicle (this is part of how *exploitable* structural similarities are defined, see Shea 2018 p. 120). And we might stipulate that, when such a sensitivity is in place, then the relations among the features of the vehicle represent the relations among the features of the target (Shea 2018, p.124). But then, given that receptor-using systems typically are sensitive to the temporal relations holding between receptor states, it seems correct to conclude that these relations satisfy the given definitions, just as the relations holding between the elements of a map.

One might also argue that the toy example I proposed is insufficient to vindicate the claim that Gładziejewski's account fails the job description challenge, as the system in the toy example is an imaginary (albeit plausible) one. Noticing this creates a sharp contrast with the sort of analogies Ramsey (2003; 2007) deployed to deny receptors fail the job description challenge, as all these analogies mentioned *existing* systems. This difference should not be ignored. However, I mentioned one existing system composed just by receptors that satisfy (1) to (4), namely Bovet's control architecture for robotic agents.

One might further argue that my toy example is unfit to trivialize the notion of SR proposed by Gładziejewski because, since the capacitor clearly meets (1) to (4) it *is*, according to Gładziejewski's definition, a simple SR. Hence, what my argument has shown is, at best, that SRs can be a lot simpler than Gładziejewski originally thought.

I think that the problem with this line of argument is the following: when I added the control system, so as to enable the capacitor to meet condition (4), the *functioning* of the capacitor was not modified by the addition of the control circuit. The control circuit that I added in the final version of the system enabled the whole system to "figure out" the instances in which fires and heavy smokes where "hallucinated"[35], without thereby modifying the functioning of the capacitor. The capacitor

[35] Or, in more mundane terms, the cases in which the system malfunctions.

itself functions as it functions in the version of the system that has no control circuit, and that is thus unable to meet condition (4). And, in Gładziejewski's own view, *that* way of functioning is not representational.

Moreover, I must confess that it is not clear to me *why* adding the control circuit would transform the capacitor in a SR. The addition of the control circuit does not modify the way in which the capacitor functions, nor its overall role within the system. If the way in which the capacitor functions when the control circuit is absent is non representational (and, on Gładziejewski's account, that is true), why, then, the addition of the control circuit, which *does not* modify the way in which the capacitor operates in the system, makes its functioning representational? Surely, we can *stipulate* that it does, but why should we? Gładziejewski (2015b, pp. 78-79) simply asks us to accept condition (4) without offering any substantial[36] justification for it. And the reasons as for why Bickhard (1993; 1999; 2009) deems error detection *necessary* for genuine representations seem to be fairly alien to the theoretical commitments of cognitive science. For instance, Bickhard greatly stresses the fact that, in order for some internal state to count as a representation, it must be a representation *for the organism* "consuming" it. But such a requirement is by no means necessary in the theoretical framework of cognitive science; indeed, many paradigmatic examples of representations (e.g. syntactic trees, Marr's 2 ½-D sketches) are not representations for the organism consuming them. And, in fact, cognitive scientists do not simply introspect them or somehow intuit their presence: they *posit* them as explanatory tools deemed necessary to account for the functioning of our cognitive system and the production of intelligent behavior.[37] Now, I do not wish to *simply* rule out (4) as a necessary condition. Perhaps it is. But if it

[36] Here, by substantial I mean "non pragmatic". The pragmatic rationale behind (4) is fairly straightforward: (4) makes the account of SRs more robust, protecting it from trivializing counterexamples. Notice further that Gładziejewski (2016) simply takes error detection for granted, without offering any substantial justification for it. In fact, his own brief discussion of error-detection might be leveraged as an argument *against* (4). If as Gładziejewski (2016) insists, one *cannot* determine whether one's own pragmatic failures are determined by the presence of misrepresentations or by the misapplication of *correct* representations, one is not able to detect *representational* errors. Rather, one able to detect *pragmatic failures*, which might be due either to representational errors or to misapplications of correct representations.

[37] To be fair to Bickhard, it is important to point out that the idea that genuine representations are representations for whole organisms is not the sole reason as for why he deems error-detection a necessary condition. The prospect of

is, then there must be a way to spell out why error detection is necessary. As far as I can see, this reason has not yet been spelled out.

Lastly, one might object that my argument is a *reductio* of the job description challenge. The reasoning behind this objection seems to be as follows. Any successful naturalistic account of representation *should* cast representations (more precisely, their vehicles) as causal mediators, whose causal role is systematically related to their semantic properties.[38] Now, it is widely assumed that, in the case of SRs, the relevant semantic properties *just are* properties of the vehicle; namely the features that make the vehicle structurally similar to a relevant target (see O'Brien 2015; Williams 20017; Williams and Colling 2017; Lee 2018). And, if the relevant structural similarity is exploited, these properties are *guaranteed* to be the properties that are causally relevant to the system's behavior (Gładziejewski and Miłkowski 2017). So, SRs seem to be exactly the kind of posits that *should* meet the job description challenge. If, as I've argued, they do not meet it, then there is probably something wrong with the job description challenge itself. Maybe it is too demanding.[39] Maybe it still hangs to a non-naturalistic conception of intentionality and content. At any rate, if *no* candidate representational posit is able to meet the job description challenge, then the problem is likely to be the job description challenge itself, rather than any candidate representational posit in question. Compare: if *all* the students *always* fail their tests, we would be inclined to think that the problem is the tests, rather than the students.

However, I do not think that my argument entails a *reductio* of the job description challenge. To start, my argument, if correct, only shows that SRs do not meet the job description challenge.[40] It is silent on whether other types of representations meet it. Maybe they do or maybe they don't, but

avoiding the problems of content indeterminacy seems to play an important role too. I do not see, however, how acknowledging this challenges my point: it still seems to me correct to say that, in the theoretical framework cognitive science offers, genuine representations do not *need* to be representations for entire organisms.

[38] Notice that I do not actually dispute this claim. Above I have denied *only* the fact that the accuracy conditions of a posit are causally relevant to a system's success is *sufficient* for that posit to qualify as a representation. But this clearly does not exclude that having causally relevant semantic properties is *necessary* in order for a posit to qualify as a representation.

[39] Importantly, Egan (2020, pp. 43-45) seems to articulate precisely this idea.

[40] Given the fairly widespread assumption that receptors do not meet it.

adjudicating this issue lies significantly beyond the scope of this essay.

Moreover, alongside SRs, there is another kind of representation that is widely supposed to meet the job description challenge, namely *input-output representations* (see Ramsey 2007 pp. 68-77).[41] These are representations of the values and arguments a computational system is supposed to compute upon. For instance, if *really* feedforward artificial networks acting as recognition models compute the probability of a label given (i.e. conditioned over) an input vector, they will need to manipulate vectors (arrays of variables or values) and probabilities (a value ranging from 0 to 1), which are mathematical objects. Since physical systems cannot manipulate (at least *prima facie*) mathematical objects, they must manipulate something that stands-in for them, and that represents, in an appropriate way, the relevant mathematical objects. These are input-output representations.[42]

As far as I can see, my argument does not change this state of affairs: if really input-output representations meet the job description challenge[43], they meet it whether my argument is correct or not. And, if input-output representations meet the job description challenge, the job description challenge *can* be met. It would thus be false that *all* students *always* fail the test.

But what if it turns out that input-output representations fail the job description challenge too?

[41] Some readers might be shocked by this statement, as SRs and input-output representations are sometimes taken to be identical (e.g. Sprevak 2011). But to identify SRs with input-output representations seems to me a mistake. For one thing, input-output representations are *essentially* linked to computational accounts of cognition, whereas structural representations are not (e.g. Tolman 1948). Moreover, structural representations are *necessarily* structurally similar to their targets. But input output representations need not *necessarily* structurally resemble what they represent. In fact, they might be arbitrary symbols.

[42] I believe some "historical" clarifications are in order. As Ramsey (2007) presents them, input-output representations *need not* (albeit might) represent mathematical objects. The claim that the relevant representations involved in computational processes represent mathematical objects (namely, the arguments and values of the functions computed) is, to the best of my knowledge, a claim articulated independently by Frances Egan in a number of publications (e.g. Egan 2014). Recently, Ramsey (2020, p. 72-73) has declared that Egan's account captures, in a more sophisticated way, his notion of input-output representations. Here, I'm following Ramsey (2020).

[43] Importantly, this at least partially depends on the theory of computational implementation one endorses. Here, I will stay neutral on the issue. Notice, however, that many (I suspect the majority of) theories of computational implementation try to avoid *pancomputationalism*; namely, the view that any complex physical system implements a number of (or perhaps all) computations (see Searle 1992 for the pancomputationalist challenge; see Copeland 1996; Scheutz 1999; Rescorla 2014; Piccinini 2015 for some ways to defuse it). The important point to notice, for present purposes, is this: that many accounts of computational implementation *would not* deem sufficient, for a physical system to compute a function, that the causal goings-on internal to the system systematically "mirror" the transition between computational states. Thus, if the idea common to these accounts is correct, input-output representations need to be *more* than causal mediators allowing a system to "march in step" with some relevant computable function.

Wouldn't that show that there is something wrong with the job description challenge? Maybe yes. Yet notice: I'm not claiming that input-output representations fail the job description challenge. The claim that input-output representations fail the job description challenge might be a *reductio* of the challenge, but that claim is not defended here, and so the argument offered in *this* essay is, as far as I can see, no *reductio* of the challenge.

Moreover, even if it turns out that *no* candidate class of representational posits meets the challenge, the charge of *reductio* strikes me as excessive. Discovering that no representational posit meets the job description challenge would be a *reductio* of the challenge only given a strong prior representationalist assumption. But one could also have some prior inclination towards antirepresentationalism, and conclude that the job description challenge yielded a correct result in each case. Now, I do not wish to adjudicate here whether one should be inclined more towards representationalism or antirepresentationalism. I will only notice that, insofar representationalism and antirepresentationalism are not taken to be *a priori* truths, but rather empirical research programs (or at least the conceptual bedrocks of empirical research programs), we should be open to revise our representationalist or antirepresentationalist inclinations.[44] Thus, even if it were true that no candidate representational posit meets the job description challenge (a strong claim that this essay does not support), that fact alone would not *necessarily* lead to a *reductio* of the challenge. It might also lead to a revision of one's representationalist commitments.

Having deflected a number of objections to my argument, it seems to me correct to conclude that, according to our most demanding account of them, SRs do not meet the job description challenge. Or so, at least, I argued.

[44] This claim is typically made by philosophers leaning towards antirepresentationalism (e.g. Chemero 2009; Ramsey 2017; Hutto and Myin 2020). But the rationale behind it works both ways: if antirepresentationalism is *not* an *a priori* truth, one ought to revise one's own antirepresentationalist commitment in the light of the relevant empirical evidence.

Founding: This work has been funded by the PRIN Project "The Mark of Mental" (MOM), 2017P9E9N, active from 9.12.2019 to 28.12.2022, financed by the Italian Ministry of University and Research. Early versions of this essay have been presented in a number of conferences; namely the OZSW Graduate conference in theoretical philosophy (Tilburg University 19/12/2019); The workshop *Representation in Cognitive Science* (Ruhr-Universität Bochum 3-4/02/2020) and the 10th European Congress of Analytic Philosophy. I wish to thank the audience of these conferences for helpful comments. A special thanks goes to Krys Dolega, for his encouraging words. I wish also to thank both anonymous reviewers for their insightful comments on earlier versions of this essay.

References

Anderson, M. & Chemero, T. (2013). The problem with brain GUTs: conflation of different senses of "prediction" threatens metaphysical disaster. *Behavioral And Brain Sciences*, *36*(3), 204-205.

Anderson, M. & Chemero, T. (2019). The world well gained. In M. Colombo, E. Irvine, M. Stapleton (Eds.). *Andy Clark and His Critics* (pp. 161-173). New York: Oxford University Press.

Artiga, M., & Sebastián, M. A. (2018). Informational theories of content and mental representation. *Review of Philosophy and Psychology*, https.//doi.org/10.1007/s13164-018-0408-1.

Bickhard, M. H. (1993). Representational content in humans and machines. *Journal of Experimental and Theoretical Artificial Intelligence*, *5*, 285-333.

Bickhard, M. H. (1999). Interaction and representation. *Theory and Psychology*, *9*, 435-458.

Bickhard, M. H. (2009). The interactivist model. *Synthese*, *166*(3), 547-591.

Bovet, S. (2007). *Robots with Self-Developing Brains*, Dissertation, University of Zurich https://www.zora.uzh.ch/id/eprint/163709/1/20080298_001884101.pdf. Accessed 25 February 2020.

Bovet, S., & Pfeifer, R. (2005a). Emergence of delayed reward learning from sensorimotor coordination, *Proc. IEEE/RSJ Int. Conf. On Intelligent Robots and Systems*, https://doi.org/10.1109/IROS.2005.1545085

Bovet, S., & Pfeifer, R. (2005b). Emergence of coherent behaviors from homogeneous sensorimotor coupling, *ICAR '05 Proceedings 12th International Conference on Advanced Robotics,* https://doi.org/10.1109/ICAR.2005.1507431

Braitenberg, V. (1984). *Vehicles: Experiments in Synthetic Psychology*, Cambridge, MA.: The MIT Press.

Brooks, R. (1999). *Cambrian Intelligence*, Cambridge, MA.: The MIT Press.

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

Cheney, D. L., & Seyfarth, R. M. (1985). Vervet monkey alarm calls: manipulation through shared information?. *Behavior*, *94*(1-2), 150-166.

Churchland, P. M. (2012). *Plato's Camera*, Cambridge, M.A: The MIT Press.

Clark A. (1993). *Associative Engines*, Cambridge, MA.: The MIT Press.

Clark, A. (1997). The dynamical challenge, *Cognitive Science*, *21*(4), 461-481.

Clark, A. (2010). Memento's revenge: the extended mind, extended. In R. Menary (Ed.), *The Extended Mind*, (pp. 43-66). Cambridge, MA.: The MIT Press.

Clark, A. (2013a). *Mindware. An Introduction to the Philosophy of Cognitive Science* (2nd edition), New York: Oxford University Press.

Clark, A., & Toribio, J. (1994). Doing without representing?, *Synthese*, *101*,3, 401-431.

Clark, A., & Grush, R. (1999). Towards a cognitive robotics, *Adaptive Behavior*, *7*(1), 5-16.

Copeland, J. B. (1996). What is computation?. *Synthese*, *108*(3), 335-359.

Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory*, New York, Wiley & Sons.

Downey, A. (2018). Predictive processing and the representation wars: a victory for the eliminativist (via fictionalism), *Synthese*, *195*(12), 5115-5139.

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

Dretske, F. (1988). *Explaining Behavior*, Cambridge, MA: The MIT Press.

Dretske, F. (1994). The explanatory role of information. *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*. *349*(1689), 59-70.

Egan, F. (2014). How to think about mental content. *Philosophical Studies*, *170*(1), 115-135.

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

Eliasmith, C. (2005). A new perspective on representational problems, *Journal of Cognitive Science*, *6*(97), 97-123.

Fodor, J. (1989). Semantics: Wisconsin style. In J. Fodor (1990). *A Theory of Content and Other Essays*, (pp. 31-49). Cambridge, MA.: The MIT Press.

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

Gallistel, C.R., & King, A. P. (2010). *Memory and the Computational Brain*, Oxford: Wiley-Blackwell.

Gładziejewski, P. (2015a). Action guidance is not enough, representations need correspondence too: a plea for a two-factor theory of representation, *New Ideas in Psychology 40*, 13-25.

Gładziejewski, P. (2015b). Explaining cognitive phenomena with internal representations: a mechanistic perspective, *Studies in Logic, Grammar and Rhetoric, 40*(1), 63-90.

Gładziejewski, P. (2016). Predictive coding and representationalism, *Synthese*, *193*(2), 559-582.

Gładziejewski, P., & Miłkowski, M. (2017). Structural representations: causally relevant and different from detectors, *Biology and Philosophy*, *32*(3), 337-355.

Goodman, N. (1969). *The Language of Arts*. London, Oxford University Press.

Gorman, R. P., & Sejnoski, T. J: (1988). Analysis of hidden units in a layered network trained to classify sonar targets. *Neural Networks*, *1*(1), 75-89.

Gosche, T., & Koppelberg, D. (1991). The concept of representation and the representation of concepts in connectionist models, in W. Ramsey, S. P. Stich, D. E. Rumelhart (eds.), *Philosophy and Connectionist Theory* (pp. 129-163). New York, Rutledge.

Grush, R. (1997). The architecture of representation, *Philosophical Psychology*, *10*(1), 5-23.

Harvey, I, Husbands, P., & Cliff, D. (1994). Seeing the light: artificial evolution, real vision, in D. Cliff, P. Husbands, J. A. Meyer & S. W. Winson (eds.), *From Animals to Animats 3* (pp. 392-401). Cambridge, MA.: The MIT press.

Harvey, I., *et al*. (1997). Evolutionary robotics: the Sussex approach, *Robotics and Autonomous Systems*, *20*(2-4) 205-224.

Haugeland, J. (1991). Representational genera, in n W. Ramsey, S. P. Stich, D. E. Rumelhart (eds.), *Philosophy and Connectionist Theory* (pp. 61-91). New York, Rutledge.

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

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

Husbands, P., Harvey, I., Cliff, D. (1995). Circle in the round: state space attractors for evolved sighted robots, *Journal of Robotics and Autonomous Systems*, *15*, 83-106.

Hutto, D., & Myin, E. (2020). Deflating deflationism about mental representations. In J. Smortchkova, K. Dolega, T. Schlicht (Eds.), *What are Mental Representations?* (pp. 79-100). New York: Oxford University Press.

Kiefer, A., & Hohwy, J. (2018). Content and misrepresentation in hierarchical generative models. *Synthese*, *195*(6), 2397-2415.

Lee, J. (2018). Structural Representation and the two problems of content, *Mind & Language*, *34*(5), 606-626.

Levittin, J. Y.; Maturana, H. R.; McCulloch, W. S. & Pitts, W. H. (1959). What the frog's eye tells the frog's brain, *Proceedings of the IRE*, *47*(11), 1940-1951.

Lyre, H. (2016). Active content externalism. *Review of Philosophy and Psychology*, *7*(1), 17-33.

Maris, M., & Schaad, R. (1995). The didactic robots, *Techreport No. IIF-AI-95.09, AI Lab, Department of Computer Science, University of Zurich.*

Maris, M., & te Boekhorst, R. (1996). Exploiting physical constraints: heap formation through behavioral error in a group of robots, in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 1655-1660) Piscataway, NJ: IEEE Press.

Miłkowski, M. (2013). *Explaining the Computational Mind*, Cambridge, MA.: The MIT Press.

Miłkowski, M. (2017). Szaleństwo, a nie metoda. Uwagi o książce Pawła Gładziejewskiego "Wyjaśnianie za pomocą reprezentacji mentalnych". *Filozofia Nauki*, *25*(3(99)), 57-67.

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

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

Moser, E. I., Kropff, E., & Moser, M. B. (2008). Place cells, grid cells, and the brain spatiotemporal representation system, *Annu. Rev. Neuroscience*, *31*, 69-89.

Nieder, A., Diester, I., & Tudusciuc, O. (2006). Temporal and spatial enumeration processes in the primate parietal cortex. *Science*, *313*(5792), 1431-1435.

Nirshberg, G., & Shapiro, L. (2020). Structural and Indicator representations: a difference in degree, not in kind. *Synthese,* https://doi.org/10.1007/s11229-020-02537-y.

O'Brien, G. (2015). How does the mind matter? Solving the content-causation problem. In T. Metzinger, J. M. Windt (Eds.). *Pen MIND*: 28(T). Frankfurt am Main: The MIND Group. https://doi.org/10.15502/9783958570146

O'Brein, G., & Opie, J. (2004). Notes towards a structuralist theory of mental representations, in H. Clapin; P. Staines & P. Slezak (eds.), *Representation in Mind: New Approaches to Mental Representaion* (pp. 1-20). Oxford: Elsevier.

O'Keefe, J., & Nadel, L. (1978). *The Hippocampus as a Cognitive Map*, New York: Oxford University Press.

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness, *Behavioral and Brain Sciences*, *24*(5), 939-973.

Orlandi, N. (2014). *The Innocent Eye,* New York: Oxford University Press.

Pezzulo, G. (2008). Coordinating with the future: the anticipatory nature of representation, *Minds and Machines 18*(2), 179-225.

Pfeifer, R., & Bongard, J. (2007). *How the Body Shapes the Way we Think*, Cambridge, MA: The MIT Press.

Piccinini, G. (2015). *Physical Computation: a Mechanistic Account*. New York: Oxford University Press.

Plebe, A., De la Cruz, M. V. (2017). Neural representations beyond "plus X", *Mind and Machines*, *28*(1), 93-117.

Ramsey, W. (1997). Do connectionist representations earn their explanatory keep?, *Mind & Language*, *12*(1), 34-66.

Ramsey, W. (2003). Are receptors representations?, *Journal of Experimental & Theoretical Artificial Intelligence*, *15*(2), 125-141.

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

Ramsey, W. (2015). Untangling two questions about mental representation, *New Ideas in Psychology*, *40*, 3-12.

Ramsey W. (2017). Must cognition be representational?, *Synthese*, *194*(11), 4197-4214.

Ramsey, W. (2019). Maps, models and computational simulations of the mind, in M. Sprevak & M. Colombo (Eds.), *The Routledge Handbook of the Computational Mind* (pp. 259 – 271). New York.: Tylor & Francis.

Ramsey, W. (2020). Defending representation realism. In J. Smortchkova, K. Dolega, T. Schlicht (Eds.), *What are Mental Representations?* (pp. 54-78). New York: Oxford University Press.

Ramstead, M. J. D., Kirchhoff, M. D., & Friston, K. (2019). A tale of two densities: active inference is enactive inference, *Adaptive Behavior*, https://doi.org/1059712319862774

Rescorla, M. (2014). A theory of computational implementation. *Synthese*, 191(6), 1277-1307.

Rupert, R. (2018). Representation and mental representation. *Philosophical Explorations*, *21*(2), 204-225.

Scheutz, M. (1999). When physical systems realize functions. *Minds and Machines*, *9*(2), 161-196.

Searle, J. (1992). *The Rediscovery of the Mind.* Cambridge, MA.: The MIT Press.

Segundo-Ortin, M., & Hutto, D. (2019). Similarity-based cognition: radical enactivism meets cognitive neuroscience, *Synthese*, https://doi.org/10.1007/s11229-019-02505-1

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

Shannon, C. E., & Weaver, W. (1949). *The Mathematical Theory of Communication*, Urbana, IL.: University of Illinois Press.

Sharot, T. (2011). The optimism bias. *Current Biology*, *21*(23), R491-R945.

Shea, N. (2007). Content and its vehicles in connectionist systems, *Mind and Language*, *22*(3), 246-269.

Shea, N. (2014). VI – Exploitable isomorphism and structural representation, *Proceedings of the Aristotelian Society, 114*(2,2), 123-144.

Shea, N. (2018). *Representations in Cognitive Science*, New York: Oxford University Press

Shepard, R. N., & Chipman, S. (1970). Second order isomorphism of internal representations: shapes of states, *Cognitive Psychology*, *1*(1), 1-17.

Shi, Y. Y., & Sun H. (2008). Image and Video Compression for Multimedia Engineering. Fundamentals, Algorithms and Standards (2nd ed.). New York: CRC Press.

Smortchkova, J., Dolega, K., & Schlicht, T. (2020). Introduction. In J. Smortchkova, K. Dolega, T. Schlicht (Eds.), *What are Mental Representations?* (pp. 1-26). New York: Oxford University Press.

Spratling, M. W. (2015). Predictive coding. In D. Jaeger, R. Jung (Eds.), Encyclopedia of Computational Neuroscience, (pp. 2491-2494), New York: Springer.

Spratling, M. W. (2017). A review of predictive coding algorithms. Brain and Cognition, *112*, 92-97.

Sprevak, M. (2011). Review of Representation reconsidered. *The British Journal of Philosophy of Science*, *62*, 669-675.

Swoyer, C. (1991). Structural representation and surrogative reasoning. *Synthese*, *87*(3), 449-508.

Taylor, S. (1989). *Positive Illusions. Creative Self-Deception and the Healthy Mind*. New York: Basic Books.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189-208.

Vold, K.; & Schlimm, D. (2020). Extended mathematical cognition: external representations with non-derived content. *Synthese*, *197*, 3757-3777.

Wiese, W. (2017). Action is enabled by systematic misrepresentations. *Erkenntnis*, *82*(6), 1233-1252.

Williams, D. (2017). Predictive Processing and the Representation Wars. *Minds And Machines*, *28*(1), 141-172.

Williams D. (2018). Predictive minds and small-scale models: Kenneth Craik's contribution to cognitive science, *Philosophical Explorations*, *21*(2), 245-263.

Williams, D., & Colling, L. (2017). From symbols to icons: the return of resemblance in the cognitive science revolution, *Synthese*, *195*(5), 1941-1967.