Mental representations as structural representations. Developing the notion

Marcin Miłkowski
Paweł Gładziejewski

(work in progress, please do not cite)

Abstract: Recently, the idea that mental representations are internal models, or structural representations (S-representations), has been steadily gaining popularity. In the present paper, we want to provide some much needed details to the claim that mental representations are S-representations. We present an account of structural similarity relation which preserves the notion that similarity is asymmetric and sidesteps the problems associated with defining similarity in terms of morphisms. Using the neomechanist theory of explanation and the interventionist account of causal relevance, we provide a precise interpretation of the claim that structural similarity serves as a 'fuel of success', i.e., relation that is exploitable for the representation using system. When developing our view, we discuss crucial differences between S-representations and (purported) indicator or detector representations, showing that – contrary to some claims made in the literature – there is an important theoretical distinction to be drawn between the two. Lastly, we provide answers to some worries that could be raised in the context of our proposal.

Key-words: S-representations, structural representation, mental models, mental representation, representationalism, antirepresentationalism, Hard Problem of Content

1. Introduction

Antirepresentationalism has been one of the major recent trends in theorizing about the mind. One argumentative strategy that modern antirepresentationalists employ is what we might call the 'trivializing' strategy. Instead of (or in addition to) developing new theories of cognition that do without the notion of representation, proponents of the trivializing strategy attempt to show that some of the most prevalent existing notions of mental representation are not suited to do the theoretical and explanatory jobs that are expected of them. Perhaps most spectacularly, the idea that representations are covariance-based indicators or detectors has been subjected to this sort of trivializing attack. In influential criticisms of the representations-as-indicators notion, Ramsey (2007) has argued that purported indicator (detector) representations boil down to mere causal mediators (and thus are not representations at all), whereas Hutto and Myin (2012) have argued that covariance, including functional covariance, cannot constitute intentional content construed as truth or accuracy conditions.

One way that representationalists might want to oppose attempts to trivialize representations is to defend the *status quo*, e.g. to argue that detectors are representations after all. But perhaps a different reaction could prove more justified and fruitful. Perhaps representationalists should treat the trivializing sort of antirepresentationalism as an opportunity to develop, strengthen, and indeed reform our mainstream understanding of what representations are, such that the resulting new notion is no longer subject to the trivializing arguments. In fact, we think that something like this sort of reaction is starting to take shape in the literature. This is because in parallel to antirepresentationalism, another trend in theorizing about representation has been gaining momentum recently, wherein people move from seeing mental representations as indicators or detectors to seeing them as internal models of the world (Bartels 2006; Gładziejewski 2015b; Gładziejewski 2015c; Miłkowski 2013; O'Brien 2014; Opie & O'Brien 2004; Shagrir 2012; Shea 2014; Ramsey 2007; Rescorla 2009). Here, the model is understood as internal, manipulable structural representation, i.e. a representation based on the structural similarity between the representation itself and its target.

We think that this move towards the model-based account of representation is exactly the way forward for representationalism. Here our job will be clarificatory. The present paper aims to elucidate the claim that structural representations (henceforth: S-representations) are paradigmatic cases of mental representations, and provide reasons to think that S-representations are not reducible to indicator, detector, or receptor representations.

Put more precisely, our aim here is threefold. First, we want to make some advancements when it comes to two key issues in the theory of S-representations. In particular, we aim to provide an account of the structural similarity relation which connects the S-representational vehicle with what it represents (Section 2). At the same time we want to precisely spell out the idea that similarity acts as relation that is exploitable for some larger system or mechanism – a relation on which a larger system or mechanism can actively depend in its interactions with environment. For this purpose, we provide an interventionist account of what it means for the similarity relation to be exploited in a cognitive mechanism (Section 3). Second, following Cummins (1996), Ramsey (2007), and Burge (2010), we want to develop our proposal in a way that shows that there is a crucial distinction between representations proper – S-representations in particular – and mere indicators or detectors. In this, we will oppose authors who have argued or assumed that all S-representations are indistinguishable from indicators or receptors, or that they are subject to the same trivializing arguments (Morgan 2014; Hutto & Myin 2012). In particular, we want to show that structural similarity relation is different from covariance (Section 2), as well as that there are crucial differences between mechanisms that exploit structural similarity and ones that simply have the ability to adaptively respond to the changes in the environment (Section 6). Third, we want to answer some possible worries about our proposal, namely ones that point to the fact that (allegedly) it is not similarity itself, but rather just structural properties of the representational vehicle that are of relevance for the workings of cognitive systems (Section 4); that the relationship between similarity and success is more nuanced than we give justice to (Section 5); and that S-representations as we construe them do not have contents or accuracy conditions (Section 7).

2. Covariance vs. structural similarity

There are several reasons why similarity seems better suited to explicate the structural relationship between the vehicle of representation and its potential targets than mere morphisms. But to claim that, we need to show that similarity is indeed different from indication. The most important fact is that similarity is irreducible mere indication or simple mapping. We will first show that representations may be richer than indications. Then we show that structured representations may be actually also more selective, or contain less information than mere indicators. We also argue that indication is a symmetric relation while similarity need not be. Finally, we answer some popular objections against the usefulness of the notion of similarity, and show that they apply equally to other kinds of structural notions such as isomorphism.

Here's one reason to believe that similarity is irreducible to indication. Let's suppose, for the sake of argument, that all structurally rich representations are reducible to a set of indicators. Let's suppose further that these indicators correspond to yes/no answers to questions, or that they indicate the presence of entities or their properties.[1] A set of such decisions would fully define the structure of the representation. But it has been shown that one cannot determine even which logical connectives and operators are used in a language

[1] Notice that we don't include indicators of relations between indicators themselves, as standard accounts of indication do not include them. It's their absence that underlies our argument.

based on assent, dissent and suspension of judgments of speakers (Berger 1980). For example, one may assent to a sentence of a form $(p\ v\ q)\ v\ r$ without assenting to its logically equivalent sentence $p\ v\ (q\ v\ r)$ but the table that lists the judgments of speakers would remain the same. This is even more visible if you accept the intuitionistic logic: the inference from $\sim\sim p$ to p is not valid but assent of $\sim\sim p$ is automatically the same as assent of p. In other words, the logical structure of statements cannot be discovered using only choices between assent and dissent. But if you think that any structure, including at least logical structure, can be reduced to a set of indicators, then intuitionist connectives are encoded by a list of indicators which correspond to assent or dissent. But they cannot. Hence, we have a *reductio*.

One might object to this argument by saying that this intuitionistic logic is a special, controversial case. But the argument applies even to mere associativity of disjunction in $(p\ v\ q)\ v\ r$. Furthermore, if all mental representations are *just* sets of indicators, we would be unable to represent the rule of associativity in logic, and parentheses would not be required in the standard infix notation. (Obviously, Reverse Polish Notation requires no parentheses but it represents associativity with operator ordering). So an unordered set of indicators cannot replace any structure of mental representations. We need to retain the internal structure in the representation, and simple mappings, or unordered bags of indicators, won't do. There might be internal structure that is implied in the use of representation but never present in the target domain. Arguably, for example, most parentheses in logic do not have straightforward correspondence to any state of affairs but play a major role in inferential processes by structuring the precedence of operators. Hence, we argue that representations may be richer than indicators, and that their structure need not be fully expressed in the target domain. A cartographic map of mountains usually includes their names but that does not mean that these names are actually written on mountains. Moreover, including these names does *not* make this map less accurate.

Some kinds of similarity may be thus richer in information than mere isomorphism or homomorphism. Obviously, one *can* define similarity in terms of mappings, as O'Brien and Opie (2004) do. But that kind of similarity will have several properties that are not desirable from our point of view. Not only both the street and my percept may have properties that do not stand in any correspondence (this much O'Brien and Opie clearly assume); more importantly, the function of representation requires that my percept is more selective and arguably in some ways richer than the street itself. Full sensory information need not be fully retained for us to perceive the street; our representations may contain highly compressed and strategically distorted information. Note also that even if there are one-to-one mappings (of the selective kind that O'Brien and Opie assume) between the properties of the parts of the neural system and the represented targets (which we find highly dubious), we can treat these as standing in similarity. Just because (one-to-one or one-to-many) mappings are limiting cases of similarity, then the most general account should refer to similarity. This similarity should also be an *asymmetric* relationship just because the structure of vehicles may not fully correspond to the target, just like names of the mountains or lakes on the map are not similar to mountains or lakes themselves.

Let's now focus on the case where representations contain less information than indicators, and consider that the level of detail can be naturally accounted for in terms of the degree of similarity. All representations, including mental representations, are selective, and one may want to compare different levels of selectivity. For example, a cartographic map of a city may be very detailed to include names of monuments and restaurants, and another one highly abstract. Depending on the use of the map, we might need just the abstract one to get a general gist of the direction we should go. Importantly, we might start from a highly indefinite representation of something, and then gather additional information to have a more detailed representation of that very thing; for example, I may spot a face and to make sure if that is a familiar face, take another look. Both representations have different level of detail. The notion of similarity allows to account for differences between both representations and their information content. Technical accounts of similarity include some measures of similarity that account for gradation of similarity judgments.

The same point is related to current developments in computational neuroscience. For example, Chris Eliasmith claims that occurrent temporal patterns in the neural networks do not simply encode other patterns; they are so called Semantic Pointers that compress information contained in neural structures in a lossy way (Eliasmith 2013).[2] Lossy compression is irreversible but necessary for tractability and for further use of such pointers in inferential processes. As Eliasmith argues, the Semantic Pointer Architecture can implement symbolic operations usually considered to be problematic for neural networks (Fodor & Pylyshyn 1988); interestingly, the logical operations such as inferences are realized in a noisy way, which makes them a good but imperfect approximation of classical symbolic operations. Yet such irreversible lossy compression is automatically considered to be *less* similar, hence non-veridical, by symmetrical mapping-based accounts of similarity. But symbolic operations on non-compressed neural information are biologically implausible, so we should allow for

[2] Note that *Semantic Pointer* is a technical term used by Eliasmith to describe the crucial part of his hypothesis about the cognitive functioning of the brain.

partial similarity; our cognitive apparatus does not consider symbolic knowledge to be automatically non-veridical even if it is more abstract than perception (note that this claim can be endorsed also by proponents of highly embodied accounts of cognition).

In general, owing to the complex interactions and transformations, there is no simple one-to-one or many-to-one mapping between the structure of the target and the structure of the vehicle; some properties of vehicles may be caused by internal processes only to make further processing possible. But we still can say that there is an informational connection between these vehicular properties and the target, which can be accounted for in terms of similarity.

Let's now turn to the last consideration. We have argued above that because some structure of vehicles need not correspond to the structure of targets, the most general account of similarity that holds between both is asymmetric. Asymmetric relationships can account for richer structure in representations, and for richer relationships between individual representations than mere indicators. Therefore, one simply cannot derive that asymmetric similarity obtains just in case indication obtains.[3] Indication is a symmetrical relationship, so Morgan's (2013) claim that *all* S-representations are reducible to indicators is not true, even if *some* simple S-representations could *also* be considered to be simple indicators just because they are based on one-to-one mapping (see also Section 6).

There are multiple accounts of asymmetric similarity. For our reasons, it's enough to refer to the account probably best known to philosophers, namely Tversky's (1977) account of feature similarity. This account has been also used to account also for the modeling relationship in philosophy of science (Weisberg 2013). We do not claim that all mental representations stand in exactly this relationship to their targets. It is not a task for philosophers of science to establish empirical facts; we only claim that it's plausible to believe that a version of asymmetric similarity relationship holds between different kinds of neural representations and their targets but one needs solid empirical evidence to establish which kind exactly. Our strategy is rather to look for a most general account that would fit possible informational relationships already established in experimental and theoretical literature.

Many philosophers eschew all talk about similarity since Goodman's (1972) sustained criticism of the notion. They prefer to talk about morphisms instead; for this reason, accounts

[3] However, the fact that similarity reliably obtains between A and B suffices to show that some other indication relation obtains between A and B. It is not guaranteed that it has exactly the same informational content.

of S-representation usually refer to the notion of isomorphism (Cummins 1996; Ramsey 2007) or homomorphism (Bartels 2006). This relationship is supposed to obtain between the vehicle of the representation and its target (if there is any). However, all structural notions, including isomorphism or homomorphism, are prone to the same set of problems that plague structuralism in general. A simple mapping, be it defined in terms of similarity, or in terms of morphism, between structure and reality cannot avoid indeterminacy. The problem was noticed by M.H.A. Newman (1928), who carefully showed that Bertrand Russell's causal theory of perception fails precisely because one cannot define the relationship between the structure of the world and the world itself without avoiding triviality or incoherence. The problem raised by Newman still appears in many guises, which include Goodman's troubles with similarity. To wit, structuralists face the following problem: How to show that a structure corresponds to all and only entities that have that structure? Newman showed that under an intuitive, set-theoretic interpretation of relations, all sets with proper cardinality correspond to just *any* structure in a one-to-one fashion. That leads to a massive indeterminacy. Hence, we will call this *indeterminacy* problem. If the problem remains unsolved, the similarity-based account of representation will trivialize the notion of representation, as anything might be taken to correspond structurally to something else, even if you define the correspondence relation precisely, as Russell did.

To make claims about similarity meaningful, one needs also to specify the candidate arguments of the similarity relationship, and the type of the similarity you want to establish. Even a raven always resembles a writing desk in some respect (Carroll 1900). If you want to reject a claim that a raven is similar to a writing desk, you need to make sure which properties of the raven can be included, and which properties of the writing desk ever come to play. If you talk about color properties, then a black raven does not resemble a white writing desk; if you include just any properties, then ravens are just as spatiotemporal as writing desks. So one desideratum for a theory that makes use of structural notions is to determine the arguments of the structural relationship. For example, Ramsey in his recent (2007) book proposes to include only the properties that have explanatory value for an explanation of a cognitive task to avoid problems in Cummins's account of meaning. Though this is somewhat vague, the aim of his constraint is to avoid indeterminacy of claims about isomorphism; and it can be straightforwardly applied to similarity as well. In the next section, we show that similarity can be causally relevant and this relevance can be rather straightforwardly accounted for using the popular interventionist account of causality. Similarly, the relata of this relation should be causally relevant for an explanation at hand; in other words, respects of similarity and the relata of similarity relationship can be fixed using fairly standard explanatory tools. More on that to come in Section 3.

Tversky claims that the similarity relation must fulfill several conditions. One of them is the matching condition, which says how similar two entities are is a function exactly of the set of their common features and the two respective sets of their distinctive features (the set of features that belong to the first but not to the second, and the set of features that belong to the second but not to the first). For all a, b $\in$ D,

$S(a, b) = F(A \cap B, A - B, B - A)$,

where S is the similarity relationship, F some ternary real-valued function, A the features of a, and B the features of b.

The monotonicity condition requires the following:

$S(a, b) >= S(a, c)$ if (1) $A \cap C \subseteq A \cap B$; (2) $A - B \subseteq A - C$; (3) $B - A \subseteq C - A$.

Similarity increases with the addition of shared features or with the removal of their distinctive features. There are several models that fulfill Tversky's condition but the exact details need not concern us here (for details see the original paper). What is important is that Tversky proved that relationships that satisfy his conditions will have a similarity scale and a salience scale; again the details are not important for this paper.

It would be preposterous to claim that mental representations are similar to their targets in that they have exactly the same properties. This version of a pictorial account of representation, which is as old as the ancient metaphor of memory as ring on the wax, is not plausible at all. When I see a red tomato with green leaves, my neurons do not turn red and green. However, the features that the neural systems represent are probably much more complex than simple sensory qualities. Thomas Metzinger (2003) argues that sensory systems represent highly relational properties in the first place because Ganzfeld experiments – homogenous stimulation of a sensory modality – do not produce homogenous percepts but a complete lack of perception. For us, this means that sensory systems represent relational properties of targets, so in this case, the similarity relationship holds between relational properties of observable entities and relational properties of the neural vehicles, usually transformed in a highly non-linear way.

Famously, Roger Shepard claimed that "second-order" isomorphisms between (a) relations among alternative external objects and (b) relations among internal representations are more important for mental representation than first-order ones that hold between (a) an individual object and (b) its corresponding representation (Shepard & Chipman 1970). This might be true but this is not our position. We think that internal representations of considerable interest are ones that are richly structured themselves, not *only* the ones that are related to other representations (which is implied by Shepard). We also do not think that only external objects are represented, which is probably an unintended part of Shepard's account. In addition, there actually are some cases in which geometric relational properties are represented in the nervous system geometrically. Grid cells in the hippocampus are the case in point (Bechtel 2014; Redish 1999). Instead of supposing that there is a particular kind of property or entity that is always represented by all cognitive systems in the same way, which seems less than plausible, we claim that if there is a rich representation, there will be similarity relationship. This relationship holds between neural (or artificial, if you accept the computational theory of mind) vehicles and whatever the system represents. What the system represents is to be established empirically, not in an a priori fashion. In our view, sensory systems, for example, usually represent relational properties in a complex way, so the sets A and B in Tversky's conditions should contain relational properties that are explanatorily relevant for the task that is best explained by referring to the similarity relationship.[4]

In their criticism of representationalism, Hutto and Myin (Hutto & Myin 2012) have claimed that there is just one notion of semantic information that a naturalized semantics can refer to, namely the notion of information-as-covariance (as defined, for example, (Dretske 1982)). According to them, information-as-covariance is the naturalist's 'best bet' when it comes to specifying the relation that grounds or gives rise to intentional content. But by now it should be apparent that this a limited, and in fact misleading perspective. Not only do Hutto and Myin miss a notion of semantic information as defined by control relationships (MacKay 1969; Miłkowski 2015), they also do not do justice to the notion of information-as-similarity, implied by accounts of S-representation. Let us define information-as-similarity: d's being F carries information I about w's being G just in case F is similar to G. This general definition can be easily adapted to different kinds of similarity, including Tversky's feature-based asymmetric similarity. In the latter case, F is understood to include also the contrast classes mentioned above. One can also straightforwardly quantify the level of informativeness not in terms of probabilities, as it is usual with covariance approaches, but in terms of similarity

[4] We do not claim that there are only perceptual representations of external objects. There might be representations of the inner bodily states, and representations of highly abstract objects that do not exist. Whether a given neural system is able to represent such abstract objects is a matter for empirical enquiry, not for armchair speculation.

levels, where 1 stands for indistinguishability. The technical details need not concern us here (we are not using these properties of information-as-similarity further in the paper).[5]

Why is this notion important? For several reasons. One, which should be already evident, is that higher levels of similarity imply that representation is more informative about the structure of the target. It is the structure of the target which is not explicitly modeled in the notion of information-as-covariance. For example, imagine a rodent with a seriously disturbed hippocampus, which still lawfully covaries with the structure of its environment. As long as the relationship is lawful, it's highly informative. But it may still be less informative in terms of information-as-similarity than an analogous representation in the hippocampus of a healthy rodent. Another reason is that one doesn't not refer to reliability of covariance to establish informational relationship. The covariance may be quite fragile (based on a somewhat invariant causal generalization but not a law of nature) but the representation may still hold quite a lot of information about something. (The third reason is that this notion avoids the Bar Hillel-Carnap paradox – the contradictions are not informative at all, not maximally informative – but this is a story for another occasion). In other words, this notion is not only naturalistic; it's much more intuitive than information-as-covariance.

The information-as-similarity is implied in many accounts of information flow (Barwise & Seligman 1997) and modeling (Rosen 2012). In particular, Barwise and Seligman use the notion of infomorphism to talk about the information flow; it obtains between two classifications. A classification is simply assigning certain tokens to a type or not. Once you have classified objects in a set, you have objects that belong to a type, and the ones that don't. Now there might be a relationship, called infomorphism, between two classifications A and B and tokens in both classifications that allows to reconstruct A from B and B from A. Our

[5] A semi-historical side note might be in order here. At least on some accounts (Dretske 1986), some detectors or indicators do have satisfaction conditions. According to Dretske, an indicator that d is G means that w is F. In other words, Dretske frames indication in predicative terms: an entity's having a certain property means that some other entity has another property. Hence a property is being ascribed to an entity just in case some other entity has another property. Such ascriptions have satisfaction conditions; for real indicators, they are always true. However, Dretske usually analyzes simple examples, and rarely goes beyond mere binary registrations that could answer a Yes/No question (is w F? Yes if d is G, No if d is not G). For that reason, it was claimed that indicators only indicate the presence of referents (Cummins & Roth 2012; Ramsey 2007). This is not entirely true; indicators do indicate the presence of properties of referents, according to Dretske, and one of his paradigmatic examples is a speedometer that involves a speed scale (which makes it more informational than a simple binary indicator). However, Dretske failed to give a detailed account of how structure is important for the contents of his indicators. For this reason, our position goes beyond his teleofunctional account.

account is very similar but more general in that instead of mappings, we talk of similarity: infomorphism is an extreme case of a similarity relationship. This might be because we're more interested in how information is operated upon and transformed rather than with the mere flow of information. However, we believe that the intuition behind infomorphisms (and many accounts of modeling) is similar to what we express as information-as-similarity: the more structure of A you can discover by looking at the structure of B, the more information B bears about A.

Before we proceed, it needs to be noted that despite concentrating on the relation of similarity, we do not mean to deny that there are causal relationships between the vehicles of S-representations and their targets. For example, we do not claim that the structure of vehicles of representations is not caused (in the ontogenetic or phylogenetic sense). Of course it is. But neither causation nor covariance define information-as-similarity. Information-as-similarity is not, generally, reducible to information-as-covariance. This fact alone is enough to reject the claim that S-representations are mere indicators. They may be both indicators and S-representations (nothing prevents that) but one should not confuse two different informational relationships: similarity and indication.

3. Exploiting structural similarity: a mechanistic-interventionist account

On the view we are advocating, the mere existence of structural similarity between two entities is by no means sufficient to confer on one of those entities the status of a representation. S-*representations* only come into play when a cognitive system depends, in some nontrivial sense, on the similarity relation in its engagements with the environment. This way, we follow Peter Godfrey-Smith (1996) and Nicholas Shea (2007; 2014) in thinking that the correspondence (in the case of S-representations, the structural similarity relation) between representation and whatever it represents should be construed in terms of fuel for success or a resource that enables organisms to "get things done" in the world. In other words, similarity should be understood as a relation that is *exploitable* for some larger representation-using system. Only when the structural similarity is usable and thus has value for a cognitive system can it give rise to representation. Although this idea strikes us as quite intuitive and is by no means new in the literature, to our knowledge it has never been elaborated in detail. Thus, we now want to address the question what it means exactly for structural similarity to be an exploitable relation. In particular, we will try to clarify this idea in the context of purely subpersonal S-representations of the sort that we could find inside a mechanical system such as a human brain.

Let us start by taking a closer look at the basic, rather commonsensical intuition that underlies the idea that is at issue here. The intuition seems relatively straightforward. Consider an external, artifactual S-representation such as a cartographic map. It seems natural that we can at least sometimes explain someone's success at navigating a particular territory by pointing to the fact that the person in question used an accurate map of this territory (and vice versa, we can explain someone's navigational failure by citing the fact that the person in question used a map that is inaccurate). Claims like "Ann found a train station in Cairo because she was using an accurate map of the city" or "Ann failed to find a train station in Cairo because she was using an inaccurate map of the city" can be truly explanatory. But what does the map's accuracy that explanations of this sort invoke consist in? It seems that it has to do with facts regarding whether the map in question bears a structural similarity to the territory – or at least whether it bears enough similarity to enable successful navigation. Users of cartographic maps owe their success, in some explanatorily illuminating sense, to the similarity relation that holds between the spatial structure of the representation and the spatial structure of the territory it represents (analogously, the failures can be due to the lack of similarity between the representation and what is represented). This link between similarity and success generalizes to all S-representations, including, we claim, *the ones that are not interpreted by full-blown human beings*.

Now, in virtue of what exactly can explanations of success that cite facts regarding similarity be true? We want to propose that what is crucial here is that facts about similarity relation can be *causally relevant* to facts about success. On our view, the structural correspondence can quite literally cause the representation-user to be successful at whatever she (or it) is using the representation for; and lack of structural correspondence can cause the user to fail at whatever she (or it) is using the representation for. Explanations that invoke S-representations should thus be construed as causal explanations[6] that feature facts regarding similarity relation as an explanans and facts regarding success or failure as an explanandum. To exploit structural similarity in this sense is, very broadly speaking, to use a strategy whose success is causally dependent on facts regarding structural similarity between the

[6] More precisely, at least in the case of *cognitive-scientific* representational explanations, they should be seen as *mixtures* of causal and mechanistic explanations, as they point to *interlevel* causal relations holding between the activity of mechanism's components and some capacity of the mechanism as a whole. More on this subject to come in the main text.

representational vehicle and what is represented. Crucially, here lies the solution the indeterminacy problem mentioned in Section 2. Of, course, everything – including the vehicles of S-representations – could be taken to structurally resemble many things in the world. But on our view, only those similarities give rise to *representation* which are causally relevant for the workings of larger agents or mechanisms. That is, a thing may resemble many, perhaps all (spatiotemporal) things in this or that respect, but, as matter of fact, only a small subset of those similarities will be such that whether this similarity to that particular thing holds (or the degree to which it holds) is causally relevant to success of some agent or mechanism. What thus demarcates similarities *per se* from similarities that are representationally relevant is the causal role played vis-à-vis success and failure. When talking about S-representations, only those similarities count which are exploitable.

Our treatment mentions two concepts that are in need of clarification, especially when applied to internal, subpersonal representations: the notion of success/failure (for which the similarity relation is causally responsible), and the notion of causal relevance. We will now concentrate on each of those notions in turn. Let us start with success and failure.

The idea that *human agents* can succeed or fail at whatever they use S-representations for seems straightforward enough and we will not dwell on it here. But how to understand success/failure in the case of internal, subpersonal representations of the sort that are of interest to us here? We propose that a fruitful way of unpacking this is by looking at the problem through the lens of the prominent neomechanist theory of explanation, as applied to cognitive-scientific explanation (Bechtel 2008; Craver 2007). Neomechanists see the cognitive system (say, the embodied brain) as a collection of mechanisms. Each mechanism is a set of organized components and component operations which jointly enable the larger system to exhibit certain capacity. Mechanisms in this sense are at least partly individuated functionally, that is, by reference to the capacity that they give rise to – they are essentially mechanisms *of* this or that cognitive function (mindreading, motor control, attention, perceptual categorization, spatial navigation, etc.). Components and operations derive their functional characterization from the function of a larger mechanism they are embedded in. That is, the function of a component is determined by an operation such that it is through the performance of this particular operation that the component in question contributes to a capacity for which the larger mechanism is responsible (see Craver 2007). This is why, say, the function of heart as a component part of a mechanism responsible for blood circulation lies in its operation of pumping blood, and not in its emitting rhythmic sounds; it is the former, and not the latter operation through which the heart contributes to blood circulation.

In this paper, we assume that internal S-representations (more precisely, the vehicles of S-representations) can be treated as components of cognitive mechanisms, and are targets of various cognitive operations, such as off-line reasoning or belief updating. Each mechanism equipped with an S-representation as its component part underlies certain capacity. S-representations construed as mechanism components owe their functional characterization to how they contribute to a capacity that the larger mechanism is responsible for. What we mean by this is, essentially, that *structural similarity* between the representation and what it represents is what contributes toward the mechanism's proper functioning. To put it more precisely, any mechanism responsible for some capacity C which includes an S-representation as its component can *fail* to realize or enable C as a result of the fact that the component in question is not (sufficiently) structurally similar to some (represented) part of the environment; and analogously, when the mechanism *succeeds* at realizing on enabling C, this is at least in part due to the fact that this component is (sufficiently) structurally similar to some (represented) part of the environment. So the structural similarity relation is causally relevant to success/failure because the ability of any S-representation-involving mechanism to perform its function depends on whether or, rather, on the degree to which there is structural similarity between the representation and something in the environment external to the mechanism. In simpler terms, success and failure are treated here as success or failure at contributing to some *function of a mechanism*.

We now turn to the question of what it means for the similarity relation to be causally relevant to success (or failure) thus understood. Here we aim to make use of James Woodward's (2003, 2008) popular interventionist theory of causal relevance[7]. It is beyond the scope of the present discussion to present Woodward's theory in details, so a rough sketch will have to suffice. The core idea behind the interventionist view is that claims of causal relevance connect two variables, say, X and Y (the variables and different values they take admit many ontological interpretations, i.e. they could stand-in for properties, processes, events, states of affairs, etc.). What it takes for X to be causally relevant to Y is that appropriate interventions into X (i.e. interventions in the values of X) are associated with changes in Y (i.e. the values of Y). Slightly more technically:

[7] Following Carl Craver's (2007) work, we take it that the interventionist account of causal relevance can be reconciled with the neomechanist view of explanation. More specifically, we assume that the component's contribution to the capacity of the mechanism as a whole can be understood in terms of this component's being causally relevant for the capacity in question, where causal relevance is construed along the interventionist lines.

> **(M)** X causes Y if and only if there are background circumstances B such that if some (single) intervention that changes the value of X (and no other variable) were to occur in B, then Y would change (Woodward 2008)

The intervention in question can be helpfully understood as an experimental manipulation of X in controlled settings, although Woodward's theory does not require human agency to be involved in establishing causal relations – any change of the value of X could potentially count as an intervention, even one that is not dependent at all on human action. Importantly, there are certain conditions that an intervention must meet in order to establish a causal connection between X and Y. For example, the intervention must not change the value of Y through any causal route except the one that leads through X (e.g. it must not change the value of Y directly or by directly changing the value of a variable that mediates causally between X and Y) and it must not be correlated with any causes of Y other than X or those that lie on the causal route from X to Y.

By employing the interventionist view, we can now understand the causal relevance of similarity for success in the following way. The structural similarity between the representational vehicle and whatever is represented is causally relevant for success in virtue of the fact that interventions in the similarity relation would be associated with changes in the successfulness of whatever action or cognitive function that is based on, or guided by the representation in question. That is, manipulations on the similarity relation would also be manipulations on the ability of the representation-user – be it a human being or some internal cognitive mechanism – to be successful at whatever she or it is employing the representation for.

To make this proposal more precise, let us reconstruct both relata of the causal relation as variables. The variable X corresponds to the similarity relation between the vehicle and what is represented. It would probably be a gross simplification if we treated X as a binary variable, with one value corresponding to the existence, and the other to the lack of similarity. Luckily, the account of structural similarity introduced in the preceding sections allows for this relation to come in degrees (as similarity scale or salience scale). This way we can treat X as capable of taking a range of values $\{X_1, X_2, ..., X_n\}$, where each increasing value corresponds to an increased degree of similarity between the vehicle and what is represented. Therefore, between the lack of any similarity and a complete structural indistinguishability there is a range of intermediate possibilities.

What about Y, the variable that corresponds to success/failure? As mentioned, we treat internal S-representations (which are of interest for us here) as components of mechanisms, where each mechanism underlies a particular cognitive capacity. So far as we can see, S-representations could turn out to feature in a diverse set of mechanisms which give rise to a diverse set of cognitive functions, like motor control and motor planning, perceptual categorization, mindreading, decision making, etc. Now, it seems rather intuitive that cognitive systems can be more or less *effective* at realizing each such capacity: they can perform better or worse at motor control and planning, perceptually categorizing objects, attributing mental states, making decisions, etc. In this sense, we can treat the variable Y as corresponding to degrees of success of the mechanism in question at enabling an effective performance of a given capacity. Increasing values of $Y = \{Y_1, Y_2, ..., Y_n\}$ would correspond to increasing degrees of success thus understood. But what sorts of values can we have in mind exactly? Here we want to remain as open as possible. We think that any scientifically respectable way of measuring success can do. For example, the success could be measured by the average frequency of instances of certain level of performance at some cognitive task, or to probability of certain level of performance at some task, or a distribution of probabilities of possible levels of performance at some task, etc. The details will always depend on the sort of capacity in question, as well as on the experimental paradigm used to test or measure this capacity.

We may now formulate our thesis as follows. For similarity to cause success, interventions in the value of X (which corresponds to the degree of structural similarity between the representational vehicle and what it represents) should result in systematic changes in the value of Y (which corresponds to the degree of success of the mechanism that makes use of an S-representation in performing its mechanistic function or capacity). In particular, by intervening in X so that its value increases, we should increase the value of Y; and by intervening in X so that its value decreases, we should decrease the value of Y.

A following empirical illustration should illuminate this proposal. In the philosophical literature, hippocampal spatial maps in rats have been proposed as a good example of an internal S-representation (Ramsey, 2015; Rescorla, 2009; Shea, 2014). The rat's hippocampus is thought to implement an internal map of the spatial layout of the environment, encoded in a Cartesian coordinate system. According to this hypothesis, the co-activation patterns of so-called place cells in the hippocampus correspond to the spatial structure of rat's environment (Shea, 2014). That is, there exists a structure-preserving mapping from co-activation relationships between place cells (roughly, the tendency of particular cells to show joint activity) and the metric relations between locations within the environment. This hippocampal map constitutes a component of a cognitive mechanism that is responsible for the rat's ability to navigate its environment (Craver 2007). The rat's capacity to find its way within the environment, even when there is no possibility of orienting in space using external cues or landmarks, depends on the fact that it has an internal mechanism equipped with a map of the terrain. This capacity for navigation is usually tested by verifying the rat's ability to find reward (food) within a maze in which the animal has no reliable access to external orientation points (see (Bechtel 2014; Redish 1999) for review).

As has been already argued in the literature, spatial navigation using hippocampal maps is an instance in which the structural similarity between the map and the territory is being actively exploited by the organism (Shea, 2014). Similarity serves a resource that the rat depends on in its dealings with problems that require spatial navigation. Our proposal provides what we think is a clear and precise interpretation of this claim. The map-world similarity is causally relevant to rat's success at finding its way in the environment. This means that we could manipulate rat's capacity to navigate in space by intervening in the degree to which its internal map resembles structurally the (relevant part of) the environment. We know for example that rats are quite efficient at constructing and storing separate maps for particular mazes (Alme et al. 2014). We may imagine an experiment in which we place the rat in a previously-learned maze and then intervene on the co-activation structure of place cells in a way that distorts (i.e. decreases) the structural correspondence between the map and the maze to a particular degree.[8] If the similarity is really being exploited, then intervention of this sort should decrease the rats ability to navigate the particular territory, and we should be able to observe and measure this decrease by investigating the change of rat's performance at finding rewards in the maze. What is more, rat's navigational capacity should be reduced to the degree which is in proportion to the degree to which we decreased the similarity relation between its internal map and the spatial structure of the maze. And crucially, our intervention should change rat's performance *only insofar as it constitutes an intervention on the similarity relation as such* (see next section for more details).

4. Is *similarity* really what's causally relevant?

[8] Interventions with this sort of precision go beyond what is currently technologically achievable. But this not a major problem, as Woodward's account of causation allows interventions to be in-principle possible only. And there is no reason to think that intervention of the sort we are postulating is not possible in-principle.

A following issue might well be raised in the context of our mechanistic-interventionist treatment of the notion of exploitable similarity. One could wonder whether it is really the similarity relation *as such* that is causally relevant for success. Notice that it is impossible to perform an intervention on the similarity relation in any other way than by intervening on the structure of at least one of its relata (that is, the representational vehicle or what is represented). But this invites a worry. Wouldn't it be much more parsimonious to simply state that what is causally relevant for success are structural properties of the vehicle and/or the relevant part of the environment? After all, it is by intervening in either of them that we can manipulate success. Why bother attributing the causal role to the *similarity* relation itself? For example, to change rat's performance at navigating mazes, it will suffice to intervene on the structure of the hippocampal map. Why not simply say that it is the structure of the map (the representational vehicle) that is causally relevant for rat's success at spatial navigation? Why treat the *relation* between the map and the environment as causally relevant?

We think that our proposal is left unscratched by this worry. It is really the similarity itself that matters. To show this, let us contrast the effects of intervening just on the structure of the vehicle with the effects of interventions in the similarity relation itself[9].

For starters, it would be helpful to distinguish interventions that change the way some cognitive system *acts* (or behaves, or cognizes) from interventions that change the *successfulness* of its actions (or behaviors, or cognitions). The change of action can, but does not have to change the successfulness of the organism at whatever it is doing. If the change of the way the system acts is accompanied with an appropriate change in the external environment, the successfulness of action can stay the same (e.g. we could change the rat's behavior in a maze without changing its ability to find food if the maze changes accordingly). At the same time, one and the same manipulation of action can change the successfulness of the organism either by increasing it or decreasing it – again, the direction influence will depend on properties of the environment (e.g. on the structure of the maze that the rat is traversing). So there is no context-free, one-to-one correspondence between action and success. The reason for this is that success and failure in the sense we are using are essentially *ecological* categories. They co-depend both on what a given system is doing, *and* on the world within which it is doing it. This, by the way, is precisely why organisms need S-

[9] Importantly, the analysis to follow could be easily extended to contrast the effects of interventions on similarity with the effects of interventions on the other relatum, that is, the structure of the represented part of the environment. Here we concentrate exclusively on the representational vehicle's structure just for the sake of simplicity.

representations: they need internal structures that correspond to the world (or function in a way that is supposed to establish such correspondence) so that they can take a course of action that is ecologically valid given the actual state of the world.

Notice now that by concentrating solely on the properties of the representational vehicle, we would completely miss the point just made. Surely, interventions on the structural properties of the vehicle (e.g. the hippocampal map) would change the cognitive system's *actions* (e.g. rat's behavior when places in a maze) in a very precise ways. That much is not debatable. But as we have just seen, manipulating action is not the same as manipulating success. Because of this, the effect that the structure of the vehicle has on action does not imply that the same sort of relationship exists between the vehicle's structure and *success*. It is impossible to say how manipulating the vehicle's structure (viz. organism's action, be it practical or cognitive) will change success independently of what the facts about the environment are; or more precisely, independently of what the facts about the relation of structural similarity between the vehicle and the environment are. In other words, interventions on the vehicle's structure change the success *only insofar as they change the degree of similarity between the vehicle and some part of the environment*. They increase success if they increase the structural fit between the vehicle and the environment. They decrease success only if they decrease the structural fit.[10] And they do not change the success if they do not bring about any change in the structural fit. In any case, what the success depends on is not just the facts about the vehicle, but the facts about the relation of structural resemblance. Of course, again, the only way to intervene on similarity is by manipulating the relata. But it is just wrong to conclude from this that the similarity relation itself is not what is causally relevant here.

To further underline out point, let us formulate it using some technicalities of Woodward's account of causal relevance. Suppose that the independent variable X corresponds *not* to the similarity relation between the vehicle and the environment, but to purely vehicular-structural properties of the representation. More precisely, imagine that each value of X corresponds to a different potential structural pattern of the vehicle, regardless of its relationship to anything outside the mechanism. The dependent variable Y remains the same, i.e. it measures the degree of success. Now, there are certain constraints that Woodward

[10] Here, we idealize a little. Actually, as we argued, highly simplified, hence less similar representations may be much more tractable for organisms, and thus contribute to their success. However, they cannot be fully dissimilar to their targets. Hence, there is a specific range of similarity typical for a given representation, and if similarity drops below that level, the success of the action will drop as well (though the relationship between two drops need not be linear). See also the next section for a more detailed treatment of this subject.

(2003) puts on any scientifically respectable causal relationships. Two of them are relevant for our present purposes. First, interventions should not simply effect *some* changes in Y. Rather, the relation between X and Y should be systematic in that we should be able to establish which values of X correspond to which values of Y. Second, the relationship between X and Y should be stable, viz. it should hold across a wide range of different background conditions. But notice that neither of those constraints is met on the interpretation of X and Y that we are now considering. First, because of the reasons we mentioned above, there is no clear mapping from values of X to values of Y, which prevents the relationship between those variables from being systematic in the relevant sense. Setting X at some value could well increase the value of Y, decrease it or even not change it all. Second, the relation between X and Y is by no means stable. In fact, it is fundamentally instable because of how dependent it is on the states of the environment. It is not possible to say how manipulation of X will change the value of Y independently of the state of the world. Again, one and the same manipulation of X (e.g. setting the structure of the spatial map in the hippocampus) could bring about drastically different results depending on external circumstances (e.g. depending on the spatial structure of the maze that the rat navigates).

Both Woodward's constraints are however met if we go back to our original view and consider the variable X to correspond to the degree of *similarity* between the representational vehicle and the represented part of the environment. The relation between X and Y is then both systematic and stable. It is systematic because we can map increasing values of X onto increasing values of Y (however, see section 'Limited representations for limited beings'). And it is stable at least in the sense that it cannot be broken down by changes in the relevant (represented) part of the environment. After all, the value of X partially depends precisely on what the environment is like[11]. Overall, we think that those considerations provide conclusive reasons to think that in a mechanism that makes of S-representation, it is really the relation of structural similarity that is causally relevant to success.

5. Limited S-representations for limited beings

[11] Of course, even in this case the causal relation between the two variables only holds within a certain set of background circumstances. For example, X can be said to be causally relevant to Y only if other (nonrepresentational) parts of the representation-involving mechanism are working properly. Nonetheless, we take it that the conditions under which *this* causal relation remains stable are wide enough for it to be of scientific value.

Another issue that one might take with our proposal is that our view of the relationship between similarity and success is too simplistic. It seems that good S-representations often, maybe even always resemble relevant parts of the world only partially. As already mentioned before, good maps never mirror the territory all of its detail; instead, they are simplified, selective, and even distorted (consider the map of London Underground which only preserves topographic, but not metric properties of the spatial layout of railway lines it represents). The same seems to apply to all S-representations, including subpersonal ones that could feature as components in cognitive mechanisms. Real biological agents have limited storage and computational powers. Furthermore, their cognitive functions have ecologically determined temporal limitations – cognition needs to be (appropriately) fast. It is doubtful that real-life cognitive systems would turn out to have the resources and time to build internal S-representations that come even close to mirroring the structural complexities of the world. It seems natural to expect that similar to cartographers, natural selection promotes simplified, incomplete, just-good-enough S-representations, as well as ones that introduce useful 'fictions' and distortions. The reason for this, of course, is that S-representations that resemble whatever they represent too much become excessively complex themselves. Thus, too much resemblance hampers, or even precludes success.

As should be apparent from Section 1, we have no gripes with the general notion that the role of similarity vis-à-vis success is nuanced in this sense. However, we believe that this can be reconciled with our interventionist treatment of the similarity-success relation. We think that the observation made above complicates, but does not cancel this relationship. Notice that the concession that too much similarity can hamper success does not mean or imply that similarity has no effect whatsoever on success. A map does not have to be a literal copy of the territory to serve its purpose, but it still has to preserve some nontrivial portion of the territory's structure in order to do its job right. Although the resemblance does not have to be complete (i.e. it does not have reach structural indistinguishability), some degree of resemblance is required for S-representation to work properly. S-representations that share no structural properties with what they represent cease to serve their purpose. The lesson to be learned here is not that similarity is functionally irrelevant, but simply that *too much* similarity can render the S-representation inefficient at serving its purpose.

This observation can be expressed using our preferred interventionist framework. Suppose that increasing values of variable X correspond to increasing similarity between the vehicle and what is represented, and the increasing values of variable Y correspond to increasing success. To accommodate the observation that too much similarity can (and usually does) harm success, we may simply say that although there is positive causal relation between X and Y (i.e. manipulating X by increasing its value increases the value of Y), it only holds within a limited range of values of X. For simplicity we may suppose that the relation holds from the lowest value of X to some specific (much) larger value, but it disappears when X exceeds this value. That is, once the value of X exceeds certain level, X becomes only randomly associated with Y; or the relationship can even get inverted, i.e. increasing the value of X may begin to decrease the value of Y. We may suppose that this breaking of the relationship between X and Y is due to the existence of a trade-off between the befits of increasing X and the costs that are associated with it. That is, when the representation reaches a certain degree of similarity to what it represents (viz. the representational vehicles reaches a certain level of complexity), the cognitive and temporal resources needed to make use of it (costs) start to exceed the success that the representation in question underlies (benefits). This way we can give justice to the fact that limited cognitive systems can only afford limited S-representations, but without giving up on the idea that similarity serves as a fuel of success.

6. S-representations vs. representations based on response-selectivity

There is yet another issue which makes the notion of exploitable similarity not quite as straightforward as it might seem at first. Namely, there are cases where it is not obvious at all whether it is really the relation of similarity between the vehicle and what is represented that is being exploited by some larger mechanism. More precisely, some structures that do not strike us as S-representations on the first sight, nonetheless seem, on closer inspection, to function on the basis of the similarity relation after all. If this is to, then one cannot help but wonder whether exploitable similarity is cheaper and more ubiquitous than it should, which could in turn blur the distinction between S-representations and other types of representation, or even render the notion of S-representation trivial (Shea 2007, 2013; Morgan 2014).

Consider the notorious thermostat. In order to do its job, the thermostat needs to function in a way that is adapted to changes in ambient temperature. For this reason, it is equipped with a bi-metallic strip whose shape reliably reacts causally to variations in the temperature. The strip in turn switches the thermostat's furnace in a way that is required to keep the ambient temperature at a certain level. It is usually claimed that *if* it is even justified to treat it as an representation (which is far from uncontroversial in itself, see Ramsey 2007), the bi-metallic strip counts as, at most, a detector or an indicator of some state of affairs. However, on closer inspection, it turns out that causation is not the *only* relation that connects the strip to ambient temperature. They are *also* related by way of structural similarity. Namely, there is a structure-preserving mapping from the relational pattern of bi-metallic strip's possible shapes to the pattern of possible variations in the ambient temperature (O'Brien 2014; see also Morgan 2014). Perhaps, then, we should regard the thermostat as a device that makes use of an S-representation after all?

It might seem that the problem can be easily resolved by investigating which of the two relations between the bi-metallic strip and the ambient temperature (i.e. the strip's being reliably caused by or it being structurally similar to variations in temperature) is *causally relevant* for the functioning of the thermostat. That is, although both relations hold, perhaps only one of those is actually being exploited by the larger mechanism. To resolve this, we could employ the interventionist outlook on causation. From this perspective, the question would be: interventions in which of those two relations reliably change the thermostat's effectiveness at regulating the temperature? But alas, things are not so easy. It seems that any intervention on one of those relations can equally count as an intervention in the other one. That is, it seems impossible to intervene in one without intervening in the other as well. By disrupting the reliable causal relation between the bi-metallic strip and the changing of ambient temperature, we also disrupt the degree to which the changes in the shape of the strip mirror the changes in the temperature – and vice versa. It seems impossible to test for exploitable similarity without thereby testing for exploitable reliable causation between the vehicle and what is represented. So is the bi-metallic strip an instance of S-representation or not? Or perhaps, contrary to what is often assumed (e.g. Cummins & Poirier 2004; Grush 1997; Opie & O'Brien 2004; Ramsey 2007), there is no difference between S-representations and mere detectors after all? Maybe the conditions of being an S-representation are so liberal that any detector can easily count as an S-representation, so there is no philosophically or cognitive-scientifically interesting distinction to be made here (Morgan 2014)?

We think that a partial resolution to the problem at hand has been recently formulated by Gerard O'Brien (2014). O'Brien points to the fact that in the thermostat case, there *is* an important counterfactual asymmetry between the role played by reliable causation and structural similarity. Namely, 'curvature correspondence without causal covariation (e.g., where a mere correlation exists) would still generate the appropriate behaviour [of the thermostat], but causal covariation without curvature correspondence (e.g., where the bimetallic strip heats up but maintains its shape) wouldn't' (O'Brien 2014, p. 8). To put this in interventionist terms, there is a possible intervention that could settle the debate after all. We could redesign the thermostat so that the states of the bi-metallic strip are correlated with the changes in the ambient temperature – in a way that preserves the structural mapping between the two – but there is no causal relation between them (i.e. after the intervention, the correlation in question still exists, but is established by means other than direct causal relation). For this reason, it is structural similarity (‘correspondence’) that gets the upper hand when it comes to the question of what exactly enables the bi-metallic strip to control the functioning of the thermostat.

Although we take O’Brien’s proposal to be a significant advancement in resolving the issue, we still think that the problem reappears nonetheless. Notice that what is necessary for the bi-metallic strip to perform its function is for it to be *adaptively selective in responding* to changes in the ambient temperature. This relation of ‘adaptive selectivity’ holds between A and B when, (1) A and B can both take on a number states or values or states, (2) each particular state of A gets instantiated (the bi-metallic strip’s takes a particular shape) in response to a particular state of B getting instantiated (e.g. the ambient temperature setting at a certain level), (3) A causally affects the behavior of some system S (e.g. it changes the behavior of the furnace), (4) the resulting behavior of S is adaptive or functional because of the fact that states of A get instantiated in response to particular states of B (e.g. the changes in furnace behavior enable the thermostat to regulate the temperature in appropriate way because the bi-metallic strip changes its curvature in response to changes in the ambient temperature). The counterfactual asymmetry that O’Brien points to in the passage cited above results from the fact that in order for A to be adaptively selective with respect to B, it is not necessary for A to be *causally* related to B. It will suffice for A and B to change in response to some common cause, or just to be contingently correlated with each other (see O’Brien 2014). *This* is why the thermostat could still do its job even if the causal connection between the bi-metallic strip and ambient temperature was broken.

But notice how all this makes the notion that the thermostat makes use of an S-representation problematic again. After all, should we think that *adaptive selectivity* is the same as (exploitable) structural similarity? It does not look like it. The crucial thing to note here is that representations based on structural similarity do not *need* to change ‘in response’ to anything in the world – at least, as far as we can see, on any useful or illuminating interpretation of what ‘in response’ could possibly mean exactly in this context. Just think of a person who manipulates an interactive map (say, of the sort that you can find on Google Maps) in order to plan a *future* trip, or even to consider a travel purely *counterfactually*. In cases like this, map’s usefulness depends on the fact it resembles the territory: it is because of this resemblance that manipulating the map actually gives one access to the way the world will be or could have been. But notice that in cases like these, the map does its (S-)representational job even though *nothing changed in the world such that the manipulation preformed on the map could possibly count as being performed 'in response' to that change*. In fact, we propose that part of what constitutes an S-representation precisely is the fact that its ability to function does not require it to change 'in response' to anything. To put it in a broad philosophical context, on our view, representations based on adaptive selectivity (if they even count as 'representations' at all) lie in the domain of receptivity understood in, roughly, Kantian sense: they are a matter of having the ability to be passively influenced by the world in certain cognitively useful ways. S-representations, on the other hand, lie in the domain of – again, very roughly speaking – Kantian spontaneity, in that they can be, so to speak, freely deployed as part of endogenous activity of the cognitive system. S-representations are actively used and manipulated for cognitive purposes; they are active, not (just) reactive.

To clarify the abovementioned idea, let us first go back to the thermostat case again. Does it count as an-S-representation-using system after all? Are its workings based on exploitable structural similarity or maybe just on adaptive selectivity? It seems like we have another case of counterfactual asymmetry here, but this time it goes *against* S-representational reading of thermostat's working. Remember that when we consider attributing the thermostat with an S-representation, we are led by the fact that there is a structure-preserving mapping from bi-metallic strip's shapes to variations in ambient temperature. So suppose that we perform an intervention which systematically breaks this relation of similarity. For example, we install some mechanism which reliably changes the bi-metallic strip's shape so that the previous mapping does not hold anymore. Could the thermostat still work properly? The answer is: yes, as long as after the change, the bi-metallic strip remains to be *selectively reactive* to changes in the environment in a way that is *adaptive* for the larger mechanism (i.e. it still changes the behavior of the furnace in appropriate ways).[12] *All* that is required for the thermostat to work is to have an internal mediator which reliably reacts to changes in the environment and generates a change in behavior that are adaptive given the circumstances.[13]

[12] For this to happen, of course, we would have to tweak the way the bi-metallic strip affects the furnace.

[13] Of course, regardless how we re-shuffle the connection, we will end up with *some* new mapping between the bi-metallic strip and the temperature. But this mapping is just nothing more than a trivial and necessary side-effect of adaptive selectivity. Every time A is adaptively selective with respect to B, we will be able to map states of A and relations among them onto states of B and relations among them. The structural similarity is, so to speak, epiphenomenal here: it results from the existence of adaptive selectivity, but by itself it is not doing any

At the same time, it seems impossible to break the relationship between the thermostat's workings and adaptive selectivity. Given how thermostats work, they *require* an internal device that track changes in the environment by selectively responding to them. Thus, although the thermostat could work without this or that structural similarity relation, it could not work without the bi-metallic strip being adaptively selective in its reactions to changing temperature. It is adaptive selectivity, not similarity that really matters here.

To put our verdict more broadly, the thermostat is not the kind of system which could be said to *use* structural similarity between its internal machinery and something in the world. Even though the similarity exists (i.e. there is a mapping from the states of the bi-metallic strip to variations in temperature), it is not actively exploited by the system in question (see Shea 2014). The bi-metallic strip plays a purely reactive function which depends on its being affected by the world. But if this is so, then, one might ask, when *does* a mechanical system count as S-representational?

Perhaps the clearest example is a system equipped with S-representations that function in a robustly off-line manner. These are S-representations that perform their representational duties in a mechanism even when there is no direct contact (causal or other) between this mechanism and whatever is represented. For example, the represented entity could be so spatiotemporally distant from the mechanism so as to count as absent for it (Clark & Toribio 1994; Haugeland 1998). In this strong sense, deploying S-representations off-line consists in manipulating them for the purpose of representing things located in the (distant) past, future, or ones that are merely counterfactual. Notice how at least in the case of (S-)representing future and counterfactual states of affair (or entities, or events, etc.), there is no possibility – at least not without some serious metaphysical gymnastics – of saying that the representation is tokened 'in response' to the tokening of what is represented.

Let us now use off-line S-representations as paradigmatic cases of S-representations and see what is involved in representing something by exploiting structural similarity. It seems that three ingredients are required. First, the S-representation is actively transformed or manipulated within the mechanism. That is, the S-representational vehicle undergoes an endogenously-controlled process in which its structure changes over time. The structure of the vehicle is being effectively put to use by some larger mechanism. Second, manipulations of this sort are employed by the larger mechanism to perform certain function. For example, the

job for a larger mechanism (see also Shea 2014). What is required for the mechanism to work is adaptive selectivity, and similarity just rides on its back. That is, the thermostat could work under *any* mapping from states of bi-metallic strip to variations in temperature because it is not the *mapping* that is being used.

effects of manipulations could serve (for some consumer component) as a basis for a decision about which course of action – out of some possible range – to take. Third, the degree to which the effects of such manipulations of S-representational vehicle's structure are actually functional should reliably depend on how well those manipulations and their outcomes resemble the processes underwent by whatever is represented. That is, the manipulations need to actually resemble or simulate the relevant processes if they are to do their job. More precisely, (1) the degree to which the manipulations performed on the vehicle successfully simulate the relevant processes in the world should causally depend on (2) the degree to which the former (dynamically) resemble the latter. This, of course, is in line with the idea of similarity as a fuel of success.

To illustrate this, take the rat's spatial navigation system again. First, it has been suggested that place cells in a hippocampal spatial map can fire in sequences in a purely off-line manner, e.g. when the rat is asleep (see Shea 2014). The map is internally manipulated and the firing sequences are supposed to correspond to possible routes that rat could take when navigating an actual territory. Second, these manipulations are functional for the navigational mechanism in that they (presumably) serve as basis for route planning. Perhaps alternative routes that could lead to a reward are simulated in order to select one that is the shortest (Shea 2014). Third, this off-line planning is effective to the degree to which internal simulations can be actually projected to actual interactions with the environment. That is, we could manipulate rat's ability to effectively plan short, cost-effective routes through the environment by intervening in the degree to which its hippocampal map resembles it structurally.

But are S-representations *restricted* to the domain of off-line cognition? Not at all. Organisms can exploit similarity also when it comes to regulating on-line interactions with the environment. In fact, the existence of reliable co-variance between some system's internal machinery and environmental states of affairs does not, by itself, exclude the possibility that it *is* an S-representation-using system (as opposed to one whose functioning is based solely on adaptive selectivity). Imagine a cognitive system whose internal states change concurrently to changes in the external environment, and control behavior so that it is adaptive given the circumstances. Someone might (mistakenly) consider it to be a detector system. However, when we investigate the system's workings, it turns out that its internal machinery is cut off from the world; it has no sensory system. What explains its successful behavior is that it is equipped with an internal structure that ongoingly simulates the changing environment. This simulation process is not a matter of responding (causally, or by having a common cause, or by being contingently correlated, etc.) to the world. Rather, it is an endogenously controlled process whose unfolding resembles the relevant dynamics in the environment, enabling the system behave in accordance the world it inhabits. The best, in fact, the only way of explaining the system manages to cope with the environment is by pointing to the similarity relation between its internal processes and processes in the environment. Hence, despite working in a purely on-line manner, the system in question turns out to employ an S-representation of its environment[14].

7. What about intentional content?

As William Ramsey (2007, 2015) argues, there are two separate questions that need to be distinguished when one is theorizing about representation. On the one hand, there is the question of what it means for something (say, a component in a mechanism) to perform a *function* of a representation . One the other hand, there is the question of how representational *content*, i.e. representation's accuracy or satisfaction conditions, is determined. These problems are not identical, as content might be determined by factors other than those that constitute representational function (for details, see Ramsey, 2015). So far, our preoccupation in this article has been with the functioning of S-representations – on what it means for similarity relation to be exploitable or the S-representations to be manipulated off-line. Before we close the discussion, however, we want say some things about how content determination might work in the case of S-representations as construed in our account.

[14] Of course, this example is given is for purely illustrative purposes. No real-life S-representational system, even one whose cognitive processes unfold in a purely on-line manner, would be able to work if it was completely unresponsive to the changes in the external environment. It would be impossible for such an encapsulated agent to detect and correct errors in the endogenous simulation of the environment, which could lead to catastrophic consequences. It is much more reasonable to postulate a *mixed* strategy which combines detector-based and S-representation-based ways of dealing with the environment on-line. What we mean is a system that simulates the environment but is at the same time equipped with response-selective detectors. The internal model could make predictions about the way the detectors will be affected by states of the world, with the mismatch between that prediction and what is actually generated in the senses serving as a way of 'measuring' the representational error (Gładziejewski 2015c). This sort of prediction-error-based cognitive strategy is postulated by theories in motor control (Grush 2004; Wolpert et al. 1998), as well as by recent predictive processing approaches to cognition (Clark 2013; Friston 2010; Friston & Stephan 2007; Hohwy 2013). Although S-representations are not detectors, they will often need detectors to help them with their S-representational duties.

The problem of content is pressing for us because the very notion of S-representations has been criticized by Eric Myin and Daniel Hutto (2015) precisely for the fact that it supposedly falls flat when it comes to accounting for representational content. These authors deny that S-representations are the sort of things to which we could reasonably attribute accuracy conditions. In fact, to put this discussion in broader context, Hutto and Myin (henceforth H&M) are similarly critical of *every* existing naturalistic theory of representation (Hutto & Myin 2012). These theories, according to H&M, fail to resolve the Hard Problem of Content, viz. the problem of accounting for accuracy conditions in purely non-semantic and naturalistic terms. Roughly speaking, H&M's claim is that the most promising naturalistic accounts of content aim to explain accuracy conditions in terms of information understood as covariance. However, since, as they argue at length (Hutto & Myin 2012), covariance cannot be equated with content, the most promising theories of content fail miserably at dealing with the Hard Problem.

Of course, the reader might notice that our own theory of S-representations does not fall under H&M's critique. After all, S-representations are based on structural similarity, and we have argued before that structural similarity is not the same as covariance. It appears, however, that this is not enough of a difference for H&M – for them, similarity fares no better as a ground for content than covariance (Myin & Hutto 2015). What they claim is, at heart, that *similarity just is not content*. In addition, for H&M the fact that the similarity relation is exploitable for some larger system or mechanism is of no help either. For them, this sort of view simply boils down to teleosemantics, a theory which, according to them, conflates having content with having a (non-semantic) function of attuning the organism to its environment.

H&M's critique of naturalistic treatments of content, including ours, naturally invites one raise a following meta-level problem. What would it even take for any theory to solve the Hard Problem of Content? What sort of criteria should such a theory meet in order for us to be in a position where we can rationally claim that this theory *really*, *truly* accounts for intentional content? How can we ever know that somebody has actually succeeded at naturalizing content?

Much could and should be said about different strategies of solving this Metaproblem of Content. Again, this is a large subject that merits a separate treatment. For the sake of brevity, let us explain how we see the solution. We claim that the question of representational content is not *fully* independent from question of representational function after all. Namely, what we want to put forward is that if some internal state or structure plays a nontrivially, genuinely representational *functional* role, then we have at least a prima facie reason to ascribe *content* to that state or structure. The idea here is that when some mechanistic component R serves as a representation within a cognitive mechanism, then there should be mechanism-external conditions such that R's functioning *qua* representation reliably depends on the occurrence of these conditions. In other words, representational vehicles have conditions of successful operation *qua* representational vehicles. These conditions could be regarded as, simply, *accuracy* conditions, or intentional contents. If this is a workable position, then a large chunk of heavy lifting in explaining content is done by showing that a structure in question actually performs a truly representational function – a project that is, we admit, by no means simple or straightforward (Ramsey 2007).

Let us see how this sort of approach works when applied to internal S-representations. We assume that cognitive structures that we call S-representations play genuinely representational functions in a cognitive system. In particular, they can be said to serve as internal, manipulable models of the environment (Gładziejewski 2015b; Gładziejewski 2015c). We take this claim for granted here and we will not argue for it, as even H&M (2015) do not take issue with idea that S-representations (mechanistically construed) can be regarded as internally modeling some parts of the world. In any case, because S-representations serve as models, and thus representations, we treat our view as providing an answer to the question of representational function.

But how can this idea work in detail? Well, once you have something subpersonal in the brain (or some other, perhaps artificial cognitive system) which is in the business of modeling the world, then it makes sense to say that there are *ways the world should be like* in order for this model to perform its representational function properly. That is, it makes sense to ask about what the world should be like in order for the model to perform its function properly. And these ways the world should be like are, in other words, the model's accuracy conditions. To take a familiar example, imagine a rat navigating a maze by employing a hippocampal map with a particular co-activation structure. In order for this map to enable successful navigation (i.e. perform its function properly), it has to correspond structurally with the maze, or at least correspond to some sufficient degree. That is, given the structure of the internal map, we can say that the relevant part of the external environment should have such and such (spatial) structure if this map is to effectively guide the rat through its environment. This latter structure constitutes the content of the map, i.e. its accuracy conditions. On such a view, the S-representation can be said to be accurate if it properly performs its function as a result of the fact that its accuracy conditions match, to sufficient degree, the conditions in which it is actually employed; for example, when the internal map that the rat uses for navigation resembles the terrain enough to enable successful navigation. The representation is inaccurate if it fails to perform its function because its accuracy conditions do not match, to a sufficient degree, the conditions in which it is actually employed; for example, when the rat fails at navigating the terrain because there is not enough structural resemblance between its internal map and the terrain (see also Cummins on content/target distinction, where content is what the representation *applies to* and target is what the representation is actually *applied to*; Cummins 1996).[15] We also require that the rat have error-correction mechanisms that are activated at least some of the time when the accuracy conditions are not satisfied (for more on the idea of system-detectable error, see Bickhard 1993).

As mentioned before, Myin and Hutto (2015) criticize this sort of approach by claiming that it confounds the functional relevance of *content* with functional relevance of *similarity*. What we at most show, according to those authors, is how similarity, and not content can explain success. But this sort of criticism begs the question. Our proposal aims to *account* for content in terms of (exploitable) structural similarity. Again, when something acts as an internal model – as internal S-representations do – then the conditions of its successful operation (i.e. the conditions to which the model has to correspond structurally in order to do its job *qua* model right) can be equated with accuracy conditions, or content. For us, it is the fact that something plays a functional role of a model, viz. a representation, that justifies attributing it with content. To explain success by the accuracy of representation *just is* to explain the success by pointing to the fact that the model being used by some cognitive mechanism resembles, to sufficient degree, the relevant part of the environment. H&M's critique strikes as based on an unjustified *presumption* that exploitable similarity cannot ground or give rise to content. What these authors would have to provide is not a simple proclamation that 'similarity isn't content', but some positive argument for the claim that

[15] It needs to be stressed that on our view, *two* ingredients are required for content to emerge: the structural similarity relation and a larger mechanism that exploits this relation for successful operation. This way our proposal sidesteps philosophical problems that plague theories that blatantly reduce representational content to either *just* the similarity relation or *just* pure teleology or conditions of success of some activity. See (Gładziejewski 2015a) for more details on this. Curiously, H&M seem to miss this point. Their critiques point *either* to the fact similarity is not sufficient for content (which is true) *or* to the fact that teleology (i.e. playing some action-guiding role) is not sufficient for content (also true). They never seem to address the possibility that both factors are needed if they are to give rise to content: if it is to play content-grounding role, the similarity has to be functional, or, in other words, exploitable from some larger system.

something functioning as an internal model within a cognitive mechanism should not (or cannot) be taken to have accuracy conditions.

Conclusion

Recent years have seen the resurgence of interest in the idea that mental representations are grounded in similarity – *structural* similarity in particular. In the present article, we attempted to clarify and enrich our understanding of the claim that mental representations are S-representations. First, we tried to elucidate the nature of structural similarity relation, and proposed a mechanist-interventionist interpretation of the idea of similarity as an 'exploitable' relation. Second, we provided reasons for thinking that S-representations are indeed a separate type of representations, distinct from (purported) indicator or detector representations. Third, we addressed some criticisms that might be raised against our view. Overall, we hope that our proposals further pave the way that leads away from seeing representations as a matter of reacting to the world detector-style, and towards the idea that representing the world is a matter of actively modeling it.

Before we close the discussion, there is one last issue that merits mention. Some authors have argued that the domain in which S-representations can be explanatory is restricted to low-level cognition, and that S-representations are not quite suited to explain more sophisticated, human-level cognitive capacities (Morgan 2014; see also Garcia Rodriguez & Calvo Garzon 2010). But notice how our own proposal presented here only provides relatively minimal, empirically uncommitted criteria of what counts as an S-representation. Because of this, our minimal characterization can be met by internal structures that vary, perhaps drastically, in terms of their cognitive sophistication. There are couple dimensions along which there could be such variance. First, the vehicles of S-representations can vary in their relational complexity (and there should be corresponding variance in the complexity of their representational objects). Second, the manipulations performed over those vehicles can vary in their dynamic or computational complexity. Third, S-representations can differ in how decoupled they are from the environment; i.e. they can function in a way that is more or less off-line (see Gładziejewski 2015). Fourth, perhaps a case could be made that components that act as consumers of S-representations can vary in how flexible and context-dependent they are in their reactions (see Cao 2012). Now, if we agree that S-representations differ along those dimensions, what we end up is a continuum of S-representations of increasing structural and functional sophistication. If this is a workable position – and we see no reason to doubt this – then it is no longer mysterious how S-representations could underlie both simple and phylogenetically old cognitive capacities, as well as complex capacities that are phylogenetically new and perhaps human-specific, such as reasoning, imagery, or mental time travel. Roughly, more sophisticated cognitive functions are underpinned by more sophisticated S-representations. In fact, our own empirical bet is that human-level off-line cognition is largely a matter of being equipped with highly sophisticated S-representations – S-representations that actually earn the status of 'mental models'. But this is a subject for another occasion.

References

Alme, C.B. et al., 2014. Place cells in the hippocampus: Eleven maps for eleven rooms. *Proceedings of the National Academy of Sciences*, 111(52), p.201421056.

Bartels, A., 2006. Defending the structural concept of representation.

Barwise, J. & Seligman, J., 1997. *Information flow: the logic of distributed systems*, Cambridge; New York: Cambridge University Press.

Bechtel, W., 2014. Investigating neural representations: the tale of place cells. *Synthese*.

Bechtel, W., 2008. *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*.

Berger, A., 1980. Quine on "Alternative Logics" and Verdict Tables. *The Journal of Philosophy*, 77(5), pp.259–277.

Burge, T., 2010. *Origins of objectivity*, Oxford: Oxford University Press.

Carroll, L., 1900. *Alice's adventures in wonderland*, New York; London: Street & Smith.

Clark, A., 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioral and brain sciences*, 36(3), pp.181–204.

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

Craver, C.F., 2007. Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience.

Cummins, R. & Roth, M., 2012. Meaning and Content in Cognitive Science. In R. Schantz, ed. *Prospects for Meaning*. Berlin & New York: de Gruyter, pp. 365–382.

Cummins, R.C., 1996. Representations, Targets, and Attitudes.

Cummins, R.C. & Poirier, P., 2004. Representation and indication.

Decock, L. & Douven, I., 2010. Similarity After Goodman. *Review of Philosophy and Psychology*, 2(1), pp.61–75.

Dretske, F.I., 1982. *Knowledge and the Flow of Information* 2nd ed., Cambridge, Mass.: MIT Press.

Dretske, F.I., 1986. Misrepresentation. In R. Bogdan, ed. *Belief: form, content, and function*. Oxford: Clarendon Press, pp. 17–37.

Eliasmith, C., 2013. *How to build the brain: a neural architecture for biological cognition*, New York: Oxford University Press.

Fodor, J.A. & Pylyshyn, Z.W., 1988. Connectionism and cognitive architecture: a critical analysis. *Cognition*, 28(1-2), pp.3–71.

Friston, K., 2010. The free-energy principle: a unified brain theory? *Nature reviews. Neuroscience*, 11(2), pp.127–138.

Friston, K.J. & Stephan, K.E., 2007. Free-energy and the brain. *Synthese*, 159(3), pp.417–458..

Gärdenfors, P., 2000. Conceptual spaces: the geometry of thought.

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*.

Gładziejewski, P., 2015b. Explaining Cognitive Phenomena with Internal Representations: A Mechanistic Perspective. *Studies in Logic, Grammar and Rhetoric*, 40(1), pp.63–90. Available at: http://www.degruyter.com/view/j/slgr.2015.40.issue-1/slgr-2015-0004/slgr-2015-0004.xml.

Gładziejewski, P., 2015c. Predictive coding and representationalism. *Synthese*. Godfrey-Smith, P., 1996. Complexity and the Function of Mind in Nature.

Goodman, N., 1972. *Problems and projects.*, Indianapolis: Bobbs-Merrill.

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

Grush, R., 2004. The emulation theory of representation: motor control, imagery, and perception. *The Behavioral and brain sciences*, 27(3), pp.377–96; discussion 396–442.

Haugeland, J., 1998. Representational Genera.

Hohwy, J., 2013. The Predictive Mind.

Hutto, D.D. & Myin, E., 2012. Radicalizing Enactivism: Basic Minds Without Content.

MacKay, D.M., 1969. *Information, mechanism and meaning*, Cambridge: M.I.T. Press.

Metzinger, T., 2003. *Being No One. The Self-Model Theory of Subjectivity,* Cambridge, Mass.: MIT Press.

Miłkowski, M., 2013. Explaining the Computational Mind.

Miłkowski, M., 2015. The Hard Problem of Content: Solved (Long Ago). *Studies in Grammar, Logic, and Rhetoric*, 41(54), pp.73–88.

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

Myin, E. & Hutto, D.D., 2015. REC: Just Radical Enough. *Studies in Logic, Grammar and Rhetoric*, 41(1), pp.61–71.

N. Shea, 2014. Exploitable Isomorphism and Structural Representation. *Proceedings of the Aristotelian Society*, 64.

Newman, M.H.A., 1928. Mr. Russell's "Causal Theory of Perception." *Mind*, 37(146), pp.137–148.

O'Brien, G., 2014. How Does Mind Matter?

O'Brien, G. & Opie, J., 2004. Notes towards a structuralist theory of mental representation. In H. Clapin, P. Staines, & P. Slezak, eds. *Representation in Mind: New Approaches to Mental Representation*. Amsterdam: Elsevier, pp. 1–20.

Opie, J. & O'Brien, G., 2004. Notes toward a structuralist theory of mental representation.

Ramsey, W., 2007. Representation Reconsidered.

Ramsey, W., 2015. Untangling two questions about mental representation. *New Ideas in Psychology,* pp.1–10.

Redish, A.D., 1999. *Beyond the cognitive map: from place cells to episodic memory,* Cambridge, Mass.: The MIT Press.

Rescorla, M., 2009. Cognitive Maps and the Language of Thought. *The British Journal for the Philosophy of Science,* 60(2), pp.377–407.

Rosen, R., 2012. *Anticipatory systems: philosophical, mathematical, and methodological foundations* 2nd ed., New York: Springer.

Shagrir, O., 2012. Structural Representations and the Brain.

Shea, N., 2007. Consumers Need Information: Supplementing Teleosemantics with an Input Condition. *Philosophy and Phenomenological Research*, 75(2), pp.404–435.

Shea, N., 2013. Millikan's Isomorphism Requirement.

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

Tversky, A., 1977. Features of Similarity. *Psychologial Review*, 84(4), pp.327–352.

Weisberg, M., 2013. *Simulation and similarity: using models to understand the world,* New York: Oxford University Press.

Wolpert, D.M., Miall, R.C. & Kawato, M., 1998. Internal models in the cerebellum. *Trends in Cognitive Sciences*, 2(9), pp.338–347.

Woodward, J., 2003. Making Things Happen: A Theory of Causal Explanation.

Woodward, J., 2008. Mental causation and neural mechanisms.