Situated Mental Representations: Why we need mental representations and how we should understand them

Albert Newen, Ruhr-Universität Bochum and Gottfried Vosgerau, Universität Düsseldorf[1]

appeared in: J. Smortchkova, K. Dolega & T. Schlicht (Eds.) (2020). *Mental Representation.* Oxford: Oxford University Press, S. 178-212.

1 Why posit mental representations?

Mental representations are a means to explain behavior. This, at least, is the idea on which cognitive (behavioral) science is built: that there are certain kinds of behavior, namely minimally flexible behavior, which cannot be explained by appealing to stimulus-response patterns. Flexible behavior is understood as behavior that can differ even in response to one and the same type of stimulus or that can be elicited without the relevant stimulus being present. Since this implies that there is no simple one-to-one relation between stimulus and behavior, flexible behavior is not explainable by simple stimulus-response patterns. Thus, some inner processes of the behaving system (of a minimal complexity) are assumed to have an influence on what kind of behavior is selected given a specific stimulus. These inner processes (or states) are then taken to stand for something else (features, properties, objects, etc.) and are hence called "mental representations". They are presupposed for two main reasons, 1. to account for flexible reactions to one and the same stimulus and 2. to account for behavior triggered when the relevant entities are not present. The latter case is highlighted by J. Haugeland: "if the relevant features are not always present (detectable), then they can, at least in some cases, be represented; that is, something else can stand in for them, with the power to guide behavior in their stead. That which stands in for something else in this way is a *representation*" (Haugeland 1991, 62 original italics). In this sense – which is a very liberal and wide sense of 'representation' that does not imply a specific structure or degree of complexity of mental states – we have to posit some kinds of mental representation so as to explain flexible behavior.[2]

If the behavior is not flexible but reflex-like and only elicited in presence of the stimulus, behavioristic explanations without the posit of mental representations will suffice. This is not to say that the details of the mechanisms do not play a role in the explanation of the behavior. For example, there are smoke detectors that contain a little light source and a light dependent resistor, so that they react on everything that leads to less light on the resistor, including smoke. These details are of course relevant to understand the details of the behavior of the smoke detector. However, they do not justify to calling

[1] For helpful critical comments on earlier versions, we would like to thank especially Robert Matthews as well as Edouard Marchery and Robert Rupert. Furthermore, we recevied helpful criticism from Matej Kohár and Beate Krickel.

[2] This idea can also be found in the pioneering work that led to the assumption of mental representations, e.g. Tolman and Honzik (1930) and Tolman (1948) (cf. Vosgerau 2010; Vosgerau and Soom 2018).

the light in the smoke detector a representation of non-smoke or anything like this. The same is true, e.g., for sea bacteria with magnetosomes or the dogs in Pavlov's experiments. This is at least the picture that we believe cognitive science to be built on[3] and the picture we want to defend here. We presuppose a standard concept of representation according to which a MR has at least two features, namely being a vehicle and having a content. If someone wants to reduce MR to just the aspect of being a vehicle and argue that the content is irrelevant, then this makes the concept of MR superfluous. The aspect of the vehicle can be adequately accounted for by the description of a mechanism without the aspect of content. According to our view, we need MRs only if the scientific explanation of flexible behavior needs to involve the aspect of content. There is one group of philosophers challenging the view that content can play an explanatory role since it seems unclear how, as Dretske (1988) puts it, content gets its hand on the steering wheel, or more precisely, how contents can play a causal role in a mechanism. Let us outline our strategy: We will argue for a systematic interdependence between vehicle and content of a MR: we accept that only the vehicle is causally involved in the relevant mechanism(s). But the content of a MR is part of a causal explanation which is available at different levels which are more coarse-grained than the level of the mechanism. Even if there is only one causal mechanism in which the vehicle plays a causal role, there are many causal explanations at different levels which are all systematically related (in different ways) to this causal mechanism which guarantees that the content explanation are also coarse-grained causal explanations (further clarification in section 5.4).

Recently, some philosophers have challenged the possibility of contents playing any explanatory role, arguing that we should get rid of the concept of mental representation. According to many enactivists (e.g. Hutto & Myin 2013, 2017) all kinds of behavior are to be described without presupposing any *mental* representation. *Linguistic* representations are still assumed to be real representations and to be needed in the explanation of some kinds of behavior; however, they are no longer classified as *mental* representations. We will come back to this challenge later, since it is deeply interwoven with the question of where to draw the line between nonrepresentational and representational cognitive abilities. To account, in principle, for the critical attitude of enactivism, we focus on nonlinguistic mental representations which are needed to explain the behavior of cognitive systems. We argue that we actually need to presuppose nonlinguistic mental representations to explain flexible behavior in animals, prelinguistic children, and also in adults. But these mental representations have to be understood in a new way, namely as real, non-static, use-dependent, and situated mental representations relative to a certain behavior or cognitive ability. MRs are real in the sense that their functional roles are realized by mechanistic relations involving the relevant vehicle and its components; and this realization can involve different levels, such as the neural level, but also the bodily or further physical and social levels. They are non-static in the sense that the variation between

[3]In fact, cognitive science started as a reaction to behaviorism, which could not explain the behavior of rats (see also footnote 2).

the realization basis of two tokens of a MR type can be huge: there are no necessary and together sufficient features determining the relevant behavior or cognitive ability to be explained, as we will show.[4] Rather, the realization basis is an integrated pattern of characteristic features that can vary greatly because a single integrated token can involve features from the package of characteristic features that are not included in another token and vice versa (e.g. as argued for emotions in Newen et al. 2015). They are use-dependent in the sense that the *coarse-grained content of a MR* (what it is about) depends on what the MR is used for, i.e. what the according behavior is targeting at (given an explanatory interest). MRs are situated in the double sense that: (i) we do not only allow for neural states as vehicles (but also for a combination of neural and bodily states), and (ii) one and the same MR can have different *specific fine-grained contents* (format of representation) depending on the detailed explanatory level. To elaborate on these four features (being real, non-static, use-dependent, and situated), we will develop them by critically delineating our account from the main alternative accounts of MRs, instead of discussing the features one by one. First we describe the functionalist framework of our account (sec. 2), then, we introduce what-where-when memory in birds and rats as a key example as in need of involving non-linguistic MRs (sec. 3). Then we analyze the central deficits of three leading theories of MR, namely the realist account of Fodor, the Ersatz-representationalism of Chomsky and the mathematical pragmatist account of Egan (sec. 4). Our alternative account (developed in sec 5) can be called *situated mental representation*: it combines a *functionalist account of representation with a relational dimension that can vary with the situation type and that allows non-static constructions of mental representations in specific contexts*. We spell out our account of MRs according to which we need to distinguish vehicle, content and format of MRs and especially account for the non-static construction of the vehicle. Of course we also deliver some fruitful applications.

2 A minimal characterization of mental representations and the exclusion of *non-functionalist and unitary feature* accounts

There is some consensus concerning the "job description" for mental representations (MRs): MRs should be introduced only if (i) we need them to explain a certain type of minimally flexible behavior of cognitive systems.[5] (For this paper, we leave aside cases relying on natural language.) But we should not adopt an ideological stance on which mental representations are presupposed as definitional for cognitive science. We need to be open-minded about alternative nonrepresentational explanations

[4] As we will see, there are of course very specific neural mechanisms in play at a low level of processing, but they are usually involved in the realization of rather many different types of flexible behavior or cognitive abilities and are thus not sufficient (nor necessary) for the specific behavior. We will elaborate on this is section 5.4.

[5] In effect, this is the same as to say that representations can be wrong. In the case of non-flexible behavior, as reflexes, there is no mistake possible. It doesn't make sense to say that Pavlov's dog dribbled mistakenly. So, this constraint is closely connected to the widely agreed upon statement that there is no representation without misrepresentation.

of cognitive abilities (Ramsey 2015). Furthermore, it is commonly agreed that (ii) MR are asymmetric (if *r* represents an object *o*, then *o* does not represent *r*) and (iii) have to allow for misrepresentation.

These three conditions allow us to exclude three radical views of MR. Ad (iii): We can exclude *a purely causal theory of MR*. MRs cannot be explained by causal relations alone because there is no such thing as miscausation. Moreover, often the same effect can have different causes, which leads to underspecification, making MR useless for explaining behavior.[6] More generally, causation is usually not connected to representation: an avalanche could be caused by a falling tree (while the latter can be caused by a strong wind), but the avalanche represents neither the falling tree nor the strong wind nor both together. Even if we have a strict and reliable causal relation, such that one type of effect can only be produced by one type of cause, there is still no representation because there would be no possibility of misrepresentation (even if we imagined smoke being caused by fire and by nothing else, it is still not true that smoke would represent fire). Furthermore, MRs of non-existing things like vampires cannot be caused by what they represent. Ad (ii): We can also exclude *a pure similarity theory of MR*: if *r* is said to represent an object *o* just because *r* is similar in certain respects to *o*, we lose the feature of asymmetry of MRs because similarity is in general a symmetric relation. Ad (i): Finally, one may argue for *a purely conventionalist theory of MR* such that *r* represents *o* if and only if there is a convention in a group such that *r* is by this convention connected with *o* – the paradigm case here are linguistic conventions, e.g. a name representing a person. But to assume that every MR is a linguistic representation would clearly be a case of over-intellectualizing behavior: many types of flexible behavior are not dependent on the existence of linguistic rules, especially flexible behavior in prelinguistic children and nonhuman animals. Thus, we need to presuppose MRs which are not anchored in social or linguistic conventions even if some MRs might be linguistic.

As already mentioned, radical enactivists (Hutto & Myin 2013, 2017) challenge the assumption that there are any non-linguistic MRs, arguing that we can explain all kinds of behavior with the help of linguistic representations or no mental representations at all. Our argument against this view is based on examples of behavior by nonhuman animals or prelinguistic children that are best explained by reference to nonlinguistic representations. What exactly is the difference between a position assuming nonlinguistic mental representations, and the radical enactivists' denial of them? Although they are anti-representationalists, radical enactivists do not deny that there are internal neural processes which cause nonlinguistic cognitively guided behavior, but they mainly deny (a) that these internal neural processes are systematically connected with a type of behavior and (b) that those internal neural processes are vehicles for contents; content; i.e. mental representations. Rather, they argue that it is adequate and sufficient to describe such behavior in terms of affordances and dispositions. In particular, it would be both unnecessary and impossible to suppose that some neural correlates have

[6]This is basically the so-called disjunction problem (Fodor 1987). His "solution" to the problem is to formulate asymmetric dependencies between the different causal relations, which is rather a sophisticated way of restating the problem than actually solving it. What we would need is an explanation of where the asymmetric dependencies come from (cf. Vosgerau 2009).

the systematic functional role of "standing in for something else", i.e. that they have content: "Fundamentally, by REC's [Radical Enacted Cognition] light, basic cognition is a matter of sensitively and selectively responding to information, but it does not involve picking up and processing information or the formation of representational contents" (Gallagher 2017. 92). On the contrary, representationalits argue that there is evidence that some neural, corporeal, or clusters of neural and corporeal states have systematic functional roles such that it is justified to ascribe them content. The ascription of content is justified because this content is explanatorily fruitful and even the best tool we can get for certain types of explanation, specifically for the explanation of some kinds of flexible behavior. This is the core of the debate. Our examples of nonlinguistic abilities are supposed to furnish evidence for systematic functional roles as being implemented in some sufficiently distinguishable neural processes or cluster of neural processes (or further conditions of implementation), and that these are best understood as mental representations.[7] So far, we have excluded purely causal, purely similarity-based, and purely convention-based accounts of MR, and promised to show that the purely linguistic account is also insufficient. We now want to set out and defend a specific functionalist account. To do this, we first introduce one key example which also has the function of demonstrating that we need to presuppose nonlinguistic MRs.

3 A key example: Spatial navigation and food finding in rats

Scrub-jays (Clayton and Dickinson 1998, 2007), rats (Crytal 2013), and probably more species have episodic-like memory. Let us focus on rats, since this case has been intensively investigated. Trained in an eight-arm maze, they are able to register three daylight situations (morning, midday, evening), distinguish different types of food (normal food and chocolate), and of course to register the spatial organization of the maze. Rats swiftly learn to expect a certain type of food in a certain arm, and can also combine this with a certain time of day. They can even learn to understand *that if there is normal food in arm2 in the morning, there will be chocolate in arm7 at midday* (Crystal 2013; Panoz-Brown et al. 2016). Why do we have to presuppose MRs to explain the behavior of the rats? Why are behavioral dispositions not sufficient? Rats in this case are shown to be able to register and systematically distinguish at least two types of food, eight locations and three day times. The best explanation is to postulate an informational state with a (partially) decoupled representation of the type of food that can be combined with representations carrying information about different day-times and different locations. The alternative would be to posit as many different internal states as there are combinations of the different parameters. This explanation, however, is not compatible with the fast learning curve of the rats. Therefore, an explanation in terms of specific dispositions or affordances, which the rat would have to acquire or learn one by one, is not only incomplete but also inadequate. The informational state of rats which have learned to behave according to a conditional in the maze is

[7]Although we focus on neural correlates as the main implementation basis, we are not committed to a strict internalism. The debate over internalism and externalism is independent of the debate on mental representations (Vosgerau and Soom 2018, Vosgerau 2018), and the latter is the sole topic of this article.

best characterized as structured into components of <object-type; location; time>. The alternative would be to presuppose a high number of independent, non-structured dispositions which need to include all the possible permutations of associations between a starting state of affairs and a type of behavior. And these dispositions would need to be learned independently of each other, since there would be no common component to be taken over. We cannot decide on the basis of intuitions which explanatory strategy counts as more parsimonious (Starzak, 2017), but we need additional facts to provide a foundation for the best explanation. Let us first look at the neural processing evidence: earlier studies have shown that rats (at least a large group of them) can spatially orient themselves based on landmarks (realized by the hippocampus), in contrast to a group of rats which rely on rigid procedural knowledge (realized by basal ganglia) (Packard and McGaugh 1996). Rats using landmark orientation are able to adjust their behavior when they have to start from different positions in the maze: this indicates that they remember and recognize landmarks and can combine that with different types of behavior (taking the second right from one direction, or taking the second left from the opposing direction). This indicates that we have to presuppose nonlinguistic representations. This is not an over-intellectualization, because the presupposition of structured representations of <object-type; location; time> is also supported by further neuroscientific evidence. There is strong evidence that rats are able to represent three-dimensional objects (Burke et al. 2012), while the relevant behavior correlates with activity of the perirhinal cortex. Furthermore, well corroborated research suggests that location is represented by either place cells or their combination with grid cells (Bjieknes et al. 2018, Moser et al. 2008) in such a way that a sequence of places which a rat has actually visited leads to a sequence of place cell activation. Even more convincingly, there is evidence that such sequences of place cell activations can be reactivated without the rat actually being in the same scene, which is usually interpreted as a replay of a spatial path that was taken before (Káli and Dayan 2004, Jezek et al. 2011). Rats also seem to represent *time* as a feature of an event (Tsao et al. 2018). Taken together, there is overwhelming evidence that we should presuppose nonlinguistic mental representations in rats structured as <object-type; location; time>. An additional argument for structured representations is drawn from observations of learning and relearning by rats in the previously discussed scenario: rats can quickly learn new conditionals in the eight-arm maze, or, even more demandingly, are able to understand the eight-arm maze as having been newly restructured (e.g. rats can also learn conditional behavior in a freely accessible eight-arm maze, and understand in a certain context that only the arms numbered 1, 3, 5 and 7 are part of the game, while the others are irrelevant for inspection (Crystal, 2013)). Thus, the informational state of the rat has components which are structured and can be flexibly combined. This allows for an interesting degree of epistemic flexibility and predicts a quick learning of new combinations that cannot adequately be described as a multiplicity of independent, non-structured dispositions. What follows is that the best description and explanation of the rats' behavior requires MRs with content.

Nevertheless, a defender of radical enacted cognition could reply that "[b]asic minds target, but do not contentfully represent, specific objects and states of affairs. To explicate how this could be so, Hutto and Satne (2015) introduced the notion of target-focused but contentless Ur-Intentionality" (Gallagher 2017, 92). The general line of argument is that basic cognition (which includes all non-linguistic cognition; Hutto & Myin 2013) can consist in target-directed behavior without involving mental representations because basic intentionality can be realized without contents (Gallagher 2017, Chap. 5,6). We partially grant this thesis which leads to a clarification: in the case of rigid mechanisms, there is indeed no need to presuppose mental representations. For example, the navigation of sea bacteria with the help of magnetosomes can be completely explained and predicted without mentioning mental representations. We also know that rigid mechanisms can produce quite impressive behavior in an adequate environment (Gigerenzer 2000).[8] But, contrary to the claim of radical enactivists, not all nonlinguistic behavior can be accounted for with this move. And this is what our key example should demonstrate: first, it has been shown that there are systematic underlying neural processes which correlate with type of food, location of the food, and the time of the day. Second, these informational states implemented by neural states are re-used by the rat when realizing memory-based behavior, when quickly learning a new conditional in the maze, or when learning to apply the same conditional in a newly structured maze. The structure of the underlying neural processes can best be described to include representations of the type of food, representations of the location of the food, and representations of the time of day since there are systematic underlying neural processes and they are best understood as bearers of information which can be easily systematically recombined and guide specifically quick learning and flexible behavior. Finally, it is obvious, yet still worth mentioning, that these are not linguistic representations. We thus take it that nonlinguistic MRs are needed, and move now to the task of detailed and consistent description.

4 Three important but unsuccessful strategies to account for representation and misrepresentation

Before presenting the three prominent positions and our criticisms, it is worth highlighting that neither of the three accounts to be canvassed runs into the problems faced by simple causal, similarity-based, or convention-based accounts as set out above. The reason is that the following three accounts are all functionalist accounts, and we will develop the details of a functionalist framework later when presenting our own account. For now, we focus on discussing the three most important explanatory strategies in the literature, and show that despite their having a functionalist framework, they all are deficient.

As explained above, MRs have (at least) two features: they are implemented as vehicles (brain states), and they have content.

[8] Thus, complex behavior is not a sufficient criterion for presupposing mental representation but still a very good indicator which forces us to investigate the structure of information processing involved and to show that it is indeed flexible.

Fodor's Realism, also called 'Hyperrepresentationalism':

> The vehicles of mental representations are brain states that are syntactically organized. They thereby implement a language of thought. Their content is a feature of the mental representations which is constituted either purely intrinsic (Fodor, 1980) or by an intrinsic syntactic organization of the symbols of this language of thought and by their causal anchoring in the environment (Fodor 1987). In the latter case, the relation between content and vehicle (brain state) *is determined by intrinsic syntactic organization of symbols and an additional causal relation from symbols to objects (properties or substances)* (Fodor, 1987, 1994). *In both cases, a purely intrinsic or an intrinsic-causal determination of the content, the relations to the world are described as use-independent, i.e. independent from an explanatory perspective.*

Jerry A. Fodor's account faces open questions and radical challenges in both versions: the main worry is that we do not have any empirical evidence for a language of thought. In the 90s of the last century, we did not have alternative explanatory strategies to account for the acquisition of concepts as well as the systematicity and productivity of language. Accordingly, the theory of a language of thought underlying our natural language ability was very popular. But theories concerning the evolution and development of language have made a lot of progress; in particular, Tomasello (2003, 2008) outlines a gradual development of language abilities anchored in basic cognition including joint attention and shared intentionality. Thus, it is no longer a convincing implication to accept an inborn language of thought. More importantly, we run into general problems with the claim that the content of an MR is an intrinsic feature or an intrinsic-causal feature since both claims involve independence from an explanatory perspective: there is the well-known problem of underdetermination of content, often illustrated with the case of the frog's fly-catching mechanism: What is the content of the activated MR in a situation when the frog detects and catches a fly? Is it: (i) moving black dot present; (ii) fly present; or (iii) food present? Is it only an informational state or is it – as Millikan (1996, 2009) argues – a pushme/pullyou representation, which also includes a motivational content such as *activate the tongue* (OR *catch it* OR *eat it*)? If content was an intrinsic or intrinsic-causal feature (independent from an explanatory perspective), there should be no underdetermination. However, pinning down the content is notoriously difficult or even impossible without already adopting a specific explanatory perspective. The reason is that the only way to constrain the otherwise underdetermined content is to appeal to the best scientific explanation (Schulte 2012). Therefore, we have to integrate the relevant explanatory perspective into the determination of content. The content of the frog's representation is so dependent on the relevant explanatory perspective: if we want to explain the activation of the tongue throwing, then it is adequate to describe the content as "black dot present and activate the tongue". If we want to explain the survival of frogs in a context of flies being the standard food available, it is adequate to describe the content as "fly present and catch it". If we aim for the general explanation of the survival of a frog, the content is adequately described as "food present and insert it". We are not

excluding that there may exist one underlying mechanism but – if it were nonrigid (further discussion below) – it would need a content explanation in addition to the mechanistic description and then the content is underdetermined and needs specification through an explanatory perspective. Thus, Fodor's view is inadequate: there is lack of evidence for the presupposed complex inborn language of thought, and the content of mental representations is either systematically underdetermined or dependent on the explanatory perspective, i.e. it cannot be neither a purely intrinsic nor intrinsic-causal feature.

Taking the argument from underdetermination very seriously, Noam Chomsky developed a view which Egan (2014) calls 'Ersatz-representationalism'.

Chomsky's Ersatz-representationalism: While mental representations are required for scientific explanation, contents are not.

Chomsky takes the argument from underdetermination to imply that 'content' is not a scientific notion at all. Nevertheless, he wants to keep the notion of MR (whatever that might be and whatever role such contentless representations could play). The price to pay for this view is that all content involving explanations becomes unscientific, i.e. they have no place in scientific explanations[9]; and this price is too high, because many explanations in psychology and psychiatry are still evaluated in the special sciences as paradigmatic cases of scientific explanation. At the same time, they essentially rely on the ascription of contents: e.g. if someone is suffering from delusions of persecution, then his rigid belief with the content that the CIA is after him enables us to explain why he never uses any name tags. And this belief is a paradigmatic case of a mental representation. To criticize Chomsky's view it does not matter that it is a linguistically based MR. Of course, delusions of persecution have some neural basis (Bell and Halligan, 2013), but this knowledge is far too unspecific to predict the enormously diverse actions of the patient in different situations. Only at the level of the content of the delusion of persecution can we hope to find specific explanations of all the different reactions of the patient. Thus, it would be like throwing out the baby with the bathwater to deny that MR have contents.

Frances Egan (2014) is impressed by Chomsky's arguments, but she wants to keep the proverbial 'baby' by introducing a scientific content for mental representations that is different from the unscientific content of folk psychology.

Egan's dual representationalism: Paradigmatic scientific explanations are explanations involving computational mechanisms. They involve mental representations with a content but only with a *mathematical* content. Any further ascription of content, especially in folk psychology, is purely *pragmatic* and thus not part of the theory proper.

Egan thinks that Chomsky is right in denying the relevance of content in scientific explanations as far as folk-psychological content is concerned (including concepts, beliefs, desires etc.). Nevertheless, she

[9]We use the expression „unscientific" to refer to explanations and terms that we will not find in a „pure" science. Some of the unscientific explanations and terms might still be used by scientists, but only for didactic reasons or because of their simplicity and elegance.

concedes that there is a level of explanation that still involves some scientifically acceptable content: this is what she calls *mathematical content* and it is implemented at the level of computational mechanisms. For many phenomena, like picking up an object, we are able to deliver a good description of the computational mechanism (Egan 2014). This computational mechanism involves mathematical content, i.e. a function from numbers as arguments to numbers as values. And this computational mechanism has its mathematical content intrinsically in the sense that the mathematical function completely and uniquely matches the mechanism. Egan then argues that if we have identified a representational vehicle, we can attribute two contents to it, namely an intrinsically attached scientific content which is the *mathematical content*, and the *pragmatic content*, i.e. an intensional content of beliefs and desires in a situation. The latter is just an unscientific gloss (i.e. they are not part of the theory proper) resulting from placing the mathematical content into a rich context and extending it into an intensional context-dependent content. The latter may still be helpful in scientific theory construction and presentation and in everyday life, but remains a gloss which has no foundation in any mechanism. Thus, MRs have two types of content, although only the mathematical content has a place in the scientific theory proper.

The main argument that Egan puts forward to exclude pragmatic content from being scientific is basically the argument from underdetermination of intensional content illustrated by the frog example above (used against Fodor). Our central criticism against Egan is twofold: First, if the argument of underdetermination is successful against pragmatic content's being scientific, it is also successful against mathematical content's being scientific. So, we are left without any intrinsic content for MR. And second, even if an MR does not have an intrinsic content (as we accept), we are able to characterize contents that are scientific since they are required by explanations in the special sciences. The latter will be developed in our own proposal of MR.

Let us elaborate on our first and main criticism: We accept that intensional contents are underdetermined in the frog case. Moreover, content is in principle underdetermined if we take into consideration only the input and output of the underlying neural processing. The reason is, ultimately, that the behavior we try to explain remains unclear if we restrict the description in this way. Is it catching-flies-behavior or tongue-throwing-behavior that we are trying to explain? The inputs and outputs of the neural mechanisms do not distinguish between these two ways of describing the behavior. However, the difference is of great importance for empirical research: if we want to explain how frogs are able to catch flies, the cases with black dots count as unsuccessful behavior (indeed as systematic failures), which give us valuable insights into the details of the mechanisms involved. So, the basic scientific description of the behavior to be explained is not given by the inputs and outputs to neural mechanisms alone, and thus the representations and contents, that are supposed to take on some explanatory role, cannot be determined by them either. In particular, even if there was a unique mathematical description of the frog's behavior, it could also not determine which cases are successful cases of the behavior in question; a fortiori, it could not determine which cases are based on

misrepresentations and thus it could not said to be a representational content at all (remember: representation implies the possibility of misrepresentation).

We want to add a general argument of underdetermination from a philosophical perspective: if we describe a physical mechanism with a mathematical operation, then the latter remains underdetermined relative to the mechanism, since usually there are many mathematical functions which fit the same data of the mechanism. Our selection of one mathematical operation for an explanation depends on other presuppositions including the selection of an explanatory interest.

Having made this very general point, we want to remind the reader that the frog example is a bad candidate for representation-hungry behavior at all: it is not flexible. This invites a short discussion of two typical cases of basic cognition, frog's fly catching and the homing behavior of ants. Indeed, the frog (and likewise the toad) throws its tongue whenever a black moving dot occurs and there maybe a rigid neural mechanism underlying it (Neander 2017). This means that the tongue-throwing behavior is not flexible and can be sufficiently explained without any representation at all – it is a simple stimulus-response pattern.[10].MRs are needed to account for flexible behavior and they are introduced to account for a triggering role of informational states without the causal source being present: an intensely discussed case concerning the latter condition is the homing behavior of desert ants: the ant is able to return to its nest although it does not get any sense information about the nest. This is based on the process of path integration or dead reckoning (Gallistel 1990), and involves a systematic representation of the animal's location relative to its nest (Gallistel 2017). One observation indicating the substitutional role of the internal mechanism is the following: if the ant has found some food during an unsystematic search, it can go straight back to its nest. Now, if you displace the ant at this very moment, it runs along a same parallel vector, of course not ending at the nest but relying on the internal representation of the relative spatial location. Thus, without a nest-specific input to the neural mechanism the ant is able to find the location of the nest. There is a systematic neural underpinning and it has an informational content that triggers the ant's behavior. What exactly is the content? Parallel to the frog example it depends on our explanatory interest: we could either say that it is the nest (when we want to explain the homing behavior of the ant, which consists in returning to its home) or that it is the vector normally directed to the nest (when we want to explain in which direction the ant moves, which is different from the homing behavior since it obviously has different success conditions). As we will argue in detail (section 5.2), in the case of the ant's cognitive system the different descriptions based on different explanatory perspectives are connected with different explanatory power and the description as homing behavior informs us about new aspects of the underlying mechanism. In the case of the frog – due to its rather rigid underlying mechanism – the

[10] As mentioned above, this is not to say that the details of the mechanism, including the details of the registration of the black moving dot – the frog's perception – do not play a role in explaining the details of the behavior. Of course they do, and they are investigated in neuroscience as much as any other reflex is a fascinating research topic for neuroscience. However, not every neuroscience is cognitive (neuro)science in a demanding sense

analog different descriptions are not coming with any increased understanding of the cognitive mechanism; thus, this may be best described as a case not involving mental representations at all.

In the second part of our criticism of Egan (and of the other accounts as well), we dispute the implicit presupposition that if a content is not attached intrinsically, i.e. as long as it involves some underdetermination, then it cannot be part of a scientific explanation proper. We take this presupposition to be unjustified. This can, for example, be illustrated with cases of multiple realization: we know many realizations from the natural science (Fang 2018). For didactic reasons, let us focus on the well-known case of a multiple realization of jade as two chemical substances, namely either as jadeite, $NaAl(Si_2O_6)$, or as nephrite, $Ca_2(Mg, Fe)_5((OH,F)Si_4O_{11})_2$. Despite this obvious chemical difference, they have similar superficial properties and thus could be used for similar purposes in everyday life. Therefore, we can describe high-level causal processes which are realized by two different low-level mechanisms. High-level explanations are underdetermined in relation to low-level explanations. Nevertheless, both can be scientific explanations. It is an important insight worked out by Sober (1999) that scientific explanations need both high-level and low-level explanations to account for a phenomenon, where the level of explanation depends on the pragmatic research interest:

> Higher-level sciences often provide more general explanations than the ones provided by lower-level sciences of the same phenomena. This is the kernel of truth in the multiple realizability argument – higher-level sciences "abstract away" from the physical details that make for differences among the micro-realizations that a given higher-level property possesses. However, this does not make higher-level explanations "better" in any absolute sense. Generality is one virtue that an explanation can have, but a distinct – and competing – virtue is depth, and it is on this dimension that lower-level explanations often score better than higher-level explanations. The reductionist claim that lower-level explanations are always better and the antireductionist claim that they are always worse are both mistaken. (Sober 1999, 560)

Let us transfer this insight to a hypothetical case: assume that the fly-catching mechanism of the frog could be steered by two different mechanisms; one being the well-known visual mechanism reacting to black moving dots, the other an olfactory mechanism reacting to certain pheromones of flies. Then, a low-level description of the behavior would differentiate between a catching-flies-by-sight-behavior and a catching-flies-by-smell-behavior. However, on a more general level, there would be only the catching-flies-behavior, and the common explanation would involve representations with the intensional content *there is a fly*. This explanation would be one of the flexible behavior of the frog that it shows both with certain visual and with certain olfactory inputs. And it would not be reducible to the low-level explanations since the commonalities of the two mechanisms are not even visible

under a low-level (maybe mathematical) description of the mechanisms (cf. Vosgerau and Soom 2018).

The result is that Egan's two contents, mathematical and pragmatic, are in the same boat: both are underdetermined. But it does not follow that they cannot be part of a scientific explanation. Explanations of this kind would depend on pragmatic factors such as explanatory interest.[11] How can we develop a theory of MR which allows for both the dependency on pragmatic factors and a role for MRs in scientific explanations? This is what our own proposal of situated MR aims to set out.

5 Defending a functionalist framework and setting out an account of situated mental representation

5.1 The functionalist framework

We want to demonstrate that the most fruitful way to capture the basic ideas of MR is through a functionalist framework. We start with the classical notion of (this kind of) functionalism that stems from the discussion of abstract automata by Putnam (1975). In the first step, we shortly characterize the minimal basis by highlighting three aspects central to our notion:

> 1) The notion of "function" used is the mathematical notion referring to a mapping from sets to sets, where every member of the first set is mapped to exactly one member of the other set. This is a very abstract description, and of course, we will go further by assuming that such functions are somehow implemented by causal mechanisms; however, this is an additional assumption.
>
> 2) Behavior can, in general, be described by functions, i.e. by mappings from inputs (stimuli) onto outputs (reactions). In this simple form, we can describe reflexes in the spirit of behaviorism. In order to account for the flexible behavior discussed above, however, we need to add internal states to the functions, which takes us from behaviorism to functionalism. This is done on both sides of the mapping: the reaction to an input is not only determined by the input but also by the internal state the system is in; and the output is not merely a reaction to the stimulus/internal state pair, but also comprises the internal state the system will enter as a result of the behavior. Thus, the function now takes two arguments – stimulus and internal state – and two values – the subsequent internal state and the reaction.
>
> 3) As in behaviorism, the relevant function containing the MR (internal states) involves behavior (as opposed to e.g. brain output alone). Only in this sense the function in which the MR "plays a functional role" can determine what the MR stands in for: it stands in for objects

[11] In line with Sober, we clearly distinguish the level of scientific explanation from nonexplanations, and the involvement of pragmatic research interests is fully compatible with scientific explanations: "The claim that the preference for breadth over depth is a matter of taste is consistent with the idea that the difference between a genuine explanation and a nonexplanation is perfectly objective" (Sober 1999, 550).

or properties that are relevant for the behavior (e.g. the nest in the case of the homing behavior of the ant).[12]

Let us elaborate on this central point by specifying that behavior should be conceived of in terms of interacting with objects, properties, and other entities in the world; and not just as relations between stimuli (sense data) and motor output. This refined version of functionalism allows us to integrate evolutionary and ontogenetic aspects of the cognitive system as they are introduced by teleological versions of MR. Due to the embedding in the real world, behavior can be described as a mapping from one state of affairs to another state of affairs, e.g. from finding normal food in arm2 in the morning to finding chocolate in arm7 at midday. Further parts of the story are adapted from Putnam's functionalism: since one state of affairs does not always lead to the same (kind of) state of affairs, we need to assume internal states that enter the mapping function as arguments and values. Thereby, the internal states (or aspects of them) can be identified as the things that play the same role as the external objects (in the states of affairs), and thus can be said to represent them. In particular, as mentioned above, one central case for introducing mental representations is the case of misrepresentation, especially when a certain behavior is elicited although the stimulus is not there. Misrepresentations can happen in two ways in our case study: firstly, if the rat finds normal food in arm2 in the morning and thus expects chocolate in arm7 at midday without there being chocolate (misexpectation of the relevant consequence); secondly, if the rat misremembered there being food in arm2 in the morning (misrepresentation due to wrong encoding of the relevant trigger for an expectation) and thus approached arm7 at midday. These cases of misrepresentation can occur because the places that the representations take in the functions described above could – in principle – also be taken by states of affairs. However, since the states are not directly available to the rat (one is to be remembered, the other to be predicted), representations have to take the places and the functional roles of these states of affairs. Given a series of control experiments, it is convincing to presuppose that a structured representation of food type, location and time is involved in the informational processing (Crystal 2013) at those places, which could – in principle – be played by the food type, location and time directly.

5.2. The use-dependence of MRs

The content of a MR cannot be determined on the basis of intrinsic properties alone; rather, the "use" of the MR that leads to a certain behavior is decisive. More specifically, the behavior of a cognitive system involves a target which is the goal of a goal-directed behavior. The target of the behavior is the coarse-grained content of the MR, i.e. what the MR is about. What exactly the goal of the behavior is, is in turn dependent on how the behavior is described, i.e. on what is chosen as the explanatory goal (see also above). In this sense, it makes a huge difference whether we describe the behavior of the ant

[12]Whether the function is interpreted as describing causal relations such that the "functional role" turns out to be a "causal role" is not crucial here. Indeed, interpreting the functional role as a causal role is the most obvious way to combine the functionalist framework with a materialistic ontology.

as a triangle completion behavior or as a homing behavior. This might sound as if we could freely choose between different equally valuable descriptions. This is, however, not the case; it is the complexity of the relevant mechanisms and how it is realized in the world that makes the difference. Some descriptions are to be favored over others, ultimately depending on whether they give us an explanatory advantage given the cognitive system in the world, i.e. if they add something to our explanation that is not available under a different description (cf. Ramsey 2007).

Let us illustrate this with the example of the ant and the frog. We show that the frog's cognitive system is so rigid that different descriptions are not helpful in contrast to the ant's cognitive situation. In the frog's case the behavior can be described as a tongue-throwing behavior with the frog's goal to hit the thing that elicited the behavior, or as a fly-catching behavior with the frog's goal to catch flies. We can systematically distinguish the two ways of describing the behavior, since different cases count as unsuccessful: catching a black moving dot is a successful tongue-throwing behavior but can be an unsuccessful fly-catching behavior if the black moving dot is not a fly. However, as far as we know, the frog's reaction to unsuccessful and successful cases is not differing in an interesting way. Thus, the more demanding description as a fly-catching behavior (in contrast to tongue-throwing behavior) does not seem to add anything interesting to our explanation of the frog's behavior with respect to the underlying mechanism; it informs us about the normal evolutionary embedding of the mechanism without adding anything about the structure or functioning of the mechanism. (The addition that the tongue-throwing behavior is advantageous for the frog because most black moving dots around it are flies in its natural habitat works independently of the type of description.)

In the case of the ant, things are different: the behavior can be described either as a triangle-completion behavior with the ant's goal to return to its starting point, or as a homing behavior with the ant's goal to return to its nest. And these descriptions have a different explanatory value: The description of the ant's behavior as a triangle-completion behavior leaves no room to describe the systematic difference between cases in which the nest is reached, and cases in which the triangle is completed (the "correct" vector has been followed, e.g. after being displaced) but the nest is not reached. The difference is that in the latter but not the former case, the ant starts a typical nest-searching behavior, i.e. moving in growing circles around the endpoint of the triangle completion. Thus, describing the ant's behavior as a homing behavior brings an explanatory advantage: It allows us to systematically distinguish between successful cases and unsuccessful cases since only the latter are followed by specific "repair" mechanisms (or other additional features shaping the mechanism). Describing the ant's behavior as a homing behavior adds relevant information about the structure of the triangle completion mechanism, namely that it includes a local search mechanism activated only in the case of not reaching the nest after a rough completion of the triangle. Thus, the description of the ant's behavior as a flexible behavior is not a matter of taste, it is a matter of the best scientific description that optimally reflects the important factors. In the frog's case, there is no reason to assume that factors beyond black moving dots play a role. In the ant's case, however, we have good reason to

assume that finding the nest is a very relevant factor that determines what the ant is doing in cases of failure. Thus, the latter case involves a mental representation of the nest while the frog can be (properly) described without involving mental representations.

5.3 The format of MRs: The adequacy relation

While the coarse-grained content of a MR is determined by actual behavior (under normal conditions) being directed at the target of the behavior,[13] the adequacy relation is determined by the function in which the representation is used as a stand-in for the target. Let us elaborate our view by using the general idea of a substitution-relation within our functionalist framework: An internal state (or process) functions as a substitute for an object, property etc. if the activation of this internal state alone triggers an informational processing that *is equivalent to* the processing that follows the registering of the actual presence of the object (or property).[14] The substitution-relation (e.g. the internal state substituting the registering of an actual object) explains the coarse-grained content of a mental representation and thereby allows us to differentiate between representations and non-representations: any inner state (or process) that can substitute something equivalently in a function describing the system's behavior is a mental representation, everything else is not. To determine the coarse-grained content of a mental representation thus involves determining equivalence in information processing. At this point, a pragmatic research perspective comes into play: the behavior has to be described as a function, which is in principle possible on different levels (see above); thus there are many levels of equivalence (even if the represented object stays the same). To account for this conceptually, we introduce a third feature of MR, namely the format of MRs. "Format" is not used here to refer to different modalities (e.g. visual in contrast to tactile); our aim is to highlight different ways of structuring information about the same object or situation (s. below). The vehicle of an MR is typically a neural state of the cognitive system. Its coarse-grained content is typically the target of the behavior to be explained (given an explanatory interest) and it is determined by the functional role the vehicle takes over in the function that describes the behavior of a cognitive system. The format of the representation is the fine-grained structure of the MR, which is determined by a certain level of equivalence in the substitution relation. The respective relation between the substitute (representation) and the substituted (represented) object is called the *adequacy relation* for the substitutional role. Intuitively, something can be a substitute for something else only if it has the right properties to fulfill the same role as the represented entity. And what are the "right properties"? They are the properties

[13] This is the place where teleosemantic ideas kick in into our account.

[14] In the case of the ant, the nest cannot be registered; instead, the number of steps and the angle to the sun is registered, and this registration then stands in for the location of the nest, thus represents the nest. Or take Millikan's famous example: the beaver uses the registration of the sound of another beaver's splash in a process that would equally well function with a registration of an endangering predator; thus, so the idea, the registration of the splash represents danger of being attacked by a predator.

that the function operates on.[15] In this sense, a mental representation is adequate to substitute whatever it represents just in case it has the right properties to fulfill the same role in the function describing the behavior. The right properties are determined by the relevant function itself. And since there are functions on different levels of cognition, we can distinguish different kinds of adequacy relations leading to different kinds of mental representations at different cognitive levels depending on the explanatory interest (Vosgerau 2008, 2009; Vosgerau 2011).[16]

5.4 Different formats of MR and the ontology of MR

A mental representation can be characterized by the vehicle, the format, and the content which is related to a type of behavior of a cognitive system.[17] The relation between these notions is as follows: if the vehicle of an MR (e.g. a neural correlate) is fixed relative to the type of behavior that is to be explained, and if we determine the relevant format (which involves a pragmatic selection of a relevant level of adequacy condition), then the content of the MR is constrained enough to play an explanatory role in scientific explanation. In more detail, our view presupposes that there is one (or more) relevant underlying mechanism(s). The vehicle of the MR is explanatory since it is part of a causal mechanism. The content of the MR is explanatory since it is systematically related to the vehicle, even though it is not part of the mechanism. The content-involving explanations are coarse-grained causal explanations which can be found at higher levels of description: either the content-involving explanation is a coarse-grained description of many (possible and real) underlying low-level mechanism sharing explantorily relevant factors (Vosgerau and Soom 2018), or it is a description of a higher level causal mechanism and is in this sense coarse-grained (cf. Krickel 2018).[18] Content involving explanations are causal explanations since they are finally related to and anchored in the causal mechanism(s).

We need to elaborate on the format of representation. To do this, we describe three formats of MR in a general way as correlational, isomorphic, and propositional. Different formats are connected with different types of representation relations which allow us to describe the same behavior in different ways.

The correlational format of a mental representation (MR): *r* (realized by vehicle *v*) is a MR of an entity *x* with respect to the type of behavior *B* (which is described by a certain function) in a

[15]Crucially, we here depart from a purely mathematical understanding of the function towards a more causal or mechanistic understanding of it. However, we try to stay as abstract as possible in order to be compatible with different interpretations of the function.

[16]Our notion of the format of an MR is not introduced to account for a difference between a sensorimotor-based and a perception-based vehicle of an MR. This distinction is not irrelevant, but the adequacy relation described above is more basic.

[17] For an adequate explanation, we do not always need a specification of all aspects of a MR. Sometimes a coarse-grained description without specifying the format is sufficient.

[18] The cited literature describes detailed accounts to prevent the famous causal exclusion argument: It is compatible with either having only low-level mechanisms but several levels of causal explanations or with also allowing for high-level causal mechanisms within a mechanistic framework.

correlational format iff *r* is substituting *x* in the function describing *B* (i.e. if *r* is taking over the functional role of *x*), and *r* is adequate with respect to its substituting *x* only if it covaries with *x*.

The isomorphic format of an MR: *r* (realized by vehicle *v*) is a MR of an entity *x* with respect to the type of behavior *B* in an isomorphic format iff *r* is substituting *x* in the function describing *B*, and *r* is adequate with respect to its substituting *x* to the degree that it is isomorphic with the relevant part of the structure of *x*, where the relevant part is determined by the functional roles in *B*.

The propositional format of an MR: *r* (realized by vehicle *v*) is an MR of a state of affairs *x* with respect to the type of behavior *B* in a propositional format iff *r* is substituting *x* in the function describing B, and *r* is adequate with respect to its substituting *x* only if *r* can be attributed the propositional content that *p* and *p* refers to the state of affairs *x* (or equivalently: *p* is true iff *x* is the case).

We can apply these different types of representational formats to explain the same everyday behavior, e.g. Sandra grasps her car key and her phone on top of which the key is positioned, in order to drive home. This goal-directed grasping behavior can be explained with a MR of correlational format by highlighting the proximal goal, i.e. the key-phone unit. Then the claim is that the cause of the behavior involves an MR of the key-phone unit: this means that there is a neural correlate as part of the causally relevant neural processes which stands in for the key-phone unit as the goal of the grasping behavior, e.g. the neural correlate of a percept or an efference copy.[19] Thus, the relevant explanation would basically run as follows: Sandra is grasping the key and the phone at once because her visual apparatus presents both to her as a key-phone unit. This is a narrow explanation of the grasping behavior focusing on one central correlation with the relevant object-unit of the behavior.

The behavior can also be *explained using an MR in an isomorphic format.* Then the claim is that the cause of her behavior involves at least an MR of the key and an MR of the phone: this means that there are at least two neural correlates as part of the causally relevant neural process, one standing in for the phone, the other for the key. This would be a minimal structure to allow for an isomorphic substitution. Usually this would also involve the isomorphic substitution of relevant properties, e.g. the property of the key to open her car, the property of the phone to enable her to make a call while driving home. The objects and properties are thought to be implemented as object files in the neural system (Kahnemann et al. 1992; Spelke 2000; Carey 2011). Thus, the relevant explanation at this level would run as follows: Sandra is grasping the key and the phone at once because there is an activation of a neural correlate of a key and its property to open her car and of the neural correlate of the phone and its property to allow for phone calls. Thus, the affordances of the objects, to open a car and to enable phone calls, are a relevant aspect of grasping them.

[19]The underlying cognitive vehicle could be e.g. an efference copy of the goal of the grasping movement as is presupposed according to the comparator model of explaining simple goal-directed grasping behavior (Wolpert and Flanagan 2001; Synfozik et al. 2008).

The most complex level of explanation uses a *propositional format of MRs*: then the claim is that the cause of Sandra's behavior involves at least an MR of a proposition *p* (a state of affairs) such that there is a neural correlate that has a functional role adequately constrained by the propositional content or equivalently a cluster of behavioral dispositions, i.e. a neural correlate of a thought is activated that can be characterized by the same behavioral dispositions that would be activated if Sandra had registered the state of affairs expressed by *p*. This results in the usual folk-psychological explanation: Sandra grasps her phone together with her key because she wants to open her car to drive home and she wants to make a phone call during her way home and she has the relevant background information about the affordances of her key and her phone.

We can explain the same behavior on different levels using specific formats of mental representation for each of these explanatory levels, thereby describing different contents of MRs for the same behavior. What does this imply for the ontology of MRs? Isn't it clear that MRs can only be pragmatic stories – just a narrative 'gloss', according to Egan (2014), and not something closely related to and systematically anchored in causal mechanisms? Here we argue that despite the different levels of explanation, and the three types of MRs which can be used to explain the same behavior of a cognitive system, MRs remain systematically coupled to reality. Explanations using MRs can indeed be scientific explanations. The reason is that different aspects of reality can be described on different levels, which in turn requires different levels of explanation.

Take, for example, a car traveling at 50 km/h. We can describe this "behavior" on different levels and accordingly explain it on different levels. The question "Why is the car traveling at 50 km/h?" could be answered by saying that a force in the direction of travel is acting on the car that is equal to the friction of the car at exactly this speed. This is quite abstract, but not unscientific, and it applies to a lot of cases, including cases of towed cars. Of course, a more detailed answer could be that an air-gas-mixture is exploding in the cylinders, thus moving the pistons such that they transfer the force onto the driving rod etc. This level of explanation is surely more detailed, but it is not more scientific: it only applies to fewer cases. If we drill down into more detail about the molecular structure of the gas and the rod, our explanation will still not gain scientific-ness; it rather loses generality, which means that it gets more and more specific. In a similar way, behavior can be described at different levels – the more detailed the description gets, the fewer cases are subsumed and the more detailed the explanation has to be. However, this does not mean that the explanation also becomes more scientific.

To clarify the role of mental representations, let us add a general remark concerning the relation between mechanisms and MRs.[20] If a behavior seems to be flexible but has an underlying rigid mechanism combining only the same stimulus with the same output behavior, then a generalization which goes beyond that mechanism itself is not explanatorily fruitful – i.e. we do not need to presuppose mental representations in a scientific explanation of the behavior. Then we would define

[20] We do not have an own theory of mechanisms in the background but our theory of mental representations fits with the characterizations in Krickel (2018).

the resulting behavior not as flexible but as rigid, although it seems flexible. But if we actually have good reasons to suppose that the behavior either relies on variable or even multiple mechanisms, then this multiple realization makes a generalization at the level of common interrelations between triggering situations and the resulting behavior explanatorily fruitful (at least if we are interested in explaining and predicting the behavior not only for one specific situation but also for a large variety of situations and challenges).

Let us illustrate the fact that MRs are nonstatic and thus have multiple realizations by an example that clarifies our view about the vehicle of MRs: for quite a while it was thought that amygdala activation is a neural mechanism that is constituting the emotion of fear. Then it was discovered that it also is activated when observing someone else‘s fear. Furthermore, it was discovered that it is often active in cases of intense negative emotions in general, and that amygdala activation can be observed in extroverted people observing happy faces, but not in introverts doing so (Canli et al. 2002). It is now commonplace that the amygdala is involved in several functions of the body including arousal, autonomic responses associated with fear, emotional responses, hormonal secretions, and memory. To make the picture even more complex, there is a study that demonstrates that a person can experience intense panic with strong damage of bilateral amygdala, namely the patient S.M. who lost her amygdala due to Urbach-Wiethe disease. Interestingly, S.M. was first mentioned as evidence for amygdala as the center of processing fear since S.M. showed an absence of overt fear manifestations in typical fear inducing situations (including exposing her to living snakes and spiders, taking her on a tour of a haunted house, and showing her emotionally evocative films) and an overall impoverished experience of fear (Feinstein et al. 2011). But first of all, fear was still processed although to a diminished degree, and secondly, under certain conditions she experienced intense panic (Feinstein et al. 2013). Thus, we can conclude that amygdala activation is neither necessary nor sufficient for experiencing fear.

While starting out with a candidate for a neural correlate of fear, we ended with a neural correlate which is important to process fear but nevertheless is involved in the realization of many other mental abilities. If we complement this bottom-up perspective with a top-down one, we should start with the phenomenon of fear and try to work out the underlying features realizing a token of fear: We have argued elsewhere in detail (Newen et al. 2015) that fear is realized as an integrated pattern of characteristic features none of which is sufficient and most of which are not necessary to process fear. Thus, there is no unitary mechanism that realizes fear. And we think that this observation is neither exceptional nor due to a premature state of neuroscience but typical for the architecture of neural mechanisms realizing a minimally flexible behavior or cognitive phenomenon.

Without being able to comprehensively defend this view as a general view in this paper, we want to mention two further reasons why this observation of multiple realization which still remains systematic is generalizable: 1. Most biological systems displaying flexible behavior are those that are

able to learn; thus, their neural vehicles change systematically during the learning processes. 2. More generally, the neural processes have a high level of neural plasticity which is due to permanently ongoing reorganizations in the brain (Hübener and Bonhoeffer 2014). Thus, complex biological systems regularly implement flexible behavior. Hence, the generalized level of content explanations is epistemically fruitful and is indeed often the only scientific means to access minimally interesting generalizations in order to explain the behavior.

5.5 Features of situated mental representations, some applications and advantages

Let us summarize the main features of situated mental representations: a MR is individuated by three aspects, namely vehicle, content and format relative to a relevant minimally flexible behavior. We have already seen the interdependence of the three aspects which is always relative to the main job of MRs, i.e. to explain a generalized type of flexible behavior. Furthermore, we have illustrated that the same type of behavior *B*, e.g. Sandra grasping her car key and her phone in order to drive home, can be explained on three different levels by using different formats for the MRs which share the same vehicle and the same coarse-grained content. This flexible way of explaining a behavior B has its limits. The first danger is that of anthropomorphizing or over-intellectualizing nonlinguistic behavior of animals and babies, e.g. the homing behavior of desert ants, by ascribing propositional content to them. This would be empirically inadequate since we do not have any additional evidence that their behavior can be based on such a generalized cluster of mechanisms which is compatible with belief ascription. The mechanism of dead reckoning or path integration is much more constrained. The second danger is that of over-intellectualizing behavior of adults. Grasping the car key and the phone can be based on an explicit consideration of planning to drive home now and to make an urgent phone call during the drive or it can be just a well-established habit. If the latter is the case, then a correlational or isomorphic MR may be much more adequate to describe the behavior than a propositional MR. Thus, despite the flexible way of using different formats of representation to explain the same behavior, in many situations it is still possible to determine the most adequate description which makes this behavior typical for a certain representational format. Here are some examples to illustrate this shortly:

Typical Correlational MRs: This is the most adequate representational format to explain the homing behavior of desert ants: there is a neural process which is systematically correlated with the spatial vector and thus with the location of the nest under favorable conditions.[21] A further example is the rather complex bee behavior. Bees make also use of path integration – as described already for desert ants – but to perform the famous waggle dance, they clearly rely on memorized representations of the

[21] Our interpretation of the complexity of the neural processing in desert ants and bees is such that we take it to involve MRs. Furthermore, we interpret what science knows about the homing behavior of ants and bee navigation such that the neural vehicles are not structured in an isomorphic way with clear separable components as it is clearly the case in rats and birds with their what-where-when information. Rather, we interpret them as merely covarying with the location of the nest and the food source, respectively. Space constraints prevent us from clarifying the lower bound of MRs and the borderline between correlational and isomorphic MR.

profile of the position of the sun: This case is quite convincing to illustrate the ‘substitution relation’: The bees extract from the incoming signal of the sun in a situation a representation of the sun’s position and seem to build a profile of its positional changes. Thus, the neural processing of the bees involves an internal informational memory of this changing positions. Even if the actual information about the relation between the direction of the source from the hive and the direction of the sun from the hive was acquired many minutes earlier, bees use the correct calculated actual position of the sun when doing the waggle dance. This is based on a presupposed substitutional memory that is required to carry this direction vector forward in time (Menzel and Eckoldt, 2016, see also the description in Gallistel 2017). This representation is even accurate and adjusted in time when bees only are exposed to the sun in the afternoon for quite a while and then are tested in a morning situation (Menzel and Eckoldt 2016). Thus, bee behavior can best be explained by presupposing a correlational mental representation realized by a cluster of neural processes which implement a vector representation standing in for the direction and the distance of the food source.

Typical Isomorphic MRs: Here we only remind the reader of the example of episodic memory realized in rats (Crystal 2013) or in scrub jays (Clayton and Dickinson 1998, Seed et al. 2009). Our key example illustrates that rat behavior can be best described by a cluster of neural processes implementing substitutes of the time, the location and the type of food. Thus, we have to presuppose isomorphic mental representations to explain and predict their behavior.

Typical Propositional MR: The so-called explicit false belief task tests whether children are able to distinguish their own belief, e.g. about the location of an object, from the belief about another person concerning the location of the same object. If the other person has left the room before the object was transferred from box A to box B and we ask four-year-old children where the other person will search for the object when coming back, then the four year olds are able to account for the false belief of the other person, answering that the other person will search in box A even if the object is in box B. This behavior typically demands the ascription of a belief, considering what the other person knows in contrast to me and thus is a paradigmatic case of typically involving a propositional format of representation.

Mental representations can be used in two strategies of content-explanations which we want to call ‘top-down’- and ‘bottom-up’ strategies. *A top-down strategy* investigates a certain flexible behavior, e.g. grasping a bottle of water from the fridge, and wants to explain this by relying on the relevant content, e.g. the content of the person’s belief that a bottle of water is in the fridge combined with the desire to drink some cold water. Using such a folk-psychological explanation, we aim to constrain the vehicle of this MR in a scientific way. Thus, we want to disclose the neural correlate of the relevant type of belief (and the relevant desire). But in the case of a type of belief, we did not find and cannot even expect to find any rigid underlying neural correlate embedded in a rigid mechanism, but neuroscientific evidence is much more complex: the best we can expect is a cluster of neural correlates embedded in one quite contextually varying mechanism or even embedded in a plurality of

mechanisms. Thus, even if we succeed to characterize scientifically the underlying cluster of neural correlates as well as the relevant mechanisms, the folk-psychological content-explanation (constrained by the format of the MR), remains explanatorily fruitful as describing the possible range of underlying neural correlates and mechanisms at a general level.[22]

A bottom-up strategy of a content explanation would ideally start with the discovery of neural mechanisms that are only known to be associated with the relevant flexible behavior of grasping a bottle of water from the fridge, e.g. the comparator mechanism (Synofzik et al. 2008; Newen 2017). Then the challenge is to constrain the content of some neural correlates or cluster of neural correlates involved in the mechanism. This can be done by first fixing a level of explanatory interest, i.e. by determining the format of MR, and then aiming to disclose the correlation between some neural states and processes and the external states and properties in the world. Then we can develop a content-explanation which is again having the function of constraining the possible underlying mechanisms in the cognitive system. Both strategies are used in cognitive sciences and illustrate the role of content explanations as systematically connected to underlying (clusters of) neural correlates and mechanisms such that content explanations are scientific but only with a rather broad range of generality in contrast to mechanistic explanations involving neural mechanisms which are rather specific.

Let us finally take up the challenge of radical enactivists who argue that we do not need MRs to explain flexible behavior and that dispositions would suffice to do this (Hutto and Myin 2017). Using the case of episodic memory in rats introduced above, it is supposed to be sufficient to describe the disposition *to search for chocolate in arm7 at midday if there is normal food in arm2 in the morning.* According to our view, a dispositional explanation is not in conflict with an explanation of behavior based on MRs. A dispositional explanation can be sufficient if one is not interested in the architecture of cognitive processes underlying the dispositions. However, if we are able to discover the main features of this architecture, we can, in principle, (i) make more fine-grained and more general explanations and predictions about the behavioral dispositions, and (ii) change the perspective from observation to intervening into the cognitive system. It follows from our assumptions that, (ad i) if one knows the (relevant) vehicle (e.g. the neural correlate) of episodic memory, one can predict that the rat will show or remains disposed to show a certain search behavior as soon as this vehicle (or a certain group of vehicles) will be activated again even in a rather different context. Furthermore, (ad ii) knowing the (relevant) vehicle in a form of a neural correlate (eventually plus some embodied or social properties), allows us to modify the behavioral dispositions. The latter is what we aim for in the case of behavioral deficits. E.g., a person with deviant behavior due to a brain tumor, (the patient developed pedophilia due to a tumor), is of course treated by removing the brain tumor, which is a

[22] These explanations can be causal explanations since we can allow for different relevant levels of causal explanation even if there is only one level of a causal mechanism. Causal explanations need to be anchored in a causal mechanism but this can be the case more or less precisely depending on the level of description of the causal explanation (If there are worries about the classical problem of mental causation and overdetermination see for discussion Newen and Cuplinskas 2002).

change of the neural basis and thereby the behavioral dispositions; in this case after the removal of the tumor the pedophilia was gone (Burns and Swerdlow 2003). This coarse-grained successful intervention only illustrates the general strategy of intervening; other methods include psycho-pharmacological drugs in the case of mental disorders to treat positive symptoms of schizophrenia. We know that these are nowadays still rather coarse-grained interventions and we also take the idea of embodied and social anchoring of MRs serious, e.g. social interventions are still the most effective means of treatment of mental disorders; but the general conclusion is that it is fruitful to presuppose MRs since they can be used to explain and predict behavior and to systematically intervene. Our account predicts that the development of neuroscience will open new possibilities of intervention; however, we also predict that these possibilities have to investigated in close relation to relevant aspects of the whole body and the social environment, which is made possible by integrating them into explanations of behavior with MRs.

6 Conclusions

We are in need of mental representations to explain or predict minimal flexible behavior. This is behavior which by definition cannot be explained by a rigid mechanism. Given the evidence from neuroscience, this is very often the case since many mental phenomena are realized either by a contextually rather variable mechanism or by a plurality of mechanisms. The latter observation is a main reason for enactivists to take an anti-representationalist position and to constrain all explanations of nonlinguistic cognitive phenomena to explanations based on affordances and dispositions. We are not denying that explanations suggested by enactivists can be enlightening but argue that their anti-representationalism is an ideological constraint on the plurality of scientific explanations. Furthermore, we developed a new notion of situated mental representations according to which a mental representation is individuated by three aspects, namely vehicle, content, and format relative to a relevant minimally flexible behavior to be explained. Thus, mental representations do not have their content intrinsically but only relative to an explanatory level; and this content can be characterized by (at least) three formats of representation.

Egan (2014) argues that any content that is not intrinsically attached to a mental representation cannot be taken scientifically serious but is only a pragmatic gloss. We argue that all types of content are suffering from the problem of underdetermination and thus there is never an intrinsic connection between content and mental representation. But we argued that content-based explanations can nevertheless be scientific. We need to and can distinguish non-explanations from explanations of behavior and in the remaining package of explanations we have a great variety of high-level and low-level explanations with different advantages: low-level explanations involving specific mechanisms are typically much more constrained while high-level explanation involving mental representations are typically much more general: the level of generalization can be described in more detail with our notion of representational format. Thus, mental representations can be part of scientific explanations if

they are characterized by vehicle, format and content relative to a flexible behavior of a cognitive system. Mental representations understood in such a way are dependent on the explanatory level of interest, allow for multiple realization in a cluster of vehicles and thus have to be characterized as nonstatic and flexible. If we want to account for mental representations as being used in scientific explanations and realized by neural correlates in biological systems, then we have to give up the traditional Fodorian view of rigid symbolic mental representations. At the same time, we do not have to throw out the baby with the bathwater and accept anti-representationalism. Instead we offer a notion of mental representations that enables them to take part in scientific explanations while they are real, nonstable, use-dependent and situated. They are dependent on the levels of explanation, allowing for multiple realization and modifications of the realization basis due to learning and plasticity. Furthermore, such mental representations are a necessary part of the progress of cognitive science in bringing bottom-up and top-down explanations of the same phenomenon closer together.

References

Bell, V., and Halligan, P.W. (2013) The neural basis of abnormal personal belief. In: Krüger, F., and Grafmann, J. (2013). *The Neural Basis of Human Belief Systems*. Psychology Press.

Bjerknes, T. L., Dagslott, N. C., Moser, E. I., and Moser, M. B. (2018). Path integration in place cells of developing rats. *PNAS*, *115*(7), 1637–1646.

Bolhuis, J. J., Tattersall, I., Chomsky, N., and Berwick, R. C. (2014). How could language have evolved? *PLoS Biology, 12(8)*, e1001934.

Burke, S. N., Maurer, A. P., Hartzell, A. L., Nematollahi, S., Uprety, A., Wallace, J. L., and Barnes C. A. (2012). Representation of three-dimensional objects by the rat perirhinal cortex. *Hippocampus*, *22*(10), 2032–2044.

Burns J. M., and Swerdlow R. H. (2003). Right orbitofrontal tumor with pedophilia symptom and constructional apraxia sign. *Arch. Neurol. 60*, 437–440. DOI: 10.1001/archneur.60.3.437.

Canli, T., Sivers, H., Whitfield, S.L., Gotlib, I.H., and Gabrieli J.D.E. (2002). Amygdala Response to Happy Faces as a Function of Extraversion, *Science*, *296*, 2191.

Carey, S. (2011). *The origin of concepts* (1. iss. Oxford Univ. paperback). *Oxford series in cognitive development*. Oxford: Oxford Univ. Press.

Carruthers, P. (2006). *The architecture of the mind*. Oxford: Oxford Univ. Press.

Clayton, N.S., and Dickinson, A. (1998). Episodic-like memory during cache recovery by scrub jays. *Nature* 395, 272-274.

Coolidge, F. L., and Wynn, T. (2011). The implications of the working memory model for the evolution of modern cognition. *International Journal of Evolutionary Biology*.

Crystal, J. D. (2013). Remembering the past and planning for the future in rats. *Behavioural Processes*, *93*, 39–49.

Dretske, F. (1988). *Explaining behavior: Reasons in a world of causes.* Cambridge, MA: The MIT Press.

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

Fang (2018). The case for multiple realization in biology. *Biology & Philosophy* 33:3.

Feinstein, J.S., Adolphs, R., Damasio, A., and Tranel, D. (2011). The human amygdala and the induction and experience of fear. *Current Biology, 21*, 34–38.

Feinstein, J. S., Buzza C., Hurlemann R., Follmer R.L., Dahdaleh N.S., Coryell, W.H., Welsh, M.J., Tranel, D., and Wemmie. J.A. (2013). Fear and panic in humans with bilateral amygdala damage. *Nature Neuroscience, 16(3),* 270-272.

Fodor, J. (1980). Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioral and Brain Sciences, 3(01),* 63.

Fodor, J. (1987). *Psychosemantics. The Problem of Meaning in the Philosophy of Mind*. The MIT Press: Cambridge/London.

Fodor, J. (1994). Fodor's Guide to Mental Representation. In S. Stich, eds., *Mental Representation. A Reader* (9–33). Cambridge MA, Oxford: Blackwell.

Gallistel, C.R. (1990). *The organization of learning*. A Bradford Book: Cambridge, MA.

Gallistel, C.R. (2017). Learning and Representation. In *Learning and Memory: A Comprehensive Reference* (pp. 141–154). Elsevier.

Gigerenzer, G. (2000). *Adaptive thinking: Rationality in the real world*. New York: Oxford University Press.

Haugeland, J. (1991). Representational Genera. In W. Ramsey, S. Stitch, and D. Rumelhart, eds., *Philosophy and Connectionist Theory* (61–89). Hillsdale, NJ: Erlbaum.

Hübener, M., and Bonhoeffer, T. (2014). Neuronal Plasticity: Beyond the Critical Period. *Cell 159(4),* 727-737.

Hutto, D.D., and Myin, E. (2013). *Radicalzing enactivism: Basic Minds without Content.* Cambridge, MA: MIT Press.

Hutto, D.D., and Myin, E. (2017) *Evolving enactivism: Basic Minds meet Content*. Cambridge, MA: MIT Press.

Jezek, K., Henriksen, E. J., Treves, A., Moser, E. I., and Moser, M. B. (2011). Theta-paced flickering between place-cell maps in the hippocampus. *Nature 478(7368)*, 246-249. doi: 10.1038/nature10439.

Kahneman, D., Treisman, A., and Gibbs, B. J. (1992). The reviewing of object-files: Object specific integration of information. *Cognitive Psychology*, *24*, 174–219.

Káli, S., and Dayan, P. (2004). Off-line replay maintains declarative memories in a model of hippocampal-neocortical interactions. *Nature Neuroscience, 7*, 286–294.

Krickel, B. (2018). *The Mechanical World - The Metaphysical Commitments of the New Mechanistic Approach*, Springer.

Menzel, R., and Eckoldt, M. (2016). *Die Intelligenz der Bienen: Wie sie denken, planen, fühlen und was wir daraus lernen können* (2. Auflage). München: Knaus.

Millikan, R.G. (1996). Pushmi-pullyu Representations. In J. Tomberlin, eds., *Philosophical Perspectives* 9 (185-200). Atascadero CA: Ridgeview Publishing 1996.

Millikan, R.G. (2009). Biosemantics. In B. McLaughlin, A. Beckermann, and S. Walter, eds., *The Oxford Handbook of Philosophy of Mind* (394–406). Oxford University Press: Oxford.

Moser, E. I., Kropff, E., and Moser, M. B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annual Review of Neuroscience*, *31*, 69–89.

Neander, K. (2017). *A Mark of the Mental*. Cambridge: MIT Press.

Newen, A. and Cuplinskas, R. (2002). Mental Causation: A real phenomenon in a physicalistic world without epiphenomenalism or overdetermination. *Grazer Philosophische Studien* 65, 139-167.

Newen, A. (2017). What are cognitive processes? An example based-approach. *Synthese, 194 (11).* 4251-4268 (online first 2015). DOI: 10.1007/s11229-015-0812-3.

Newen, A., Welpinghus, A., and Juckel, G. (2015). Emotion Recognition as Pattern Recognition: The Relevance of Perception. *Mind & Language 30(2)*, 187-208.

Panoz-Brown, D., Corbin, H.E., Dalecki, S.J., Gentry, M., Brotheridge, S., Sluka, C.M., Jie-En Wu, J. E., and Crystal, J.D. (2016). Rats Remember Items in Context Using Episodic Memory, *Current Biology*, *26*, 2821–2826.

Packard M.G., and McGaugh, J.L. (1996). Inactivation of Hippocampus or Caudate Nucleus with Lidocaine Differentially Affects Expression of Place and Response Learning. *Neurobiology of learning and memory, 65*, 65–72.

Polger, T. W., and Shapiro, L. A. (2008). Understanding the dimensions of realization. *The Journal of Philosophy, 105(4)*, 213-222.

Putnam, H. (1975). "The Nature of Mental States." In *Mind, Language, and Reality*. Cambridge: Cambridge University Press.

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

Ramsey, W. M. (2015). Must cognition be representational? *Synthese*, 1–18. [Deutsche Übersetzung: Muss Kognition repräsentational sein? In: T. Schlicht, and J. Smortchkova, eds., *Mentale Repräsentationen: Grundlagentexte* (4197–4214) (2018). Suhrkamp Verlag.]

Schulte, P. (2012). How Frogs See the World: Putting Millikan's Teleosemantics to the Test. *Philosophia, 40(3),* 483–496.

Seed, A. M., Emery, N., and Clayton, N. (2009). Intelligence in corvids and apes: A case of convergent evolution? *Ethology, 115*, 401–420.

Sober (1999). The Multiple Realizability Argument against Reductionism. *Philosophy of Science* 66, no. 4 (Dec., 1999): 542-564.

Spelke, E. S. (2000). Core knowledge. *American Psychologist, 55*(11), 1233–1243.

Starzak, T. (2017). Interpretations without justification: a general argument against Morgan's Canon. *Synthese, 194(5)*, 1681–1701.

Synofzik, M., Vosgerau, G., and Newen, A. (2008): Beyond the comparator model: A multifactorial two-step account of agency. *Consciousness and Cognition, 17*, 219-239.

Tolman, E.C., and Honzik, C.H. (1930). Insight' in Rats. *University of California Publications in Psychology* 4: 215–32.

Tolman, E. C. (1948). Cognitive Maps in Rats and Men. *Psychological Review* 55: 189–208.

Tomasello, M. (2008). *Origins of Human Communication.* MIT Press.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

Tsao, A., Sugar, J., Lu, L., Wang, C., Knierim, J. J., Moser, M. B., and Moser, E. I. (2018). Integrating time from experience in the lateral entorhinal cortex. *Nature*, *561*(7721), 57–62.

Vosgerau, G. (2008). Adäquatheit und Arten Mentaler Repräsentationen. *Facta Philosophica 10*, 67–82.

Vosgerau, G. (2009). *Mental Representation and Self-Consciousness. From Basic Self-Representation to Self-Related Cognition*. Paderborn: mentis.

Vosgerau, G. (2010). Memory and Content. *Consciousness and Cognition, 19*, 838–46.

Vosgerau, G. (2011). Varieties of Representation. In A. Newen, A. Bartels, and E. Jung, eds., *Knowledge and Representation* (185–209). Stanford, Paderborn: CSLI Publications, mentis.

Vosgerau, G. (2018). Vehicles, contents and supervenience. *Filozofija i drustvo (Philosophy and Society) 29*, 473–488.

Vosgerau, G., and Soom, P. (2018). Reduction Without Elimination: Mental Disorders as Causally Efficacious Properties. *Minds and Machines, 28(2)*, 311–330.

Wolpert, D. M., and Flanagan, J. R. (2001). Motor prediction. *Current Biology: CB*, *11*(18), R729-32.