Semanticizing the brain

Reiteration as a key concept in understanding representation and behavior

Roos Geerse[1]

ABSTRACT

In the past decade neuroscientists have arrived at an understanding of neural representation that potentially sheds a whole new light on a range of phenomena in philosophy as well as psychology. Concretely, I argue that their work on so called engrams suggests that semantic representations are involved in almost all behavior seen in humans and other animals that are sensitive to reward and punishment. Although in the brain there is a division of tasks, the way information is processed at the level of neurons is the same throughout the (mental) system. Part of the information that a neural element codes for 'carries over' to the neural elements it is connected to. I argue that it is thanks to this phenomenon, that I propose to call reiteration, that the various mental subsystems can interact as readily as they do without the help of a central system to regulate their activity. I further argue that this reconceptualization of representation can be combined with insights from the predictive processing and semantic roles approaches and outline the resulting model of intelligent behavior. I submit that the reiteration model shows promise in that it 1) suggests new explanations for important phenomena such as innate concepts and unconscious reasoning, and 2) offers a novel way to harmonize embodied and representationalist approaches to cognition. I conclude by arguing that my analysis shows how empirically informed philosophy can help naturalize the mind by semanticizing the brain.

Key words: engram, mental representation, reiteration, information processing, core concepts

1. Introduction

These are exciting times for everyone who is interested in the workings of the mind as, thanks to the development of new, non-correlational, techniques such as optogenetics, we are at long last starting to see how individual concepts and memories are encoded in the brain (e.g. Ortega-de San Luis & Ryan, 2022). This is good news because it seems reasonable to assume that an empirically informed conceptualization of these encodings, also known as *engrams*, can help us evaluate and in as far as that is possible harmonize the extant theories of representation and behavior. Thus, if we want to arrive at a unifying theory about intelligent behavior that is at once empirically informed, sufficiently detailed, and in terms philosophers

[1] Email: roosgeerse@gmail.com

and psychologists can work with, reconceptualizing representation using these new insights from neuroscience seems a good way to start. In a slogan, naturalizing the mind may well require ‘semanticizing’ the brain.[2] So far, however, very few philosophers have offered their perspective on the topic of engrams with reference to the empirical evidence I will review below, and the ones who have tend to focus on the nature of engrams and their relation to memory, rather than on their workings and their role in behavior (e.g. Najenson, 2020; Robins, 2020). I would maintain, however, that if we want to understand what engrams are, we need to understand how they work. Moreover, the experiments that give us the most information about neural representation investigate (parts of) neural pathways, so the routes from stimulus to behavior, and not memory per se. That is why in this paper I will reconceptualize mental representation primarily by looking at the role engrams play in the generation of behavior. More concretely, I will review some of the new insights into neural representation and combine them with insights from other theories about representation and behavior into a comprehensive account of intelligent behavior that I will complement with a tentative account of the way representations come into being. This sounds more ambitious than it is. As I will look at the way engrams generally work in all animals that are sensitive to reward and punishment, and engrams can be assumed to have a causal role in all behaviors these animals are capable of, the result should be a general account of behavior in these animals, whether this behavior is based on learning or the result of an innate reflex.

Of course, one might still wonder whether it is possible to arrive at any theory by investigating the explanantia without first identifying the explananda. I maintain, however, that at least in this case we can and should. Churchland & Sejnovski (1990) already argued that theorizing at one level of analysis is always constrained by data from all other levels, thus suggesting that identifying the explananda as a first step may be as challenging as taking the explanantia as a starting point. Piccinini & Craver (2011) argue in a similar vein that “psychologists ought to let knowledge of neural mechanisms constrain their hypotheses just like neuroscientists ought to let knowledge of psychological functions constrain theirs.” However, if a conceptualization in psychology is not in line with the neuroscientific evidence, they clearly think that is too bad for psychology. Moreover, they argue that the internal states one posits in a functional analysis should be the states of the system’s components. But how are psychologists or philosophers to come up with good hypotheses at this level of analysis if they not already have a good theory about these components? I therefore submit that, if at all possible, what Piccinini & Craver call mechanism sketches should not be based on a functional analysis of the phenomenon we want to explain but rather on the relevant neuroscientific insights. That way we do not run the risk of reproducing old ideas when we

[2] Note that such a project seems closer to what Dennett (1987/1989) calls “[showing] how a system described in physiological terms could warrant an interpretation as a realized intentional system” (p. 68), than to what Dretske (1995) proposes in his *Naturalizing the mind*, as the latter is willing to assume that “[p]hysically indistinguishable heads can harbor different thoughts” (p. 127).

have empirical evidence that suggests better ones. Moreover, identifying underlying mechanisms may be the only way to reliably carve nature at its joints (Boyd, 1999). For clarity, this is not saying that neuroscientists should ignore what has been found by psychologists or philosophers. They should not, and in many cases it is clear they do not, as they will often use terms, theories, procedures and/or measures that have been developed in neighboring fields. For instance, and as we will see, in many of the studies that have recently been done on engrams researchers used fear conditioning as a method.

Now the question becomes whether our current knowledge of neural mechanisms is such that we may at least try this. I maintain that it is. Techniques such as optogenetics have made it possible to block very specific neural pathways at different points and relate this to the animal's behavior. In other words, the time that neuroscientists had to rely on correlational methods has past. Granted, the most detailed engram studies are done with animals. However, the neural mechanisms and structures we are talking about here may be assumed to be the same in all species that are sensitive to reward and punishment, or at least in mammals. That means that the multiple realizability argument which is often given as a reason to ignore brain processes does not apply.

It would be a different matter if we could assume that the mental forms of representation and information processing are like the non-mental ones, as in that case we could base a theory of the former on our knowledge of the latter. But we cannot. First of all, information has a quantitative and a qualitative side, making neural information processing a form of computation that is unlike other forms of computation (Piccinini & Bahar, 2013). Secondly, something similar may be said about mental representation. Mental representations are not like other representations in the way they are causally relevant (e.g. Rupert, 2018). And because mental representations can function as models, brains are able to solve problems even if there is no system that interprets them (e.g. Ramsey, 2016). In conclusion, we will not be able to understand the role of representation in behavior without some notion of the underlying mechanisms. That is why in this paper I will take as a starting point what neuroscientists in the past two decades have found in their study of engrams and neural pathways, and use these and related insights to answer the two questions about neural representation that I hold to be basic. How is information processed in and by neurons? And how do neural representations come to model whatever they model?

As I will be looking at neural mechanisms to explain mental processes, I will take as my background theory a simplified version of what I consider to be the standard model in neuroscience. In this simplified model all the capacities of the individual are the result of processes that come down to neurons activating other neurons. The neurons that are found in the peripheral nervous system form the link to our bodies, while the ones in the brain handle the information processing involved in all behavior that is not as mechanical as the knee jerk reflex. These brain neurons typically come in ensembles which in turn are organized in at least three ways: in anatomical structures such as the prefrontal cortex and the amygdala, in neural pathways such as the ones for the different modalities in perception (e.g. vision and audition), and in large scale networks such as the default mode network and the salience network.[3] Furthermore, whether a brain neuron will fire depends on the activation of the neurons it is connected to. In other words, what neurons can code for is constrained by their location in the nervous system. Note that, while this model presupposes materialism, it can accommodate embodied theories as well as cognitivist ones. Furthermore, it takes as mental representations, that is representations in the mind/brain, not just semantic representations such as (semantic) concepts and beliefs, but also non-semantic ones such as mental images and motor programs.[4] Somewhat on a side note, that is why in this paper I will use the term *concept* for any mental representation that is binding or bound by other concepts, whether these are semantic elements or not.[5] Only when a structure of concepts is best characterized as an informational element with a truth value, I will use the term *proposition*.

Concretely, my aim is to reconceptualize mental representation in the following way. I will start by summarizing the recent findings on engrams that are most relevant to our project. As a way to characterize these elements I will introduce the concept of reiteration, that is the phenomenon that part of the information that is carried by a (neural) representation carries over to the (neural) representations it is connected to. Next, I will briefly discuss two theories about mental processing that are in line with this conceptualization of representation, namely the predictive processing approach and the semantic roles approach. I will then propose a model in which I combine these theories and the proposed conceptualization of representation, which I will further call the reiteration account of intelligent behavior and complement this with the reiteration account of concept formation to explain how we come by important concepts such as *cause, agent, good* and *bad*. Lastly, I will sum up and briefly discuss this way of combining neuroscience and philosophy.

2. Reconceptualizing representation

In this section I will focus on the studies neuroscientists have done in the past twenty years on what at least some of them call engrams.[6] I will first discuss the older work on what some have called grandmother cells by Quiroga and colleagues and then relate this to the work others have done in the past decade using optogenetics and other techniques that make it

[3] Other terms for *ensemble* are *assembly* and *population*.

[4] For readability, I omit the phrase *representation(s) of*. For instance, when it is clear I am referring to the representation of a concept X, I will simply write *concept X*.

[5] Note that the literal meaning of the word *concept* is “taken together”.

[6] Although the term engram has been in use for over a century, for reasons of scope I cannot go into the turbulent history of this concept. Relatedly, I will not address the question whether representation in natural systems is localist or distributed in the sense these terms are used in the field of artificial intelligence, as it may be argued that the now available neuroscientific evidence has made that discussion largely irrelevant (Roy, 2017).

possible to track individual representations. Lastly, I will introduce the concept of reiteration to explain the processing of information in and by neurons and other neural elements.

The hypothesis that representation is atomic in the sense that for every concept the individual has there is at least one neuron that fires always and only to instances of the category in question was never taken as seriously as is sometimes suggested (Bowers, 2017).[7] Nevertheless, in a series of experiments with participants who had to undergo brain surgery for medical reasons Quiroga and colleagues (2005; 2008) found neurons that almost always fired to one specific concept, but not to similar concepts.[8] Moreover, it did not matter whether a concept was probed in the form of a name or a picture. To give an example, in one participant a neuron was found that almost always and only fired to the name and pictures of Jennifer Aniston and the picture of at least one other *Friends* actress, but not to the names or pictures of other actresses, nor to a picture of her with Brad Pitt, to whom she was married at the time but who was not in this television series. According to Quiroga (2019) this can be understood by assuming that concepts are coded for by sets of neurons all of which code for a particular set of aspects of the concept in question. This way neurons can be shared between concepts that are in some way related, meaning that every neuron can code for a number of very specific (aspects of) concepts and thus be part of a vast number of representations. I propose we call this type of representation *quasi-atomic*, as in this coding scheme representations can be seen as atomic at one level, while at a lower level they are clearly not.[9] As we will see below, such a conceptualization suggests an explanation, not only of what Quiroga and his colleagues found, but also of the way representations are organized, how they work and even how they come about. A potential problem with these findings was that many of the neurons in question were found in a part of the brain that is not associated with semantic processing, namely the amygdala, but as will become clear below in the current thinking about neural representation this is no longer an issue.[10]

In the almost twenty years that have passed since the discovery of the Jennifer Aniston cell neuroscientists (e.g. Brodt & Gais, 2020; Guskjolen & Cembrowski, 2023; Josselyn & Tonegawa, 2020; Ryan et al., 2021) have converged on the following conceptualization of the engram. Firstly, an engram is commonly seen as an enduring change in the brain's wiring as the result of learning. Concretely, at least some of the neurons in the engram that are involved

[7] Another term for *atomic* in this sense is *localist*.

[8] In studies like these the term *concept* can refer to anything that is coded for by a ensemble of neurons.

[9] Note that this is not just another way of saying that representation has a symbolic and a subsymbolic level. Symbols are atomic in the sense that they are unstructured (cf. Chalmers, 2023), whereas in the case of quasi-atomism representations are structured and it is this property that explains how they can activate one another. It can be seen as a form of sparse coding, but as there are more forms of sparse coding I prefer the term quasi-atomic.

[10] Regrettably, in Quiroga (2020) there is no more mention of the amygdala, only of the hippocampus. Moreover, in this opinion article the author claims his alternative to the patterns separation hypothesis is only found in humans. I think this is highly unlikely.

in the encoding and consolidation phases of a memory are thought to become active again as soon as enough of these engram cells are activated as the result of an appropriate cue, thus leading to the retrieval of the memory. Some of the most convincing evidence for this comes from so called gain-of-function studies in which the neurons for a learned response are artificially activated as a result of which the animal reacts as if the cue for this response is given, even when it is not.[11] Secondly, engrams are seen as sets of neural ensembles (or engram components), each of which code for a different aspect of the memory or concept.[12] Put differently, a memory or concept should not be thought of as a thing that can be stored and retrieved, but rather as a network of networks within the greater network of the brain. As a first example, take an experiment in which an animal learns to associate a particular stimulus with pain. In this case, the engram components in the amygdala are thought to code for the emotional aspects of the memory, the engram components in the hippocampus are thought to code for the sensory information about the stimulus, and the engram components in the prefrontal cortex are thought to code for something like an abstract summary of all these things.[13] As a second example, take an experiment in which people listened to stories or watched movies while their brains were scanned. In this case, the engram components for concepts in the same class (e.g. tools, places, emotions) were found in close proximity, but either within the visual cortex or close to it, depending on the modality (verbal or visual) in which the concept was probed (Popham et al., 2021). Thirdly, engrams take time to form and are never completely fixed. In the encoding phase, cells in the hippocampus and other structures seem to form *indexes* that will keep firing, for instance during sleep. This way the neurons that were active during the experience as part of different engram components are regularly co-activated, as a result of which they can form enduring connections and the memory will be consolidated. This phenomenon, called *systems consolidation*, explains why recently formed memories are dependent on the hippocampus, whereas older memories require activation of the engram cells in the prefrontal cortex. Evidently, even after consolidation a concept or memory can undergo significant changes.[14] One way this is thought to happen is that neurons in an engram after re-activation sometimes form entirely new connections, although how this could happen is still largely unclear.

[11] Much of our understanding of engrams is based on animal experiments that involve reflexive behavior and different forms of conditioning. These experiments are relevant for the understanding of human behavior as there is ample evidence that the mechanisms that underlie these phenomena are the same across species.

[12] For readability, I am ignoring the fact that, as every aspect of a concept can be seen as a concept in its own right, every engram component can also be seen as an isolated engram.

[13] For clarity, in all of these brain structures neuroscientists distinguish a number of parts and thus a number of engrams.

[14] In fact, even during encoding and consolidation the structure of the engram is subject to change. As such, the terms *encoding, consolidation* and *retrieval* have become somewhat misleading (cf. Robins, 2020).

In conclusion, if this emerging view is correct, the mechanisms underlying mental representation are very dissimilar to other types of representation. Mental representations are not like the representations we use in communication (e.g. in the form of texts or maps) because they are not generated by one system and then somehow interpreted by that same or another system. They are also unlike the ones we see in computers, as the information they contain is not stored at one site and then processed in another (Brodt & Gais, 2020). However, as we already saw, this is not to say that mental representations should not be seen as proper representations (Ramsey, 2016; Rupert, 2018). In mental systems information flows, if we can call it that, because representations activate other representations. This is possible because the information a neuron or neural ensemble codes for is given by the elements it is connected to. Thus, the to be signaled information carried by the neural element that gives the signal is at least in some sense already present in the element that receives this signal, or it will be after consolidation, and inversely, at least some of the information carried by the element that receives the signal is already present in the element that gives the signal, or it will be after consolidation. In other words, every representation echoes some, but typically not all, of the information encoded by the representations it is connected to, as every representation carries over some, but typically not all, of the information it holds to the representations it is connected to.[15] I see this as a special case of what I propose to call *reiteration*, that is the phenomenon that (under the right circumstances) information in one element or system is caused to be present in another element or system as well.[16]

To get a better grasp of this phenomenon, I propose we first take another look at the Jennifer Aniston cell (Quiroga et al., 2005; 2008). It seemed strange that this neuron was found in the amygdala, that is a region that is associated with emotion, rather than semantic processing or memory. The reiteration hypothesis, however, suggests that it was actually part of the engram component coding for the way this participant felt about the actress and related concepts (e.g. her character and co-stars in *Friends*). This explains why it fired to the stimuli to which it was found to fire. The engram component it was a part of reiterated much of the information that was encoded in the corresponding engram component in the prefrontal cortex.

For a second example, one that I will come back to in the following sections, I propose we look at the so called shock responsive cells researchers have found in fear conditioning experiments with animals. Shock responsive cells are neurons that although they may also

[15] Note that post-synaptic neurons are as important as the pre-synaptic ones in what philosophers often call determining content. For instance, in a person who is inclined to react aggressively in the case of social threat, the neurons involved in the representation of social threat will have more projections to neurons involved in the representation of aggression than in a person who is inclined to react in a submissive way. As a result, their concepts (in the sense of ideas) of social threat will be different as well.

[16] Note that there is also reiteration between the world and the mind/brain, between words and concepts, and so on.

code for a particular context or stimulus, they are only activated to a high degree once the animal has received a shock in that context or paired with that stimulus. Kitamura and colleagues (2017) mention finding such neurons in the prefrontal cortex, and Jimenez and colleagues (2020) found them in the hippocampus.[17] Moreover, both studies suggest the shock responsive cells 1) got their input from neurons in regions that code for features in the environment (e.g. a chessboard pattern), and 2) projected to cells in the amygdala that (via other cells in the amygdala) project to the region that will cause the animal to freeze. Thus it may be as hard to say what the engrams formed by these cells coded for as it was in the case of the Jennifer Aniston cell. They can be argued to have coded for a wide range of concepts, such as shock, the chessboard pattern or whatever else it was that was used as the conditioned stimulus, pain, aversive stimulus, something that calls for freezing, threat, or any combination of these things. I submit, however, that it is exactly this indeterminacy that makes mental representation possible and as powerful as it clearly is.[18]

So far we have focused on acquired representations, but that is not to say all neural representations are the result of learning. Some concepts are innate and some behaviors are instinctive, in which case the neural representations involved are sometimes called ingrams. Examples are the (sets of) neural ensembles involved in what one might call emotional reflexes such as ducking for looming objects and the neural ensembles that code for simple features such as the pitch of a tone or the orientation of a line. Another interesting example could be the cortical shock responsive neurons mentioned above. At least as I see it, these could be seen as forming (a major part of) the innate concept of threat that is needed for fear learning. This is an exciting idea for two reasons. Firstly, it suggests that the difference between instinctive fear and the fear that is the result of an aversive experience is that the latter requires an innate semantic concept. Although in both types of fear there are threat neurons found in the hippocampus, these are not the same neurons, nor do they project to the same region (Jimenez et al., 2020). Moreover, in cases of instinctive fear the neural pathway does not involve the prefrontal cortex (Gross & Canteras, 2012). Secondly, threat is one of the so called core relational themes hypothesized by emotion theorists (e.g. Lazarus, 1991). This suggests that, since threat is innately encoded, the same may be true for at least some of the other core relational themes such as loss.

An even more important reason for looking at ingrams as well as engrams is probably that, taken together, the ingrams that an animal is born with form a network of grounded concepts, which can serve as a framework for new concepts (cf. Ryan et al., 2021). To my knowledge there are no explanations of how such a thing might work on offer yet, but this

[17] Note that the fact that the prefrontal cortex and the hippocampus code for the same state of affairs is predicted by the theory of systems consolidation mentioned above.

[18] In language this indeterminacy is smaller, but I would argue that is also why putting things into words is often both hard and helpful. As we will see in Section 4, verbalization requires and facilitates the selection of the corresponding semantic concepts.

might well be because understanding how concepts come about requires a worked out theory of the role of representations in the generation of intelligent behavior. That is why below I will suggest an answer to the question of how concepts can be grounded, but not before I have presented a theory of the role of representation in the generation of behavior, based on the insights I have summarized above.

Here it may be good to stress that, as rapidly as our knowledge of neural representations has accumulated, it is still limited. For instance, most optogenetic studies involve fear conditioning with pain as the unconditioned stimulus, and although that form of classical conditioning is already a lot more complex than when natural cues such as predator odor are used (cf. Gross & Canteras), such studies cannot tell us much about operant conditioning or semantic reasoning for example. Furthermore, although the engram studies using optogenetics provide convincing evidence for the indexing theories of systems consolidation, the way in which concepts can be bound remains unclear (cf. Yu & Lau, 2023).

One thing that we can conclude, however, is that the ‘reiterationalist’ view on representation is trouble for the radical enactivists and other philosophers who argue that in most cases the information that is relevant for the animal’s response to what is happening is already available in a non-semantic format and a doubling of this information would be inefficient (e.g. Hutto & Myin, 2012; Prinz, 2008). For instance, Prinz views emotions as perceptions of bodily changes that “represent via the body, rather than via the disembodied, freely recombinable concepts that we use in thought” (p. 709) and argues that appraising a situation in both ways would be “bad engineering” (ibid.). I submit this is no longer tenable, as this ‘bad engineering’ is exactly what was found in the optogenetic experiments done by Kitamura and colleagues (2017) I refer to above. What these experimenters found concretely was that in the days and weeks after a painful experience a fear reaction is dependent on the representation in the subcortical hippocampus, but after consolidation, this reaction of the amygdala becomes reliant on the representation in the so called shock responsive cells in the prefrontal cortex. Thus, even in relatively simple cases as this type of fear learning the formation of associations is a brain-wide process involving one or more arguably semantic concepts. That is why in this paper I claim that in neuroscience the emerging picture is that there is a mirroring between semantic and non-semantic processing and it is the interaction between these two types of representations that makes the system efficient, meaning it is not an either/or. For completeness, I am not saying that all behavior requires semantic processing, nor that the generation of inferences requires the processing of symbols by some central system (cf. Clark & Torribio, 1994; Fodor, 1983) or that learning requires conscious processing (cf. Mitchell et al., 2009).

In sum, we have good reason to assume that the different aspects of a conceptual structure all have their own representations at their own locations and that the information in any such engram components is reiterated in the components it is connected to. Now let us see how this conceptualization of mental representation can help explain behavior.

3. Complementary theories

Although it would be interesting to know to which degree the extant theories about behavior or representation are in line with the reiteration hypothesis, here we cannot examine them all. That is why in this section I will confine myself to two approaches, the predictive processing approach and the semantic roles approach, that I believe complement our analysis so far.[19]

Within the still relatively new but promising predictive processing account of behavior the brain is seen as a hierarchy of systems that represent whatever it is in the world that can cause the animal to have a particular experience (Barrett, 2017; Clark, 2013; Friston, 2010). This representational structure, referred to as the animal's generative world model (GWM), is continuously updated so that the animal's needs are anticipated and met as much and as possible. More concretely, predictions what the animal will experience or do are generated in a system such as the default mode network, typically in reaction to signals from other systems, while at the same time making the processing of these signals more efficient. Activation of the representations of what the animal experienced or did in similar situations in the past will lead to the generation of a number of representations of what it will experience or do next, and the one of these that is most fitting given the situation will lead to the corresponding experience, behavior or action. Somewhat on a side note, this is an unconscious process. Thus, what the individual does is determined outside of awareness, and as we will see below, this holds as much for humans as it does for animals.

Putting it very simply, whereas the engram literature focuses on how information may be stored, namely in the form of engrams and ingrams, the predictive processing literature offers an explanation of how the thus stored information is used. In both literatures representations do not need to be read out, nor are they somehow manipulated. Instead, information is reiterated between non-modular systems. Now we still have to come up with a way to conceptualize the predictions that are central in the predictive processing approach that researchers from various disciplines can use for their investigations. In the more worked out computational and neuroscientific versions of the predictive processing account predictions are seen as neural patterns that may be captured in mathematical models, not as conceptual structures with a truth value. However, there is no reason to assume that a more semanticized version, in which predictions are taken as propositions, is not possible as well. So let us assume it is and try and get a picture of what propositions and predictions in such a version could look like.

Proponents of the predictive processing approach typically assume that part of the animal's GWM is an amodal model of the self in its context that is associated with big parts

[19] The reason that here I am not considering teleosemantics (e.g. Millikan, 2022) is that this approach may be seen as a foreshadowing of the reiteration hypothesis without adding much to it.

of the default mode network (Koban et al., 2021; Kleckner et al., 2017). Importantly, other authors argue something similar, namely that complex forms of information processing are explained by the fact that the sensorimotor and affective information in lower-level representations is summarized in higher-level amodal representations that can easily form combinations with one another (e.g. Barrett & Satpute, 2017; Binder, 2016; Brodt et al., 2018; Brodt & Gais, 2020). Moreover, there is little controversy about the importance of the default mode network when it comes to semantic processing (e.g. Binder & Desai, 2009) and the representation of the animal's body (e.g. Lyu et al., 2023). And lastly, there is growing evidence that the what we might call index cells in the hippocampus are typically connected to index cells in cortical regions (e.g. Cowansage et al., 2014; Guskjolen & Cembrowski, 2023; Lee et al., 2023). This suggests that the non-semantic features of an experience are bound in the hippocampus, and that during the consolidation phase of the memory process this information is reorganized so it can be stored in a more schematic form in the cortical parts of the default mode system (cf. Sekeres et al., 2018).

The big question seems to be how the neural representations of propositions are structured. In the various versions of the language of thought hypothesis and most other representationalist theories semantic processing is typically seen as the manipulation of atomic elements (Chalmers, 2023). Importantly, this seems to make them incompatible with the reiteration hypothesis. In symbolic systems we typically have two sets: one containing the informational elements or symbols themselves and one containing the rules that specify how these elements can be combined into bigger structures. But if the reiteration hypothesis is true, then rules can be followed only in the sense that things happen in a systematic way as the result of representations activating other representations.

Luckily, there is at least one type of semantics in which this does not constitute a problem, namely the semantic roles approach, (e.g. Dowty, 1991; Jackendoff, 1987).[20] This is because in these theories the informational elements are thought to have roles (e.g. agent, patient, cause, manner) that can be seen as informational elements themselves. After all, we not only have words and other linguistic means to express that an action is performed by one person and undergone by another for instance, it seems plausible that we use the same categories to make sense of the situation we are in. Thus, even if these categories are only real from the cognizer's perspective, seeing they are amodal concepts that all cognizers use, we may reasonably assume that these relational concepts are represented in the same system as referential concepts are, and even in the same format. Indeed, the most important semantic roles are seen as elements of core cognition in psychology (Cain, 2021; Carey, 2011; Rissman & Majid, 2019). Much is still unclear about their workings and origin, but these are things I will come back to in the remainder of this paper. Here it may be more important to note that there may be properties and relations that cannot be represented in the way the semantic roles

[20] Other terms are *event roles, thematic roles* and *thematic relations*.

can (e.g. quantity, modality), in which case to represent the causal structure of the world a system would need more than a way to represent referential and relational concepts (cf. Yu & Lau, 2023).

To summarize, in the previous section we saw that on the reiteration account behavior is the result of representations activating other representations, and that even relatively simple behaviors such as freezing in contexts that are associated with pain involve representations in the part of the brain that is associated with semantic processing, namely the default mode network. In this section we looked at the possibility that in this network predictions are made of what the animal will do and that the neural representations of these predictions include representations of the animal itself, elements in its context, and the semantic roles. Now let us see how we can combine these insights into one account of intelligent behavior.

4. The reiteration account of intelligent behavior

In this section I will give the key tenets that together constitute a first version of what I propose to call the reiteration account of intelligent behavior, or the reiteration model for short. For clarity, I will say that behavior is intelligent if it is found only in humans or other animals that are sensitive to reward and punishment. For my tenets I have two criteria. The first is that a tenet is supported by the evidence, and the second is that together with the others they form a coherent theory, meaning that tenets cannot always be convincingly argued for in isolation. After all, a comprehensive theory is often not empirically testable in its entirety, but that does not mean it cannot be evaluated on its explanatory power and compared to other theories.

As the first key tenet of the reiteration model, the *GWM tenet,* I suggest we take the idea that is at the core of the predictive processing approach, namely that animals have a GWM they continuously update. In terms of the reiteration model I would say that in the mind/brain, on the basis of both the incoming information and the information that is stored in memory, possible *situational conclusions* are generated, that is inferences are made about what is and will be happening, not just in the world but also in the animal's body. The representations of these conclusions are all activated to some degree, but only when the activation level of one of these representations exceeds a certain threshold, the corresponding perception, thought, behavior or emotion is actualized. Alternatively, when none of these conclusions are thus confirmed, or when there is another problem, the model is adapted so a new situational conclusion can be generated that is confirmed, and/or a behavior is generated that will solve the problem. An example of the latter is looking closer at an object that could not be identified before.

As the second key tenet, the *quasi-atomism tenet*, I propose that mental representations are distributed in a quasi-atomic fashion. In a quasi-atomic coding scheme neurons typically code for one or a limited number of related concepts. In some cases this will mean that a neuron will only fire to one thing. For instance, some simple hardwired concepts such as horizontality will be coded for by a ensemble of engram cells that typically code for (different variants or aspects of) one and the same thing. This means that, at least in theory, they will always fire to this thing and only fire to some other thing if they are activated by a neuron that they are connected to and that will fire to this other thing. In such a case the engram cells are commonly referred to as simple cells (e.g. Bowers, 2009). In other cases an object, event or some other element may be represented in the form of an engram complex, that is the set of engram components in the form of neural ensembles that together represent this element. In this case, there may be no neurons that only code for this element only.

For clarity, overlap between (active or inactive) concepts can occur within in a system, as well as between two or more systems. As an example of the latter, consider the overlap between the semantic concept of a cat and the modal concept of a cat (in a person who is no way thinking of a cat). I would consider this a case of (inactive) vertical reiteration, namely between the semantic and at least one non-semantic system. To give an example of the former, within the semantic system the semantic concept of cat will have overlap with the semantic concept of dog, meaning this would be a case of horizontal reiteration. I submit that this type of overlap can be seen as reiteration because concepts are networks and thus systems by themselves. Most importantly, it is at least in part because of this type of overlap that one semantic concept can activate another semantic concept automatically, which as we will see below can lead to the automatic generation of conclusions.

Relatedly, we are still far from understanding how neural ensembles that code for semantic concepts can interact while they also remain separate. The reiteration model does suggest some ways to approach this version of what is known as the binding problem (cf. Yu & Lau, 2023), however. For instance, I can imagine there being relatively complex representations that are nevertheless coded for by index cells. Take the concepts we frequently use and have words for. Most of these will not be coded for by neurons that are linked to specific receptors or effectors, so they must get their meaning, so to speak, from their links to other concepts. I hypothesize, however, that within the semantic engram component of such a representation some neurons will always (and maybe only) fire when this particular concept applies, either because they were among the first ones to be recruited in the development of this engram or because they gradually developed the function of indexing this content. I would hypothesize further that in other cases index cells could be temporary.

As the third key tenet, the *reiteration tenet*, I propose that a representation in one mental subsystem can be reiterated in another. As an example, let us take the representations of the smell and taste of coffee. These can activate and be activated by the representation of the concept of coffee, that in turn can activate and be activated by the representations that are involved in the production or perception of the corresponding word form. All these representations refer to different aspects of the same thing and are found in different, but at least functionally connected parts of the brain.

To be sure, there are more and maybe far better ways to carve up the mind-brain, but for now I propose to distinguish the systems in have visualized in Figure 1. The idea that underlies this first proposal is that like content, the mental subsystems can be modal, amodal, or transmodal. Thus, I distinguish a (modal) primary system and a (transmodal) associative system, in which for instance visual and auditory information is encoded and combined into *Gestalts*. Additionally, I distinguish an (amodal) semantic system, in which the encoded concepts, rules and conclusions are less concrete, meaning they can be readily combined, even when they are very disparate. Finally, I propose there are three other (transmodal) systems, namely the symbolic system, that deals with verbalizations and other symbolic representations like mathematical formulas, the affective system, that deals with objects and events that relate to a goal or concern of the individual, and the indexical system, that 1) locates objects and events in space and time and 2) facilitates the binding, consolidation and retrieval of representations by keeping elements in the other systems co-active (e.g. Sekeres et al., 2018). Together these six subsystems form the mental system, that is complemented by the somatic system, that is the rest of the body.[21]

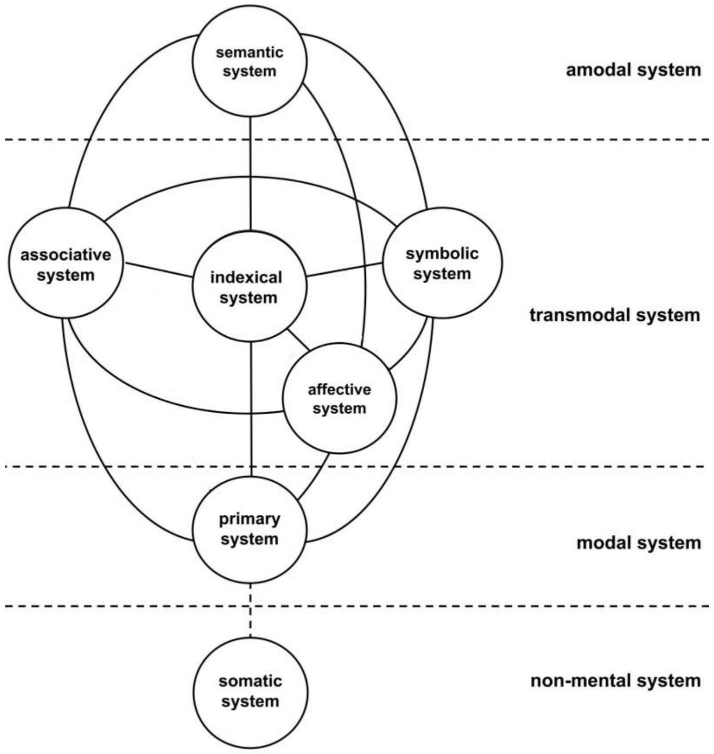

Figure 1. The subsystems in the reiteration model
The solid lines show the most important reiteration paths within the mental system, the vertical dotted line shows the reiteration path between the mental and the somatic system.

[21] Note that here I use the term *mental* in contrast to *somatic*, whereas in the most part of the paper I use it in contrast to *neural*. Thus *mental system* can be read as *mind/brain*, whereas the term *somatic system* is meant to refer to the body save the central nervous system.

To be clear, the mental systems I distinguish here are not modules in which information is encapsulated (cf. Fodor, 1983), nor are they competing systems like the automatic and controlled systems in the dual mode view (cf. Evans & Stanovich, 2013). On the contrary, their neural correlates may be hard to distinguish for two reasons. Firstly, as we have seen, the proposed systems are highly interactive, meaning that corresponding structures in different systems can keep each other activated. Secondly, they are functional systems, meaning they can have shared circuitry. For instance, the anterior cingulate cortex can be seen as part of both the semantic and the affective system. I would argue, however, that the proposed systems can be identified at least to some degree.

As already mentioned, it is reasonable to locate the semantic system in the default mode network that overlaps with the prefrontal cortex and includes a structure that is associated with the bodily self (e.g. Binder et al., 2009). The primary system I have named after the primary sensory and motor areas, but I propose more structures are involved, such as the structures that have a gating function, notably the thalamus (LeDoux, 1996) and the basal ganglia (Guo et al., 2018). The associative system I have named after the association areas in which information from different modalities for instance is assumed to be integrated further. Notably, even in the primary system information is already integrated, in the sense that for instance some related sensory and motor representations are connected, but arguably in the association areas the information will be organized in a way that it may be reiterated in the semantic system, rather than in the somatic system. The indexical system can be assumed to involve the medial temporal lobe, and especially the hippocampus and entorhinal complex (e.g. Sekeres et al., 2018), but also some cortical regions (e.g. Cowansage et al., 2014; Guskjolen & Cembrowski, 2023: Lee et al., 2023). As I see it, this system is at the core of the mental system, as it typically co-activates representations in other systems, as a result of which these representations become connected, either temporarily in working memory (Hannula et al., 2017), or permanently (Sekeres et al., 2018). The affective system may be roughly equated with the salience network (Seeley et al., 2007) and/or the mesolimbic reward network (Berridge & Robinson, 2003), but will always involve the amygdala. Lastly, the network underlying the symbolic system can be assumed to involve linguistic structures like the visual word form area (Dehaene et al., 2015) and areas that have since long been associated with language like Broca’s and Wernicke’s.

To conclude, adding the reiteration tenet to the GWM tenet and quasi-atomism tenet leads us to the conclusion that the situational conclusions that enable animals to behave in an intelligent way are represented in a multimodal quasi-atomic fashion. However, we still have to explain how the concepts an animal has can combine into a prediction. So let us now turn to the tenet that relates to that question.

As the fourth key tenet, the *semanticity tenet*, I propose that the semantic roles are represented as concepts in the semantic system. We already saw that the semantic roles can be seen as relational concepts, that is concepts that express the relations between the referential concepts. For an example of this, let us take the representational analysis in Figure 2. In this visualization of a representation both types of concepts are shown, as well as some of the propositions they either help form or characterize. It shows a part of an engram such as a participant in a conditioning study might develop. In it we see at the bottom an example of a pair of relational concepts, namely *object* and *attribute*, that are linked to *key* and *to the left* respectively, so that the proposition is formed *the key is to the left*. We also see examples of relational concepts that identify aspects of propositions. For instance, *(the left) key* is the part of the proposition *I press the left key* that fulfills the role of object. Lastly, we see examples of relational concepts that identify the roles propositions can have when they are combined into a complex proposition such as a rule. For instance, *I press the left key* is an action related to a condition in the rule *If I see XXX, I press the left key*, and an action related to an effect in the rule *If I press the left key, I will hear a low tone*. Thus, all referential concepts and propositions in this engram are related to one another, but not in an arbitrary way. As a result of their links to the relational concepts they form qualified relations, and not simple associations. Moreover, the resulting network reflects the structure of the world as the animal knows it.

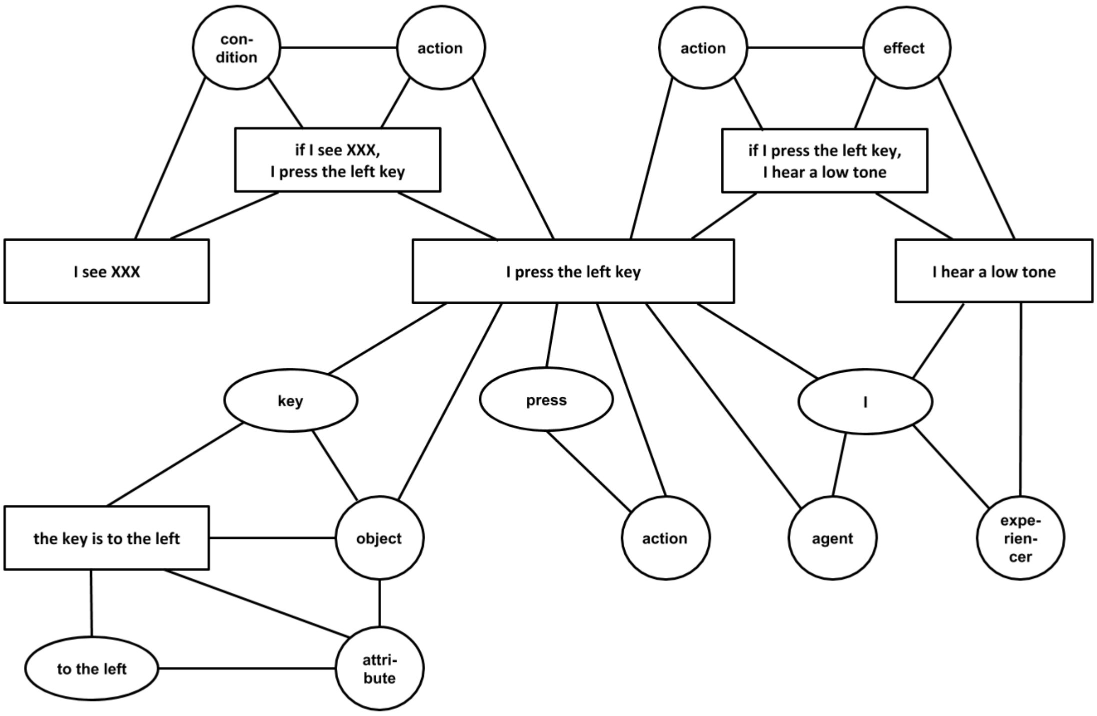

Figure 2. Representational analysis of possible engrams in a participant of a conditioning experiment
Referential concepts are represented as ovals, relational concepts as circles and the (complex) propositions as rectangles.

Here it may be good to note a number of things. Firstly, relational concepts typically come in pairs. For example, in Figure 2 we see an object-attribute pair and a condition-action pair, with the relation between the two complementing elements visualized as a line. Examples of relational pairs that are not made explicit in Figure 2 are the event-action and the event-agent pairs. Possessing concepts such as *object, agent, event, cause* and *effect* means among other things knowing that objects and agents have attributes, and that events typically have causes and effects. Admittedly, the network shown in Figure 2 is just one of the possible visualizations of the information a participant in a conditioning experiment can be expected to have. Additionally, as humans have linguistic tools to specify the relations between referents, we are likely to have more complex representations than other animals.

Secondly, representational analyses such as the one in Figure 2 can be misleading. For one thing, the use of the *I*-concept may suggest the animal has a sense of self and/or is aware of everything it perceives and does. In the animal's GWM, however, the agent here denoted as *I* is rather the entity about and from which the system gets information it does not get about anyone else. As a second example, the distinction that is made between propositions and concepts is not a principled one. Admittedly, we normally see concepts as elements that could refer to something that exists or not, meaning they are typically expressed in the form of (a word group that centers around) a noun, whereas we see propositions as elements with a truth value, meaning they could be expressed in the form of a sentence. Nevertheless, in natural systems many relational structures may generate propositions as well as concepts, an example of which are the elements *the left key* and *the key is to the left* in Figure 2. After all, as long as a representation in the semantic system is not (about to be) reiterated in a (trans)modal system, it has a structure that may be cast in a number of verbal forms and images.[22]

Thirdly, it may seem a stretch to assume animals actually encode *if-then* rules. However, let us assume that some of the engram cells for the representation of the condition part of a rule code for the fact that this representation is indeed a condition, and that some of the engram cells for the representation of the action part code for the fact that this representation is indeed a action. Then, if we assume further that these two sets of engrams cells can connect to one another, which we can as connecting to actions is a property of conditions and vice versa, we already have an engram for the corresponding *if-then* rule.

Lastly, in newborn animals, including human babies, the semantic system will not yet be fully functional, that is few or none of the elements that represent the core concepts it contains will perform all of the functions that they will later in life. However, in the next section we will see that arguably at least the concepts that are of the greatest importance to us, such as some of the relational concepts and the *I*-concept, will start to develop around the time they are born. This is because these concepts can be derived from the representations that are involved in the reflexes animals that are sensitive to reward and punishment are born with, although they will need time and input to develop fully. Of course, in the case of non-human animals even after they have developed fully the individual will have no way to express them,

[22] Especially psychologists tend to equate propositions with sentences, thus ignoring the important distinction between amodal concepts, transmodal words and modal word forms.

meaning it cannot use them to reflect on their own behavior or solve complex problems. However, for the remainder non-human animals will use them just like we do, namely to use past experience to act and react in an adaptive way.

In conclusion, I propose that the more complex (situational) conclusions in the GWM are summarized in a propositional format. However, I do not conceptualize these propositions as (the result of) computations over symbols, but rather as semantic structures, that is structures that reflect the elements and their relations to one another as we perceive them to be. The addition of this tenet to the others gives us an idea of how situational conclusions can be realized and structured. Now let us turn to the tenet that specifies how these structures can be generated.

As the fifth key tenet, the *automatic reasoning tenet*, I propose that inferences are made automatically through the spreading of activation in the semantic system. As the neurons in a brain are all connected to one another, activation of one will always lead activation of others. Importantly, this never leads to activation of the entire network because activation is a phenomenon that takes time and energy (cf. Anderson, 1983; Collins & Loftus, 1975; Mirman & Magnuson, 2009).[23] Thus, given the fact that the neurons that activate each other in the semantic system are connected to one another in a way that reflects the structure of the world as the animal knows it, the routes that are thus followed will reflect logical inferences about the situation as the animal perceives it. As spreading of activation is a very fast and automatic process, there may be many routes followed in parallel, only a few of which will enter awareness in the form of a verbalization or image. When and how an inference will enter awareness is something I will come back to below. For now the important thing is to see is that the phenomenon of activation spreading can help explain the unconscious reasoning that is assumed by philosophers such as Carruthers (2017) and Jackendoff (2012), and for which at least some psychologists have found evidence (e.g. Dijksterhuis & Strick, 2016). It may seem as if reasoning requires awareness, because we associate it with formulating reasons. However, to come up with an expression of these reasons, in the form of an utterance or image, the necessary inferences need to be made first.

Obviously, not all conceptual structures that are unconsciously activated and formed are equally relevant to the animal in the given situation.[24] However, an activated or newly formed structure that is not relevant will have very few connections to structures that are relevant and thus highly activated, for instance because they are reiterations of concepts in the modal system. Thus, the irrelevant structure will not stay active for long, meaning only the relevant alternatives (cf. Dretske, 1970) remain active, that is until an inference is made that ‘wins’ the competition. For instance, if you have a white cat and as you come home you see a white fluffy sweater on your couch, you are more likely to mistake this for your cat than for a rabbit.

[23] Thus, we need not fear mental holism (e.g. Fodor & Lepore, 2002).

[24] I use the term conceptual structure to denote structures that are concepts or combinations of concepts. So they are representations, but from a specific perspective.

This is because the concept of your cat is closely related to the highly activated concept of your home.

I submit that in some cases reasoning amounts to no more than activating an *if-then* rule, whereas in other cases the system will have to construct a number of propositions and *if-then* rules on the spot by combining existing propositions and *if-then* rules.[25] Such ad hoc elements may become permanent, but that need not be the case. At any rate, their construction will involve the indexical system. I submit, that in this system, neurons are recruited that keep the neurons in the other systems that together will form a new structure co-active so that they can indeed establish the necessary connections, at least for the time being. As we already saw, this mechanism has been proposed as an explanation of the encoding of memories (e.g. Kitamura et al., 2017; Sekeres et al., 2018), but as encoding a memory implies making inferences, it is assumed to work in reasoning as well (e.g. Hannula et al., 2017).

In conclusion, the automatic reasoning tenet gives us an explanation of how we may arrive at situational conclusions. Because these are generated automatically, they may or may not lead to a conscious experience. And because they are in competition with one another, they may or may not lead to an (overt) emotion or behavior. So, let us now consider what it takes for a situational conclusion to have a reportable or otherwise observable effect.

As the last key tenet of the reiteration model, the *intelligence tenet*, I propose that intelligent behavior requires semantic processing. More specifically, I propose we make a distinction between basic cognition, semantic cognition and symbolic cognition. Semantic processing is sometimes confused with symbolic or linguistic processing, I suspect, partly because semantics is a branch of linguistics, and partly because of the logical turn in philosophy. However, as humans seem to be no less sensitive to reward and punishment than other animals, I would say the only difference between humans and other conditionable animals is our use of symbols. Furthermore, seeing that even such a simple type of learning as fear conditioning in animals involves the semantic system, I would say that all behavior that is the result of learning that goes beyond the most simple forms of classical conditioning is intelligent. This leaves us one type of cognition and thus behavior that is basic or automatic, rather than intelligent.

So what kind of doings are automatic rather than intelligent, meaning they do not involve semantic processing? Clearly, (emotional) reflexes such as ducking for looming objects are an example. Another example are the simple forms of classical conditioning such as seen in Pavlov’s dogs, that salivated when they were presented with the bell they had learned to associate with food.[26] Another example are the things you do when you trip. The

[25] I submit that if information that is needed to do this is not immediately available the system may even develop the equivalent of a (topical) question, but that is for another time.

[26] Note that, although this form of Pavlovian (or classical or respondent) conditioning is not routinely distinguished from more complex forms of Pavlovian conditioning such as used in the engram studies we saw in Section 2, their underlying mechanisms are very different (cf. Gross & Canteras, 2012).

representations involved in these behaviors will all be in what I have called the primary, associative and affective systems. As a last example, take reaching for your cup of coffee. After you have decided to take a sip, the remaining part of this action will necessarily be automatic, if only because semantic processing is too slow for this type of thing. In this case I submit the intelligent part of your behavior is the part where you decide to take a sip. This is something that arguably involves the semantic system, as the situational conclusion *I will take sip of my coffee* will have to become activated to the degree that the neurons that realize this representation in the semantic system will activate the corresponding neurons in the primary and somatic systems. Additionally, the corresponding neurons in the symbolic system may be activated enough to lead to a conscious thought, but this is not necessary for the action of picking up your cup.

Coming back to intelligent behavior, I submit that whether or not a situational conclusion leads to a distinct memory, experience, act or other behavior, will depend on its relevance. As I already indicated, in the semantic system a great number of inferences can be made in parallel and outside of awareness. However, when a situational conclusion becomes highly activated, then the corresponding transmodal representations (corresponding to words and emotions) and modal representations (corresponding to different perceptions and actions) will become also highly activated, thus leading to a thought or utterance, a conscious perception, a conscious image, emotion or action. In the case of a thought or utterance, this will involve reiteration between the semantic an symbolic systems, in the case of perception, mental imagery and behavior this will involve reiteration between the semantic and the primary systems, and in the case of emotion this will involve reiteration between the semantic and the affective systems.

Notably, there are at least three ways in which a situational conclusion can be drawn. The first is that the elements involved in this reasoning have been activated outside of awareness. The second is that the animal has already been aware of these elements for some time. And the third is that a representation is activated that is always reiterated in the affective system. In fact, this is how emotions work. Even if an animal's attention is directed at something else, as soon as something happens that really requires its attention, the representation of that event activates the representation of the relevant core relational theme, that is the engram component that is part of the neural pathway that leads to a bodily reaction that will, at minimum, direct the animal's attention to the threat or opportunity that evoked the emotion. On a side note, in all three cases the affective system is likely to be involved to some degree. After all, according to the GWM tenet doing something requires seeing it as something that is good to do. Moreover, doing something requires arousal, which is taken care of in the affective system.[27]

[27] The idea that cognitive and emotional systems work in tandem can also be found, in one way or the other, in for instance Barrett (2017), Kriegel (2014) and Morton (2009).

That in humans the activation of thoughts in the semantic system leads to activation of the related phrases in the symbolic system, and vice versa, may not be particularly surprising. It may be good to note, however, that when a thought activates a phrase, there is also reiteration between the symbolic and primary system, after which through the (simulated) perception of the formed phrase the original thought is attended to and thus activated extra.[28] Moreover, the reiteration account explains why thoughts that are verbalized are about the potentially important or salient things. Of all the situational conclusions we draw, the ones that will enter awareness are the ones that are more likely to require at least some mental action, for instance because there is no automatic explanation. To see how this works, suppose you are in a supermarket, thinking about what you will have for dinner. In this situation you will not pay much attention to the other people there, as long as they are all behaving like you expect them to. However, if you see someone sitting on the floor while reading a book, you will start to wonder. In this case the representation of the situational conclusion that there is someone sitting on the floor and reading a book in the middle of the supermarket will remain active, while your mind/brain will keep generating inferences until one is found that explains the situation (or you give up). As this will take some time, the concepts involved will activate the words that they are linked to and a question and answer may be generated that you will be able to hear in your mind's ear.

Now let us consider how it is possible that unexpected events will catch our attention in the first place. After all, when we notice something is off, this means we have already inferred that something was off. However, as we could not know beforehand that that would be the case, we had no special reason to process the information that we needed to infer this. Thus, the fact we did process this information implies we routinely process a lot of information that never makes it to a thought or experience. Additionally, given the fact that representations that are linked to one another will automatically activate one another, it follows that if we have concepts relating to the elements in a situation, then it is likely that they will be activated. Thus, if an element deserves our attention, it will catch our attention. Alternatively, if an element does not deserve our attention, a modal representation may be reiterated in the semantic system, and thus speed up the interpretation of the situation, without entering awareness. At any rate, as concepts are activated, they will almost inevitably influence the perception process. For instance, as we have concepts such as *agent* and *object*, we see people and things, and not just colors and shapes.[29] And of course, sometimes important things escape our attention or we get distracted by trivialities. The important thing is that, overall,

[28] It may be that in people with aphantasia, who often do not experience an inner monologue, the activation in the symbolic system adds to the activation in the semantic system, even though it is not enough to activate the corresponding words in the primary system (cf. Lennon, 2023).

[29] Of note, that we have these concepts is in part thanks to our experience with agents and objects, but as we will see, this does not mean they are the result of this experience only.

the automatic activation of representations in one system by the corresponding ones in the other may well be what drives intelligence.

To summarize, with the reiteration model we have gotten a better picture of how situational conclusions are evoked by stimuli and lead to behavior. Now let us examine firstly, how the concepts that make out these conclusions may be formed, and secondly, how the concepts in the representational network that the newer of these concepts build on could have come into existence.

5. The reiteration account of concept formation

So far, we have looked at mental representation in adults, that is in individuals who have had a wealth of experience on which they can draw in interpreting the situations they find themselves in. They know that the world is crowded with objects and agents, that there are events located in space and time, that they themselves have desires and can perform actions, etcetera. Even if they would want to, they could not interpret a situation without using this knowledge. However, we cannot simply assume that to a newborn the world appears in a similar way, as it is at least as likely that in its first days and weeks the individual experiences the world as an unordered mix of sounds, moving shapes, feelings, smells etcetera. That is why in this section I will propose six principles that could play a role in the construction of complex concepts, that is concepts that are not directly linked to sensory or motor neurons, whether these concepts are modal, transmodal or amodal. First, I will discuss the four principles that follow from the reiteration model that I outlined in the previous section and may explain how new concepts come into being, and next, I will discuss the two principles that may explain how core concepts such as *agent* and *cause* come into being. On a side note, here I give these principles because they can explain how we form new concepts, and thus our ability to categorize, but they are equally applicable to the formation of propositions, including situational conclusions and rules, that is our ability to reason. As noted in the previous section, in many ways the distinction between concepts and propositions is not relevant in this type of analysis.

My first assumption regarding the development of new concepts is that seeing representations as reiterations means accepting that at least some complex representations have no clear boundaries and are subject to change. As a first example, consider how the way you feel about a particular person may vary, depending on the last interaction you had with them. As a second example, consider how the change or development of a concept will always depend on changes in the network around it, meaning that the structure of the concepts that are related to the changing concept will change with it. After all, the idea behind reiteration is that connecting concepts are realized by neural ensembles that in part code for the same things. Notably, in this respect mental concepts are not very different from non-mental concepts, seeing that words do not have completely fixed meanings either. Let us call this *fuzziness principle*.

My second assumption is that, because reasoning and learning is automatic, new concepts will develop automatically as well. According to this *logicality principle* the connections in a representational network reflect the rules we also use in reasoning. This is possible as a result of the capacity of concepts to bind and be bound by other concepts. For instance, if there is a commonality or a clear difference between two or more frequently activated instances of a object category, and there is not already a concept that binds the related concepts by capturing this difference or commonality, such a concept will develop. An example of this would be a human infant discovering that a number of balls that all feel and look different can all be rolled from one place to another, thus either acquiring the concept of ball or further developing it. Other rules that the system seems to follow are transitivity and of course implication.

On a side note, the logicality principle helps to explain, not only the development of new concepts, but also automatic or habitual behavior as well as more deliberate action. Both types of behavior can be the result of the activation of an *if-then* rule in the semantic system, but the more automatic types could also be the result of the activation of a link between two representations, one of a perception and one of an action, without involvement of the semantic system. Examples of such automatic behavior are the emotional reflexes.

Furthermore, the fact that new concepts develop automatically, does not mean this cannot be a gradual process. We say that we grasp a new concept, suggesting that it is there for the taking and that after having grasped it, it will stay available to us. I maintain, however, that that can only happen after previously acquired concepts have connected to one another in the way necessary to form the new concept, which will only happen if these concepts have been co-active repeatedly and to a large enough degree. This co-activation may have led to conscious thought on more than one occasion, but it will have started as all new thoughts outside of awareness and much of the consolidation of the concept in question will have happened outside of awareness as well. Moreover, even if such a new connection is made very fast, it will always take time before it is consolidated (cf. Sekeres et al., 2018).

Lastly, as automatic processes are typically quick rather than precise, concepts may develop that serve no function. This does not mean, of course, that the system that produces them is not efficient. Moreover, if a developing concept is superfluous, or wrong even, there is less chance that it will be consolidated.

Relatedly, we already saw that there is converging evidence that concepts are multimodal (e.g. Barrett & Satpute, 2017; Binder, 2016; Brodt et al., 2018; Brodt & Gais, 2020). In terms of the reiteration model, (aspects of) concepts in one system are reiterated in at least one other system. For instance, putting it simply, the form of a word is reiterated between the symbolic and the association system, whereas the meaning of a word is reiterated between the symbolic and the semantic system. Let us call this the *cross-modalarity principle*.

What a neuron codes for is dependent on the neurons that activate it and/or the neurons that it activates itself. Furthermore, for a neuron to serve any function, it must have a direct or indirect relation with a neuron in peripheral nervous system, that is a neuron that is linked to a receptor cell, for instance in the eye, or an effector cell, for instance in a gland. Thus every neuron can be said to get information from the body and/or relate information to the body. After all, there can be no neural subsystems or engrams without links to the rest of the network. Let us call this the *embodiment principle*.

This brings us to the remaining two principles that concern the concepts that are already encoded in the neural system at birth. We have already seen a number of hardwired behaviors, such as ducking for looming objects or salivating on seeing or smelling food. I submit these behaviors all involve hardwired modal concepts (e.g. of looming object and ducking) and propose to see these concepts as examples of the *preparedness principle*. The term preparedness was originally meant to apply only to patterns that are more readily learned than others (Seligman, 1970). For instance, when you use snake-like stimuli to instill fear in a lab raised monkey you will need less sessions than when you use flowers. I submit, however, the principle can be extended much further.

As an example, let us take the modal concept of pleasant that is applicable to experiencing warmth or closeness to others for instance. This concept can steer neonates in the direction of adaptive behaviors, such as staying close to their mothers, through its rewarding effect. Moreover, it has a clear neural correlate, meaning we may assume it is innate. As there are a number of rather simple physical stimuli that are pleasant, the encoding of experiencing pleasure is likely to be connected to the encoding of these stimuli. If the logicality principle is applicable here, then with time and enough experiences of pleasure the individual may also acquire the semantic and thus amodal concept of pleasant. However, this concept may also be hardwired through its connections to the relevant modal concepts, meaning that it would be prepared in the semantic system before the individual had actually had a response that would work as a reward. In fact, this is what I would expect, given the analogy with the threat cells we came across earlier (Kitamura et al., 2017). To summarize, if a representation is hardwired but needs experience in order to become fully functional, it can be described as prepared. If no such experience is required, as is the case in emotional reflexes, than it is better described as innate. Granted, not all cases are this clear-cut, but that may be seen as an example of the fuzziness principle.

This brings us to what I propose to call the *evolutionarity principle.* This last principle specifies which concepts may be innate or prepared in the following way. If a concept or structure of concepts will help an animal act and react in an adaptive way and it consists of concepts that can be hardwired, then we may assume it will be hardwired as well. In other words, not only the concepts in ingrams are innate or prepared, so is any concept that can be linked to an ingram as per the other principles of concept formation, as long as it can be expected to enhance the animal’s chances of survival.

Together these principles offer the following explanation of how concepts come into being. The concepts that have direct links to the sensory and motor neurons are hardwired as a result of these links. The ingrams involved in each emotional or other reflex are combinations of concepts that consist of these simple cells and all the reiterations of these simple concepts that together form a pathway from stimulus to response. Thanks to these ingrams, the animal can develop a model of the world, even before it is born. As per the logicality and automaticity principles, with the formation of the ingrams other important concepts will start to form. Examples are *cause, action, I, object, conspecific, threat, pleasant*. After all, these are useful concepts to have and much of the information that goes into them is already available in the developing GWM. For instance, all the reactions to positive stimuli as food and sex are hardwired, meaning that the concepts of pleasant and good can be hardwired as well. They may not become fully functional until the animal is born and has interacted with others and had other important experiences, but soon enough they will be and then they will help the animal to arrive at the situational conclusions that in most cases will ensure it does what is in its best interest.[30] Moreover, as some of the thus formed concepts are normative rather than descriptive, they may even be instrumental in the individual's moral development.

In conclusion, our reconceptualization of mental representation does not just give us a new understanding of intelligent behavior, it also explains how representations come into being, including some of the most central and abstract ones. As such, it may give us a new perspective on a range of philosophical questions.

6. Discussion

In this paper I proposed a way to 'semanticize' the brain, that is to explain behavior in a way that focuses more than other approaches on the content of representations. Building on the recent neuroscientific literature about engrams, I worked out the concept of reiteration, that is the phenomenon that part of the information that is carried by a mental representation carries over to the representations it is connected to. Next, I combined this reconceptualization with insights from the predictive processing and semantic roles approaches, and outlined the resulting accounts of intelligent behavior and concept formation. Thus, we arrived at an account of behavior in which representationalist and embodied views on cognition are combined.

As the reiteration model is only a first version both accounts are mechanism sketches rather than full mechanistic explanations (cf. Piccinini & Craver, 2011). For instance, it remains unclear how indexing works, if relational and referential concepts can be combined as suggested, how neurotransmitters can be fitted in this proposal, etcetera. Nevertheless, we already came across some examples of what the reiteration model may explain. The model

[30] This may be seen as evidence for the solution Piccinini (2022) suggests for what he considers the second problem of content, that is the coordination between vehicles and their content.

suggests, for instance, how reasoning can be unconscious, and how concepts come into being. Furthermore, it offers a rationale for distinguishing between emotional reflexes (or basic emotions) on the one hand, and standard emotions on the other, thus shedding a new light on a debate that has gone on for decades (e.g. Scarantino & Griffiths, 2011). Lastly, it suggests a distinction between basic cognition, semantic cognition and symbolic cognition, thus shedding a new light on the so called representation wars (e.g. Clark, 2015) or the hard problem of content (Hutto & Myin, 2012). So, although the reiteration model raises new questions, it suggests new answers as well, meaning it does what a new theory arguably should do, and as importantly, it does so in line with the available empirical evidence.

Now one might still wonder if developing a model of mental processing should not be left to neuroscientists and other empirical researchers. I would argue, however, that scientists are trained and expected to rigorously test ideas in their own field, meaning they are often hesitant to hypothesize about a subject using ideas from other fields or other corners of their own field. This may be an even greater problem in the case of representation which arguably has a neural and behavioral as well as a semantic and logical side to it. For these reasons, I am with those who call for an empirically informed philosophy (e.g. Churchland, 1986; Dutilh-Novaes, 2023). For a better understanding of the relation between representation and behavior ideas from very different fields will have to be integrated, and this requires the skills and attitudes of the philosopher as much as the knowledge of the empirical scientist.

Utrecht, 27 December 2023

References

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of verbal learning and verbal behavior*, *22*(3), 261-295.

Barrett, L. F. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, *12*(1), 1-23.

Barrett, L. F., & Satpute, A. B. (2013). Large-scale brain networks in affective and social neuroscience: towards an integrative functional architecture of the brain. *Current opinion in neurobiology*, *23*(3), 361-372.

Binder, J. R. (2016). In defense of abstract conceptual representations. *Psychonomic bulletin & review*, *23*(4), 1096-1108.

Boyd, R. (1999). Homeostasis, species, and higher taxa. *Species: New interdisciplinary essays*, *141*, 185.

Bowers, J. S. (2009). On the biological plausibility of grandmother cells: implications for neural network theories in psychology and neuroscience. *Psychological review*, *116*(1), 220.

Bowers, J. (2017). Grandmother cells and localist representations: a review of current thinking. *Language, Cognition and Neuroscience*, *32*(3), 257-273.

Brodt, S., & Gais, S. (2021). Memory engrams in the neocortex. *The Neuroscientist*, *27*(4), 427-444.

Brodt, S., Gais, S., Beck, J., Erb, M., Scheffler, K., & Schönauer, M. (2018). Fast track to the neocortex: A memory engram in the posterior parietal cortex. *Science*, *362*(6418), 1045-1048.

Cain, M. J. (2021). *Innateness and cognition*. Routledge.

Carey, S. (2011). Précis of the origin of concepts. *Behavioral and Brain Sciences*, *34*(3), 113-124.

Carruthers, P. (2017). The illusion of conscious thought. *Journal of Consciousness Studies*, *24*(9-10), 228-252.

Chalmers, D. (2023). The computational and the representational language-of-thought hypotheses. *Behavioral and Brain Sciences*, *46.*

Churchland, P. S. (1986). *Neurophilosophy: Toward a unified science of the mind-brain*. MIT press.

Churchland, P. S., & Sejnowski, T. J. (1990). Neural representation and neural computation. *Philosophical Perspectives*, *4*, 343-382.

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

Clark, A. (2015). Predicting peace: The end of the representation wars. In *Open mind*. Open MIND. Frankfurt am Main: MIND Group.

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

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological review*, *82*(6), 407.

Cowansage, K. K., Shuman, T., Dillingham, B. C., Chang, A., Golshani, P., & Mayford, M. (2014). Direct reactivation of a coherent neocortical memory of context. *Neuron*, *84*(2), 432-441.

Dehaene, S., Cohen, L., Morais, J., & Kolinsky, R. (2015). Illiterate to literate: behavioural and cerebral changes induced by reading acquisition. *Nature Reviews Neuroscience*, *16*(4), 234-244.

Dennett, D. C. (1989). *The intentional stance*. MIT press. (Original work published 1987)

Dijksterhuis, A., & Strick, M. (2016). A case for thinking without consciousness. *Perspectives on Psychological Science*, *11*(1), 117-132.

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, *67*(3), 547-619.

Dretske, F. (1988). *Explaining behavior: Reasons in a world of causes*. MIT press.

Dretske, F. (1995). *Naturalizing the mind*. MIT Press.

Dutilh-Novaes, C. (2023, June 6). *A plea for Synthetic Philosophy.*. Daily Nous - news for & about the philosophy profession. https://dailynous.com/2023/05/30/a-plea-for-synthetic-philosophy-guest-post/

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, *8*(3), 223-241.

Fodor, J. A. (1983). *The modularity of mind*. MIT press.

Fodor, J. A., & Lepore, E. (2002). *The compositionality papers*. Oxford University Press.

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

Gross, C. T., & Canteras, N. S. (2012). The many paths to fear. *Nature Reviews Neuroscience*, *13*(9), 651-658.

Guo, Y., Schmitz, T. W., Mur, M., Ferreira, C. S., & Anderson, M. C. (2018). A supramodal role of the basal ganglia in memory and motor inhibition: Meta-analytic evidence. *Neuropsychologia*, *108*, 117-134.

Guskjolen, A., & Cembrowski, M. S. (2023). Engram neurons: Encoding, consolidation, retrieval, and forgetting of memory. *Molecular Psychiatry*, 1-13.

Hannula, D. E., Ryan, J. D., & Warren, D. E. (2017). Beyond long-term declarative memory: evaluating hippocampal contributions to unconscious memory expression, perception, and short-term retention. In *The hippocampus from cells to systems* (pp. 281-336). Springer, Cham.

Hutto, D. D., & Myin, E. (2012). *Radicalizing enactivism: Basic minds without content*. MIT press.

Jackendoff, R. (1987). The status of thematic relations in linguistic theory. *Linguistic inquiry*, *18*(3), 369-411.

Jackendoff, R. (2012). *A user's guide to thought and meaning*. Oxford University Press.

Josselyn, S. A., & Tonegawa, S. (2020). Memory engrams: Recalling the past and imagining the future. *Science*, *367*(6473.

Kleckner, I. R., Zhang, J., Touroutoglou, A., Chanes, L., Xia, C., Simmons, W. K., ... & Feldman Barrett, L. (2017). Evidence for a large-scale brain system supporting allostasis and interoception in humans. *Nature human behaviour*, *1*(5), 0069.

Koban, L., Gianaros, P. J., Kober, H., & Wager, T. D. (2021). The self in context: brain systems linking mental and physical health. *Nature Reviews Neuroscience*, *22*(5), 309-322.

Kriegel, U. (2014). Towards a new feeling theory of emotion. *European Journal of Philosophy*, *22*(3), 420-442.

Lazarus, R. S. (1991). Progress on a cognitive-motivational-relational theory of emotion. *American psychologist*, *46*(8), 819.

Lee, J. H., Kim, W. B., Park, E. H., & Cho, J. H. (2023). Neocortical synaptic engrams for remote contextual memories. *Nature Neuroscience*, *26*(2), 259-273.

Lennon, P. (2023). Aphantasia and Conscious Thought. *Oxford Studies in Philosophy of Mind Volume 3*, 131.

Lyu, D., Stieger, J. R., Xin, C., Ma, E., Lusk, Z., Aparicio, M. K., ... & Parvizi, J. (2023). Causal evidence for the processing of bodily self in the anterior precuneus. *Neuron*.

Mahon, B. Z., & Hickok, G. (2016). Arguments about the nature of concepts: Symbols, embodiment, and beyond. *Psychonomic bulletin & review*, *23*(4), 941-958.

Mirman, D., & Magnuson, J. S. (2009). Dynamics of activation of semantically similar concepts during spoken word recognition. *Memory & cognition*, *37*(7), 1026-1039.

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences*, *32*(2), 183-198.

Morton, A. (2009). Epistemic Emotions. In P. Goldie (ed.), *The Oxford Handbook of Philosophy of Emotion* (385-399). Oxford University Press.

Najenson, J. (2021). What have we learned about the engram?. *Synthese*, *199*(3-4), 9581-9601.

Ortega-de San Luis, C., & Ryan, T. J. (2022). Understanding the physical basis of memory: molecular mechanisms of the engram. *Journal of Biological Chemistry*, *298*(5).

Piccinini, G. (2022). Situated neural representations: Solving the problems of content. *Frontiers in Neurorobotics*, *16*, 846979.

Piccinini, G., & Bahar, S. (2013). Neural computation and the computational theory of cognition. *Cognitive science*, *37*(3), 453-488.

Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, *183*, 283-311.

Popham, S. F., Huth, A. G., Bilenko, N. Y., Deniz, F., Gao, J. S., Nunez-Elizalde, A. O., & Gallant, J. L. (2021). Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature neuroscience*, *24*(11), 1628-1636.

Prinz, J. (2008). Précis of" Gut Reactions". *Philosophy and Phenomenological Research*, *76*(3), 707–711.

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

Quiroga, R. Q. (2019). Plugging in to Human Memory: Advantages, Challenges, and Insights from Human Single-Neuron Recordings. *Cell*, *179*(5), 1015-1032.

Quiroga, R. Q. (2020). No pattern separation in the human hippocampus. *Trends in Cognitive Sciences*, *24*(12), 994-1007.

Quiroga, R. Q., Kreiman, G., Koch, C., & Fried, I. (2008). Sparse but not 'grandmother-cell'coding in the medial temporal lobe. *Trends in cognitive sciences*, *12*(3), 87-91.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435*(7045), 1102-1107.

Ramsey, W. (2016). Untangling two questions about mental representation. *New Ideas in Psychology*, *40*, 3-12.

Rissman, L., & Majid, A. (2019). Thematic roles: Core knowledge or linguistic construct?. *Psychonomic bulletin & review*, *26*(6), 1850-1869.

Robins, S. K. (2020). Stable engrams and neural dynamics. *Philosophy of Science*, *87*(5), 1130-1139.

Roy, A. (2017). The theory of localist representation and of a purely abstract cognitive system: the evidence from cortical columns, category cells, and multisensory neurons. *Frontiers in psychology*, *8*, 186.

Roy, D. S., Park, Y. G., Kim, M. E., Zhang, Y., Ogawa, S. K., DiNapoli, N., ... & Tonegawa, S. (2022). Brain-wide mapping reveals that engrams for a single memory are distributed across multiple brain regions. *Nature communications*, *13*(1), 1-16.

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

Ryan, T. J., Ortega-de San Luis, C., Pezzoli, M., & Sen, S. (2021). Engram cell connectivity: an evolving substrate for information storage. *Current Opinion in Neurobiology*, *67*, 215-225.

Scarantino, A., & Griffiths, P. (2011). Don't give up on basic emotions. *Emotion Review*, *3*(4), 444-454.

Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H., ... & Greicius, M. D. (2007). Dissociable intrinsic connectivity networks for salience processing and executive control. *Journal of Neuroscience*, *27*(9), 2349-2356.

Sekeres, M. J., Winocur, G., & Moscovitch, M. (2018). The hippocampus and related neocortical structures in memory transformation. *Neuroscience letters*, *680*, 39-53.

Seligman, M. E. (1970). On the generality of the laws of learning. *Psychological review*, *77*(5), 406-418.

Tonegawa, S., Liu, X., Ramirez, S., & Redondo, R. (2015). Memory engram cells have come of age. *Neuron*, *87*(5), 918-931.

Von Eckardt, B. (2012). The representational theory of mind. In K. Frankish & W. Ramsey (Eds.) *The Cambridge handbook of cognitive science*, *1* (29-50). Cambridge University Press.

Yu, X., & Lau, E. (2023). The binding problem 2.0: beyond perceptual features. *Cognitive Science*, *47*(2), e13244.