doi.org/10.53975/updh-78lg

Seeds of Science

# What does it mean to represent? Mental representations as falsifiable memory patterns

Eloy Parra-Barrero[1], Yulia Sandamirskaya[2]

**Representation is a key notion in neuroscience and artificial intelligence (AI). However, a longstanding philosophical debate highlights that specifying what counts as representation is trickier than it seems. With this brief opinion paper we would like to bring the philosophical problem of representation into attention and provide an implementable solution. We note that causal and teleological approaches often assumed by neuroscientists and engineers fail to provide a satisfactory account of representation. We sketch an alternative according to which representations correspond to inferred latent structures in the world, identified on the basis of conditional patterns of activation. These structures are assumed to have certain properties objectively, which allows for planning, prediction, and detection of unexpected events. We illustrate our proposal with the simulation of a simple neural network model. We believe this stronger notion of representation could inform future research in neuroscience and AI.**

Cognition is often described as computation over representations to yield behaviour (Gallistel, 1990; Barack and Krakauer, 2021). Although some might disagree with this explicit formulation, the notion of representation certainly plays a central role in the study of natural and artificial cognition. Neuroscientists invoke it when discussing almost all aspects of cognitive brain function, from sensory processing to learning and memory or decision making (e.g., Kriegeskorte et al. (2008), Quiroga (2012), Knutson et al. (2005)). Perhaps neuroscientific research in spatial cognition highlights this point most clearly. Over the past few decades, researchers have uncovered a zoo of neuronal cell types said to represent all sorts of spatial properties: position, heading direction, speed, the presence of borders, objects, goals, etc. (Hartley et al., 2014; Bicanski and Burgess, 2020; Parra-Barrero et al., 2021). In machine learning, researchers also concern themselves with "representation learning" (Bengio et al., 2014). However, philosophers point out that it is not at all trivial to give a naturalistic account of representation or *aboutness*. What makes my desire to drink tea be *about* or directed at tea, as opposed to coffee, or to nothing at all? After all, stones and trees and hearts are not about anything, they just are what they are. As Nicholas Shea puts it, the problem of representation is to explain how states of brains or artificial systems manage to *reach out* and be about anything outside of themselves (Shea, 2018).

A naive solution to this problem is to say that some cognitive system's representational vehicle, $R$, (e.g., the activation of a neuron) represents some state of the world, $S$,

---

[1] PhD Student at the Institute for Neural Computation, Ruhr University Bochum (Germany)
[2] Research Scientist at the Neuromorphic Computing Lab, Intel Labs, Munich (Germany)

because *R* is caused by *S* and is thus indicative of it (Adams and Aizawa, 2021). However, this is clearly not enough. Smoke is often caused by, and is thus indicative of fire, but it does not "represent" fire in any meaningful way. A problem with this example seems to be that smoke does not do anything relevant within a broader system. Could we, perhaps, fix this by requiring that a representational vehicle, *R*, influences some other part of the system it belongs to in driving a response to *S*? Think of a row of falling dominoes in a Rube Goldberg machine. The fall of the second tile is indicative of the fall of the first, and triggers the fall of the third. Could we then say that the fall of the second tile "represents" the fall of the first tile to the third? Somehow, this example still falls short of defining a representation. To explain why the third tile falls, we only need to consider the fall of the second. This is so because whatever impact the first tile had, it is already contained in the fall of the second tile. Thus, we do not gain anything by saying that the second tile represents the first. If we want to claim that *R* represents *S*, *R* should not already causally imply the occurrence of *S*, because then *S* would become superfluous in the explanation. This means that *R* and *S* should be causally decoupled, with *R* being able to represent *S* even in *S*'s absence. This causal decoupling is the key to representations and the basis for two features which are generally regarded as the hallmarks of 'true' representations. The first one is the ability of representations to participate in sophisticated cognitive capacities such as remembering the past, imagining fictitious events, or predicting the future. The second one is that the use of representations is error-prone. Since *R* represents *S*, but is not necessarily caused by it, sometimes *R* might *misrepresent* the state of the world as being *S* (Dretske, 1986; Fodor, 1990). Your mental representation of 'snake' might be triggered by the sight of a piece of rope, leading you to falsely believe that there is a snake in front of you. This highlights that representations are "normative". They are *supposed* to indicate something quite specific, and they can get it right or wrong. Here is where things get complicated, because, what does it mean for a part of a system to be *supposed* to indicate something?

A common reply to this question involves invoking some function or purpose ("telos", in ancient Greek, from which the name of this approach, "teleosemantics", is derived) (Neander and Schulte, 2021; Baker et al., 2021). *R* is supposed to indicate *S* because that is what enables the 'proper' functioning of the system that *R* was selected for. For instance, some fly-detecting neuron in the brain of the frog is supposed to indicate flies because that is what enables the proper feeding behavior of the frog that the neuron was selected for throughout evolution. Thus, if a frog is tricked into throwing its tongue at moving black dots on a screen, the fly-detecting neuron would have misrepresented the dots as flies, or so the story is supposed to go.

A problem with the teleosemantic approach is that representations tend to become defined in historical terms (e.g., the activation of fly-detecting neurons having co-occurred with flies in previous generations of frogs). This history is external to the cognitive system that contains the representation, and therefore cannot play a causal role within it (Bickhard, 2009). It might be useful to *describe* the fly-detecting neuron as representing flies, but the representing is not doing any causal work. It would drop out of a causal explanation just like the first tile in the example of the dominoes. Furthermore, because this history is external to the cognitive system, it is only an external observer who can grasp it. So for *R* to represent *S*, we would need another observer that can already represent *R*, *S* and the ways in which they have been related to each other in the past. But this would lead to an infinite regress of interpretative homunculi. For these reasons, Mark Bickhard insists on the need for representations to have their representational content (what the representation is about, which defines the representation) defined internally within the cognitive system that has them (Bickhard, 2009). This way, the cognitive system can evaluate whether it is using a representation correctly or not, and this could account for the normativity of representations. In other words, we could say that $R$ is supposed to indicate $S$ for a system to the extent that the system itself has the ability to detect erroneous attributions of $R$. Your mental representation of snakes is supposed to indicate snakes, and not ropes, precisely because when it is caused by ropes, you can, at least in principle, later recognize the mistake.

This leads us to the notion that representations correspond to inferred latent structures. Locked up in the skull, the brain does not have direct access to the identity of objects and processes outside (e.g. the rope or the snake). These objects and processes are therefore ‘latent’ or ‘hidden’ structures (also referred to as variables, factors or causes [of the sensory input]), whose presence can only be inferred based on noisy and incomplete sensations. Note that by inference we do not mean the conscious and effortful process we engage in every other day when solving puzzles. Following Helmholtz, we conceive of inference as a pervasive and mostly unconscious process that underlies all of perception (von Helmholtz, 1867). The key about inference is that the inferred latent structures are defined internally within the cognitive system, in terms of other latent structures and, ultimately, in terms of sensorimotor primitives[1]. For example, we may decide that “snake” refers to limbless, scaly and elongated animals that sometimes bite. “Bite”, in turn, could be defined in terms of certain movements of a mouth relative to a body, pain, etc., but perhaps “elongated” could already be a sensorimotor primitive based on a certain pattern of eye movements and retinal inputs. However they are defined, latent structures are assumed to have all of their components objectively, regardless of whether they are currently apparent or not. For example, when you infer that some object lying ahead is a snake, you assume that it could move and bite you—even if it is not doing it yet. This allows you to plan actions that involve those hidden properties. For instance, you might decide to take a detour around the apparent snake. The assumption that certain properties are present also allows you to falsify your hypothesis and recognize you were mistaken. If the snake turns out to be a rope, you may realize that what you thought to be the case –that this is an animal that could move and bite you– was false. Thus, identifying representations with inferred latent structures, we can account for how representations are *supposed* to indicate what they do from the perspective of the cognitive system itself, enabling representational content to play a causal role.

An account of representation based on inference therefore looks promising. But are there other reasons for postulating that cognitive systems infer latent structures? In principle, a cognitive system could get by without doing this. For example, the system could use a huge look-up table that tells it how to react to every possible combination of inputs. Improving on this slightly, the system could also use a very large feedforward neural network with only one hidden layer, thanks to the universal approximation theorem (Hornik et al., 1989). However, this will not work in practice because the number of possible input-output mappings is too vast. Luckily, there is a lot of structure

---

[1] The sensorimotor primitives are not representational in the strict sense that we are considering here, but they correlate with things in the world, and this allows the representational system to be grounded.

in the world that can be leveraged to find more compact solutions. The world seems to instantiate a compositional hierarchy, in which higher-level entities or structures are composed of lower-level ones (LeCun et al., 2015). Societies contain people, which contain faces, which contain eyes, which contain pupils. Structures at all levels appear repeatedly. Societies contain people, but so do riots and ballet companies. People have faces, but so do fish. How to recognize the presence of some structure, and what that presence affords you is largely preserved across different manifestations of that structure. Thus, it would be a waste of time and resources if we had to learn how to recognize and interact with each structure independently for each type of complex it appears in. It is much more efficient to abstract away from particular occurrences and recognize the existence of this recurring latent structure. Moreover, such recurring structures, when applied across different domains could be the basis for metaphorical thinking and the extrapolation of past experiences to novel situations—arguably some of our most sophisticated cognitive abilities.

Interestingly, organizing latent structures in a hierarchy confers one further practical advantage. We know many different facts about each structure. We know how people look and sound like, how their knees bend, how they break your heart. Thinking about one of these facts, we can retrieve any other when relevant. This ability requires the capacity to propagate activity from the representation of each fact to that of each other fact (guided and constrained by the current context so that we do not get flooded by all possible associations). To accomplish this without a hierarchy, if we knew $N$ facts about some structure, we would need $N^2$ (all-to-all) connections between their representations. With a hierarchy, we would just need $2N$ connections, going back and forth from the representation of the structure to those of each of its components.

Deep neural networks purport to exploit such a compositional hierarchy by representing increasingly more complex latent structures or factors across layers of a neural network (Bengio et al., 2014). For instance, (Higgins et al., 2021) claim that a variational autoencoder fed with images of faces learns to represent latent structures such as the gender or the age of a face. So do deep neural networks infer latent structures, thus accounting for representations? We do not think so. As we saw above, what is required to talk of representations is some justification for how some state is *supposed to* indicate something, even when it is not being caused by it. A possible justification involves the system assuming that certain hidden properties are present, which can be proved wrong by the system itself. However, there is none of this in a feed-forward neural network. A feed forward network is, in the end, a mathematical function that is sensitive to the presence of some pattern in the input. In the example just mentioned, the network could have learned to recognize the visual appearance of age, perhaps just wrinkliness and longer ear lobes. This wrinkliness is nothing hidden that the network could be mistaken about, it is either there, “in plain sight”, or it is not. Thus, if the network, for example, classifies a young face with painted wrinkles as “old”, the network did not make any mistake from its own perspective. The network just correctly indicated the presence of wrinkles.

What about approaches that focus on Bayesian inference? In these approaches, the brain is purported to (or artificial systems are constructed such that they) learn probabilistic generative models of the environment. These models are then used to infer the most probable causes of the inputs (Fiser et al., 2010; Tenenbaum et al., 2011).

Often, the inferred causes will correspond to latent structures whose identity the cognitive system could be mistaken about. But this is not necessarily so. Consider a toy world where all an agent can do is look at one of two balls of different colors (a red ball and a green ball), and then ‘infer’ which of the two it is looking at. Even though this is a trivial task, nothing prevents one from applying Bayesian inference to solve it. However, in this kind of situation, the inferred causes (e.g., ball *A* and ball *B*) are not much more than labels for visible properties (e.g., redness and greenness). As in the case of the network detecting wrinkles, the inferred causes here do not refer to anything with hidden properties that the agent could possibly be mistaken about. Thus, Bayesian inference does not seem to get to the crux of what is required to infer latent structures and form proper representations.

To develop an approach to representation that has the required properties, we must reflect on what it means for something to be a latent structure. Something is latent because it does not manifest itself completely, at least not all of the time. Take the example of salt. Its various aspects reveal themselves under different circumstances. When looking at it, you see something white. When putting it into your mouth, you taste saltiness. When eating too much, you get thirsty. When adding it to meat, it helps to preserve it. These are different patterns of interaction between the latent structure that we call salt and ourselves, other latent structures or sensorimotor primitives. Each pattern of interaction is composed of certain elements. For example, {‘tasting’, ‘salty’} is one pattern, and {‘ate’, ‘too much’, ‘thirsty’} is another. When a certain pattern of interaction is taking place, as recognized by the fact that a certain fraction of its components are present, the whole pattern is expected to be there. Thus, under the hypothesis that something is salt, if ‘tasting’ is present, so should ‘salty’, and vice-versa; if {‘ate’, ‘too much’} is present, so should ‘thirsty’, or if {‘ate’, ‘thirsty’} is present, so should ‘too much’. Perhaps we can call patterns like {‘ate’, ‘too much’, ‘thirsty’} *conditional bistable patterns*. The patterns are conditional on the assumption that salt is involved in explaining the situation. Based on this condition, the elements of the pattern should either be mostly absent (the pattern is ‘OFF’) if the interaction is not taking place, or mostly present (the pattern is ‘ON’) if the interaction is taking place, but not somewhere in between. Thus the patterns are bistable. Latent structures seem to be defined by sets of such conditional bistable patterns.

We could then conceive of cognitive systems that explicitly encode and store these conditional bistable patterns in memory, and use them to infer latent structures. A good heuristic for the inference process could be that a latent structure is recognized to be present if at least one of its conditional bistable patterns is ON, and none of them is in an unstable state. For example, you may infer that something is salt based only on looking at some white grains. If you then taste them and they turn out to not to be salty, the interaction pattern characterized by {‘tasting’, ‘salty’} is recognized as being applicable (half of elements are present) but failing to obtain (‘salty’ is missing), and you should discard the hypothesis that what you had in front of you was salt. Missing elements are not the only kind of negative evidence for your hypothesis. The presence of unexpected elements is equally revealing. Salt should not combust, or grow mold, or make people turn blue. If any of these things happen, you will also detect a violation and realize that you were not dealing with salt. However, there are infinitely many things that should not happen in any given context, so it is impossible to explicitly learn all of them. Thus, observing anything that deviates from the relatively small set of known conditional bistable patterns should by default count as evidence against something being what you thought it was. These combinations of conditions for the (simplified) example of representing salt are summarized in Fig. 1:

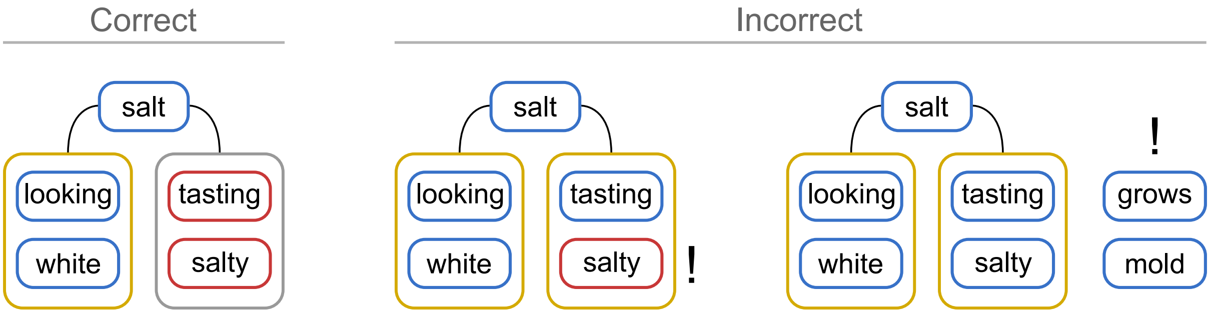
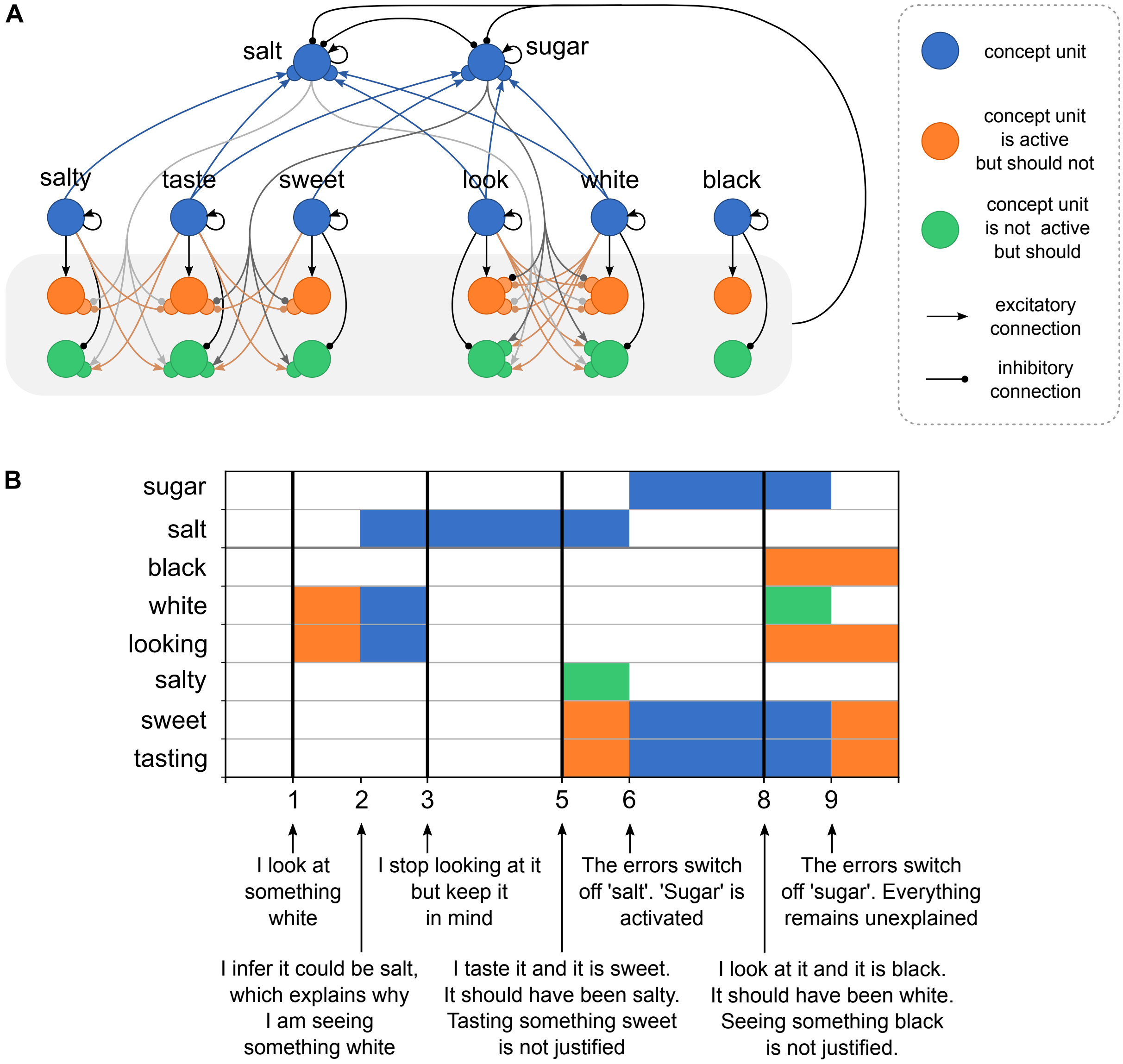


Figure 1: **Schematic representation of the concept of salt.** A node that represents a white salty substance ("salt") is defined by two conditional bistable patterns: {'tasting', 'salty'} and {'looking', 'white'}. Each element can be present (blue) or absent (red). If half of the elements in the pattern are present, the pattern becomes applicable (gold). It is correct to infer that "salt" is present when at least one of the patterns is applicable and complete, no pattern is applicable but incomplete, and no unexpected elements appear.

The unexpected sensations that count as evidence against an inferred latent structure are those that are not accounted for or "explained away" by any of the currently inferred latent structures. A natural way of detecting them is using something akin to predictive coding (Huang and Rao, 2011), where error units indicate the difference between the inputs and top-down predictions based on inferred latent structures. In our framework, such error units should be of two kinds to detect the two kinds of violations described above: either some element should be present but is not, or it is present but it should not be.

To make our proposal more concrete, in Fig. 2 we sketch a neural circuit that fleshes out the ideas put forth above, relying on the two types of error units. The neural circuit shown in the figure implements the computations required for the inference of the concepts "salt" and "sugar". The concept units representing these two inferred latent structures (blue circles in the figure) can self-sustain their activation through recurrent self-connections and compete with each other for explaining the inputs via lateral inhibition. "Salt" becomes active when either "salty" and "tasting" and/or "looking" and "white" are active. These conjunctions are detected by nonlinearities in dendritic compartments (depicted as small colored circles). This mechanism is inspired by the finding that dendrites can implement nonlinear functions, making individual neurons much more complex computing structures than conventionally assumed in artificial neural networks (Beniaguev et al., 2021). The concept units at the bottom level of the small hierarchy in our example ("salty", "taste", etc.) have corresponding error units that signal whether the concept is active but should not (orange circles), or whether it is not active but should be (green circles). These units also have nonlinear dendritic compartments which only become active when both of their inputs are active. When the patterns are in a stable state (for example, when "salt", "salty" and "tasting" are all active, or when "salt" is active but both "salty" and "tasting" are not), the excitation and inhibition arriving at the error units cancel out and the error units remain silent. Otherwise error units become active, indicating what is missing or what should not be present. The sum of all error units inhibits the concept units at the top level. One can imagine this network motif being repeated across multiple layers. The “salt” and “sugar” units could have error units themselves, which would receive top-down input from a layer higher in the hierarchy that contains, for instance units representing “anchovies” or “cake”. The lower part of Fig. 2 shows the activation of the network structures when representing different perceived situations.



I look at something white

I infer it could be salt, which explains why I am seeing something white

I stop looking at it but keep it in mind

I taste it and it is sweet. It should have been salty. Tasting something sweet is not justified

The errors switch off 'salt'. 'Sugar' is activated

I look at it and it is black. It should have been white. Seeing something black is not justified.

The errors switch off 'sugar'. Everything remains unexplained

Figure 2: **A neural circuit implementation of representations as sets of conditional bistable patterns.** **A**: The neural network implements the computations required for the inference of “salt” and “sugar”. See main text for details. **B**: A simulation of the network in action. Blue, orange and green correspond to the activation of the concept and error units, respectively, following the same color code as above. Note, however, that when the orange unit is active, so is the blue one ‘underneath’. The code for the simulation can be found here.

The presented neural circuit is just an exemplary one. We provide this suggestion here to facilitate the understanding of our proposal with a concrete example, and to stimulate further thought about possible implementations. Regardless of the details, a neural circuit that relies on calculating different types of errors between observed and expected phenomena brings important advantages for cognitive and machine learning architectures, apart from making proper representations possible. For one, checking for things out of the ordinary is arguably one of the hallmarks of common sense—something that current standard approaches in AI are notoriously bad at. This check for consistency could happen as cognitive agents passively receive inputs and form representations about them, but it could also happen actively. If time permits, or if correct identification is important, agents could trigger actions to check the validity of their emerging representations (e.g. tasting the white substance to make sure it is salt and not sugar). Second, in some cases, the error signals themselves could be used to drive behavior. For instance, if you are trying to cure meat, one of salt's error units could light up, signalling that salt is required and missing so that you can go and fetch some. And third, the error signals are useful for guiding attention and learning.

To conclude, we have proposed that mental representations can be decoupled from the input and can misrepresent (from the point of view of the system itself) because they are the result of a fallible process of inferring latent structures. Inferring these latent structures is necessary in practice, as it leads to a much more compact and flexible cognitive system that can plan actions based on hidden properties of the inferred latent structures. Deep learning and Bayesian approaches both provide valuable components for the construction of such a cognitive system. The former offers effective ways of learning hierarchies of increasingly more complex structures, whereas the latter offers sophisticated probabilistic reasoning capabilities. However, we argue that additional mechanisms are required to ensure that cognitive systems truly infer latent structures they could be mistaken about. These mechanisms could build on our proposal that latent structures are defined in terms of sets of conditional bistable patterns. Explicit knowledge of these patterns could be used in the inference process, where the detection of unexpected elements would also play a fundamental role.

We hope these views will help move the field of AI over to systems that build 'true' representations, which would lead to more intelligent systems capable of reasoning, abstraction and autonomous learning. From the neuroscience perspective, we should be more careful not to call every signal that correlates with the input a representation. A tuning curve or a receptive field is not enough. We should therefore design experiments that test whether representations can be decoupled (for example, for their use in planning or imagination), and whether they can elicit error signals. The proposed computational structure could also help explain and be validated by the connectivity patterns across cortical layers of the neocortex that form what has been dubbed a "canonical microcircuit" (Douglas et al., 1989). The wiring of neurons across layers of the neocortex seems to perform some kind of predictive-coding computation with top-down, bottom-up, and lateral connections and signal flow (Bastos et al., 2012), which might be the biological neural substrate behind the ability of the brain to represent items in the external world in the deep and fundamental way we have argued for in this paper.

# Gardener comments

**Sander Van de Cruys:**

This paper presents a clear-headed & concise overview of the issue of representation in psychology/neuroscience. I very much liked the framing of this in terms of inference, predictive coding, and latent structure. Seems on the right track to me. One small note: There is some literature on predictive coding and representation already which the author might be interested in (and could refer to). For example, see Rutar et al. (and the references therein). This one also includes a discussion on the issue of ”decoupling” (which seems in line with the current author’s ideas, but includes an interesting role for precision-weighting).

**Michael J Berry II:**

This paper presents a highly original definition of the term “representation.” Many neuroscientists use an unexamined notion in which neurons represent information when their activity correlates with external variables. However, this definition suffers from absurdities when applied more broadly, such as the atoms in a rock “representing” air pressure. The authors argue that a definition of representation that relies on causality will suffer from such situations where the term has no explanatory power.

Instead, the authors propose a framework that starts with a typical feedforward hierarchy that creates higher-level concepts as a combination of lower-level concepts but adds to it a feedback pathway that implements a set of consistency checks. These consistency checks can be thought of as predictions of other properties entailed by the higher-level concepts; these properties have, presumably, been learned through prior experience.

This framework resolves philosophical problems associated with some notions of representation, while also suggesting a concrete network mechanism. This mechanism both serves as a basis for experimental predictions as well as having the potential to be useful to AI systems.

However, the widespread use of the unexamined notion of representation in sensory neuroscience suggests that perhaps the enhanced concept proposed has a different name. Maybe “cognitive representation”? Furthermore, the specific feedback circuit proposed here would, in general, need to be learned via experience, so further thoughts on what learning rules would accomplish this goal would help extend this concept.

**Daniel Sabinasz:**

The authors present a proposal for naturalizing representational content. They start by reviewing causation-based theories of representation, and reject them via an example of falling dominoes: The fall of any given domino $D_i$ can be completely causally explained by the fall of the previous domino $D_{i-1}$, which can in turn be explained by the fall of domino $D_{i-2}$. Thus, if domino $D_{i-1}$ were said to represent the fall of domino $D_{i-2}$ to $D_i$, this notion of representation cannot do any causal explanatory work. In order for $D_{i-1}$’s representing $D_{i-2}$ to do any explanatory work, $D_{i-1}$ would have to be causally decoupled from $D_{i-2}$, so that the occurrence of $D_{i-1}$ does not causally imply the occurrence of $D_{i-2}$. This is taken to show that a representational vehicle $R$ can only represent some state *S* in a causally efficacious way if *R*'s occurrence is causally decoupled from *S*. I find the domino example a little hard to follow and to relate to representations in cognitive systems. An example pertaining to vehicles that participate in cognitive capacities while not being causally decoupled from what they purportedly represent may have been easier - e.g., a feedforward neural network.

The authors go on to argue that teleo-semantic approaches fail to provide a notion of causally efficacious representations, given that their representational content is defined historically and may thus not have a causal effect in the cognitive system here and now. Still, they grant that such approaches may be useful to describe(!) vehicles as representing something. In my opinion, the authors make implicit assumptions here regarding the optimal level at which causal interaction ought to be described - especially with respect to the admitted time scale of causal interactions, and with regard to how the cognitive system is to be circumscribed spatially and temporally. I contend that it is perfectly legitimate to say that the "snake neuron"'s representing snakes in evolutionary history is causally involved in the cognitive system's flight behavior in the here and now. I simply adopt a different scientific level of description / perspective in saying this - a perspective that I contend to be neither generally less nor generally more useful than the perspective that the authors adopt.

Next, the authors propose a theory of representations as inferred latent structures, which have the property that the representational content is defined inside the cognitive system itself - by virtue of relating a concept's representational content to other concepts internal to the system. That notion of representation is then argued to be causally efficacious, given that the cognitive system itself can identify that the representational vehicle misrepresented, and act accordingly.

The notion of representation as hierarchical latent structure is then backed up by theoretical arguments, evidence, and a computational model. The necessity of completable or falsifiable incomplete patterns and recurrent connection is identified, refuting the claim that feed-forward neural networks or Bayesian models can have representations that are causally efficacious in the here and now.

While I contend this to be a useful notion of representation, I do not agree with the authors that it is the only useful notion of representation that can be causally efficacious at some level of description, for reasons laid out in the second paragraph - namely, the commitment to a particular scientific level of description that I cannot quite make explicit yet based on the present paper. To mimic the authors' line of argumentation in refuting teleo-semantic notions of representation, I could say that the state of the atoms in the cognitive system at any given moment in time is completely causally explained by the prior state of those atoms and the laws of physics. Therefore, the representational content of latent inferred structures cannot do any additional causal explanatory work.

Despite this single point of criticism, I'm generally positive about this paper, given that it deals with the right questions and treats them thoroughly and in a well thought-through manner. Even in disagreement about some points, I still got a lot of food for thought out of the reading.

**R. Sal Reyes:**

This is an interesting exploration of how a system might create mostly-accurate &

causally-potent internal representations of things that exist in the outside world. It seems that this sort of model could have some useful applications within AI—helping such systems to better ”see” things as more than just objects that are a collection of physical attributes. However, I don’t believe it truly correlates to how these kinds of systems likely function in a human brain (an entity that ultimately works very differently from AI neural networks, which are tasked to accomplish far narrower sets of goals—to put it mildly).

First, from my perspective, the “latent structures” that are referred to here (which carry all of the relevant content that makes any “representation” capable of being “about” something) are words and language. And if language is, indeed, the basis of such “representations” having meaning, then the mix of reflexive & deliberative, sensory & cognitive processes that combine to produce any singularly & momentarily perceived mental “representation” are ultimately likely to be more complex than the ”conditional bistable patterns” model presented here. (Such as the almost absurdly complicated model of language-based cognition presented by Halliday & Matthiessen in ”Construing Experience Through Meaning.”)

One quick example that shows just one of the many aspects of this process (in a human brain) that this model leaves unaccounted for… Out of the corner of your eye, you notice a dark blur rustle behind the ferns in your yard. In reality, this blur is the product of a simple shadow. Let’s say that under normal circumstances, this exact sensory experience automatically “represents” the dark blur correctly as a shadow, allowing you to perceive it as a shadow & have any appropriately-related thoughts about it, e.g., “what a lovely backdrop for the fern.” However, if your longtime & beloved black cat has recently disappeared, this exact same sensory experience might automatically “represent” the dark blur as your black cat, allowing you to literally (& cruelly) see your lost cat out of the corner of your eye behind the fern—only to realize moments later that your current obsession over your missing companion is leading you to incorrectly (but literally) see them in every dark blur.

If I’m interpreting the model here correctly, it appears that this exact same sensory experience should always produce the same initial “representation” (correct or not) regardless of any currently missing feline friends. (Since there is no change in any of the sensory pattern data, nor any difference in any concept units that are active but shouldn’t be, or vice versa.) In fact, since the normal scenario where your cat *is not* missing is technically a circumstance in which seeing them in your yard is *more* likely to be accurate, according to a model based on these principles, you should be more prone to misperceive the blur as your cat in normal circumstances—although my personal experience with missing pets (and those I’ve heard from others) has demonstrated the opposite. In contrast to all this, a language-based associative & cognitive system has several avenues for both correctly & incorrectly “representing” & perceiving the exact same data in different ways based on any number of circumstantial or emotional factors.

**Anonymous1:**

The view that ”true” mental representations should be the results of an inference process is new, as far as I know. However, I am not satisfied by the node network model the author proposed, because I don’t see how it is different from a traditional artificial neural network. The activities of this network can still be described fully mechanistic and the content of each node (being white, saltiness, sugar, etc.) still seems to be assigned by an outside observer. I don't think the author solves the problem that any definition of ”true” representations should explain why a certain brain state is supposed to represent other things. Nevertheless, it is still a novel idea and deserves to be read by the science community. However, I think the authors should provide more background explanations about all the philosophical terms and arguments they mentioned in the first section.

**Ted Wade:**

I have a feeling that there are a lot of open, unexplored currents here from philosophy, cognitive science and AI. That and the connection from theory to practice make the paper worthwhile.

I have a couple of issues.

The paper gives its worked examples in cases where representations are basically like so-called ”concrete nouns”, whose corresponding objects are identifiable through the senses. The inferred latent structures are sort of like attributes or properties of the representation, and the examples use some properties (like “salty”) that can be determined through the senses. So, in the theory representations are *of a different kind* from latent structure properties.

But, surely representation is not just restricted to representing concrete objects. Don't we also have to “represent” other kinds of things, like processes or actions (warming, decay, tasting, looking) and attributes (saltiness, whiteness)? And can't some of these representations *also be* latent structure properties?

This becomes more of an issue when we talk about using representations in memories or counterfactuals, when direct sense impressions are not available. Don't we then have to rely on representations of tasting and saltiness? And the same would be true if it were someone else's sense impressions being used as evidence for/against the applicability of a representation

I admit much ignorance of the ways in which representation has been defined and argued in the fields relevant to this paper. However, it might be helpful to explain to readers how the paper's definition of a representation is not circular in the sense of defining a representation in terms of more representations.

My second concern is whether “checking for things out of the ordinary” has been sufficiently explained. Given the size of the set of un-ordinary things, how do we decide that a new observation is actually relevant to the use of a representation? For a concrete object like a pile of salt, some relevance can be signaled by physical and temporal co-location of the pile and unexpected sensations, like a color change, a smell, or dissolving in water. What if a pile of salt kills grass, or stains cloth and then gets washed away before we see it? Are there other or competing signals of relevance besides co-location, and if so, how could the theory incorporate them?

**Dan James:**

A compelling and superb co-ordination of accurate philosophical notions concerning mental representation and the possibilities of translating these into practical strategies for building computational structures. Consequently, though fairly brief and concise, this paper is an exemplar of Applied Philosophy and one that has potentially important implications for AI research. I found the arguments were developed convincingly and clearly, with good explanatory graphics and the whole paper to be well-researched. I have no hesitation in recommending this paper for publication as, at the very least, it is clearly a good discussion point/seed for future cognitive science.

**Fred Nix:**

I am only an interested layman in this area, but I would strongly recommend this article for publication in Seeds of Science. It seems to suggest a novel idea and the possibility of experimenting on a very interesting problem.

There are a couple of observations that popped out at me when I read this. First, is a similarity to some of the structures discussed by Marvin Minsky in The Society of Mind. The second are the similarity to ideas in Wittgenstein's Philosophical Explorations.

As to Minsky, the latent structures have a similarity to his "language frames" or the "frame" idea in general (Minsky, Society of Mind, 1986, p. 261). Our minds run a comparison to pre-existing, similar frames, and try to plug in different expected values into the empty terminals to see if there might be a good fit. Minsky looks at the sentence "Mary was invited to Jack's Party," and speculates it raises the "party invitation" frame in the mind. Which then prompts the mind to try to fill in the blank terminals: "Who is the host?", "Who will attend?", "What present should I bring?", "What clothing will I wear?" Certain terminal assignments are assigned by default based upon experience, but could be accepted, replaced or objected for the particular situation.

The way your latent structure model of comparing expected to observed variables looks to me to have much promise, as it gets away from the untestable philosophical mess and opens up a whole slew of statistical analysis tools that can be used in machine learning.

Second, was that I really like the similarity to Wittgenstein's Philosophical Investigations. He starts out his book with a similar philosophical mess coming out of Augustine's work where every word has a meaning correlated with it, and the meaning is the object for which the word stands. But Wittgenstein's approach, and yours, is much more flexible and I think closer to how our actual minds function. The emphasis is on the way language functions in context, the famous "language game." There is no perfect representation for each word. You have to identify the role the work plays in the particular language game at issue. What is my frame? What latent structures could be deployed that would work with this game? Is my split-second hypothesis on the rope vs. snake correct in the particular game I find myself in? In potentially dangerous matters, should I change my p-value and take a bigger risk on being wrong? Just fascinating how this could be done in machines.

One small editing thing that threw me off a bit was in Table B where the numbers were not consistent left to right. Some descriptions didn't have numbers, and I was looking for numbers to keep guiding me through the table.

Anyway, great job, I'd really like to hear how this idea works out!

**Michael M. Kazanjian:**

A useful collection of insights into representation and cognition. The author does not mention Kant, though Kant probably is a crucial element in representation theory: we know via categories instead of the thing itself or an objective reality out there. Thanks, Michael M. Kazanjian.

**DK:**

I enjoyed reading this article. It should be a valuable addition to the debate on mental representations.

**Mark:**

This is interesting and clearly written. There are a few small typos, most notably, casual should be causal in at least one place.

**Mike Wolf:**

Here are some comments from my review of "What does it mean to represent? Mental representations as falsifiable memory patterns.

I think that the paper contains ideas worth cultivating. I was particularly interested in the ideas of "representation" and of "inferred latent structure." By happy coincidence, I have been doing some experiments with GPT-3. I think the paper might benefit from considering its premises as applied to GPT-3.

GPT-3 is considered to be a kind of AI. It seems to have knowledge of the world, but it is designed as a "language model" that is able to predict next sentences from context. I am not an expert on GPT-3, but from what I have learned, it seems that GPT-3's representation can be understood by analogy to the way a hologram represents a scene. A hologram is a two-dimensional representation of an arrangement of objects in a three dimensional space. An image of the scene can be reconstructed from the hologram. Depending on how the hologram was created the reconstruction will represent the scene with more or with less detail. The hologram is thus a two-dimensional representation of the scene, and the reconstruction is a kind of re-presentation of the original scene. Although the objects occupy discrete areas in the 3D space of the scene, the objects are not represented in the discrete part of the hologram. If a section of the hologram is corrupted or removed, then detail in the re-presentation is lost, but no object is removed. One might say that all the objects seen in a displayed holographic image are latent inferred structures.

It seems to me that GPT-3 works in a similar way. GPT-3 does not model concepts, rather it looks for patterns. There is no part of GPT-3's world model that corresponds to "salt." Rather it has digested a large corpus of text in which words like "salt" and "sugar" and "salty" and so on appear and GPT-3 stores associations from which it can make predictions. It is not clear that it is representing information in the way described by the paper.

I did not understand the paper's model of a ”neuronal circuit.” The paradigm in my mind may be sufficiently different from the one that the paper relies on that I could not make the mapping from what I understand of GPT-3 to what I understand of the paper's ideas. I don't understand the proposed implementation. This could simply be a result of my lack of understanding of the domain. Is GPT-3 an AI? The wikipedia page describes it as ”a language model” but cites people who call it an AI. This Forbes article, What Is GPT-3 And Why Is It Revolutionizing Artificial Intelligence?, argues it is an AI.

To help me understand how GPT-3 works, I asked some questions about the content of the paper. The following interchanges are extracted from a transcript of a session. I've changed the order of prompts and completions for clarity (and because I had not yet learned how to save a session in progress). GPT-3's responses to a question depends on session context as well as its data model. It is also not deterministic, so if you asked the same question you would probably not get identical results. I've found that even when trying to repeat a sequence I've sometimes gotten wildly different results. Nonetheless, all the responses that I attribute to GPT-3 came from GPT-3. I've also made some comments between the transcript extracts.

> **How does GPT-3 represent the world?**

GPT-3 represents the world by storing information about the world in a database. This information can be used to answer questions about the world, or to make predictions about the future.

> **How is GPT-3 different from representational models?**

GPT-3 is different from representational models because it does not explicitly represent concepts. Instead, it learns to recognize patterns in data and uses this knowledge to make predictions.

> **How is its knowledge stored?**

GPT-3's knowledge is stored in a database. This database can be used to answer questions about the world, or to make predictions about the future.

> **When GPT-3 recognizes patterns in data, is that pattern stored in some way?**

GPT-3 does not explicitly store patterns that it recognizes in data. However, the knowledge that it learns from these patterns is stored in a database. This database can be used to answer questions about the world, or to make predictions about the future.

> **Is the idea of an ”inferred latent structure applicable to GPT-3?” If so, in what way?**

The idea of an ”inferred latent structure” is applicable to GPT-3 in the sense that GPT-3 can learn to recognize patterns in data. These patterns can be used to answer questions about the world, or to make predictions about the future.

(Besides asking questions, you can interact with GPT-3 by starting to type a response and having GPT-3 complete what you started. This is called a ”completion” and is actually how GPT-3 answers questions. Any answer is just the completion for the question asked.

To get GPT-3's ”thoughts” on ”inferred latent structures” I typed this to prompt it: )

>**GPT-3's inferred latent structures**

and got this completion

> ...are not always accurate. What can you do about this?

There is always a trade-off between the accuracy of a model and the interpretability of a model. GPT-3's latent structures are designed to be accurate, but they may not always be interpretable. If you find that GPT-3's latent structures are not always accurate, you can try to increase the interpretability of the model by using a different model, or by using a different method to learn the latent structures.

(Note: the ”What can you do about this question” appears to have been part of GPT-3's completion. This is surprising, but GPT-3 has done other surprising things in the past. In any case, the completion continued with a response to GPT-3's question)

## Authors' response to the gardeners

We would like to thank the gardeners for their very thoughtful summaries and critiques. They point to numerous avenues that could be explored further. How to learn these representations? To what extent is language necessary for representational content? How to determine which concept is responsible for explaining which observations? These are very complex questions and we do not have good answers for them, so we will leave them open here.

We have mostly limited our editing of the manuscript to attempting to clarify one issue. A few gardeners have described our approach as defining representations/concepts in terms of other representations/concepts, which would end up being circular. This was a fair reading of the text, but not what we originally had in mind. Therefore, we have modified paragraph 5 (starting with “This leads us to the notion that...”) to try to clarify this point.

## Acknowledgements

Many of the ideas expressed here originated in debates with José R. Donoso around the notion of mental representation, and were sharpened by further discussions with Alexander M. Hölken, Daniel Sabinasz, Xiangshuai Zeng, Nicolas Diekmann, Laia Serratosa, Olya Hakobyan, Sen Cheng, Gregor Schöner and Mark Bickhard. We thank them for their insightful engagements and friendly pushback. Finally, we thank the editors of the journal for creating this unique space that makes doing science fun again.

## References

Adams, F. and Aizawa, K. (2021). Causal Theories of Mental Content. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2021 edition.

Baker, B., Lansdell, B., and Kording, K. (2021). A Philosophical Understanding of Representation for Neuroscience. *arXiv:2102.06592 [q-bio]*.

Barack, D. L. and Krakauer, J. W. (2021). Two views on the cognitive brain. *Nature Reviews Neuroscience*, pages 1–13.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., and Friston, K. J. (2012). Canonical Microcircuits for Predictive Coding. *Neuron*, 76(4):695–711.

Bengio, Y., Courville, A., and Vincent, P. (2014). Representation Learning: A Review and New Perspectives. *arXiv:1206.5538 [cs]*.

Beniaguev, D., Segev, I., and London, M. (2021). Single cortical neurons as deep artificial neural networks. *Neuron*, 109(17):2727–2739.e3.

Bicanski, A. and Burgess, N. (2020). Neuronal vector coding in spatial cognition. *Nature Reviews Neuroscience*, (9):1–18.

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

Douglas, R. J., Martin, K. A., and Whitteridge, D. (1989). A Canonical Microcircuit for Neocortex. *Neural Computation*, 1(4):480–488.

Dretske, F. I. (1986). Misrepresentation. In Bogdan, R., editor, *Belief: Form, Content, and Function*, pages 17–36. Oxford University Press.

Fiser, J., Berkes, P., Orbán, G., and Lengyel, M. (2010). Statistically optimal perception and learning: From behavior to neural representations. *Trends in Cognitive Sciences*, 14(3):119–130.

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

Gallistel, C. R. (1990). *The Organization of Learning*. The Organization of Learning. The MIT Press, Cambridge, MA, US.

Hartley, T., Lever, C., Burgess, N., and O'Keefe, J. (2014). Space in the brain: How the hippocampal formation supports spatial cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1635):20120510–20120510.

Higgins, I., Chang, L., Langston, V., Hassabis, D., Summerfield, C., Tsao, D., and Botvinick, M. (2021). Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nature Communications*, 12(1):6456.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366.

Huang, Y. and Rao, R. P. N. (2011). Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(5):580–593.

Knutson, B., Taylor, J., Kaufman, M., Peterson, R., and Glover, G. (2005). Distributed Neural Representation of Expected Value. *Journal of Neuroscience*, 25(19):4806–4812.

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Neander, K. and Schulte, P. (2021). Teleological Theories of Mental Content. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2021 edition.

Parra-Barrero, E., Diba, K., and Cheng, S. (2021). Neuronal sequences during theta rely on behavior-dependent spatial maps. *eLife*, 10:e70296.

Quiroga, R. Q. (2012). Concept cells: The building blocks of declarative memory functions. *Nature reviews. Neuroscience*, 13(8):587–97.

Shea, N. (2018). *Representation in Cognitive Science*. Oxford University Press.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to Grow a Mind: Statistics, Structure, and Abstraction. *Science*, 331(6022):1279–1285.

von Helmholtz, H. (1867). *Handbuch Der Physiologischen Optik*. Leopold Voss.