Synthese (2021) 199:1307–1325
https://doi.org/10.1007/s11229-020-02793-y

INTERDISCIPLINARY PERSPECTIVES ON LOCATING REPRESENTATIONS IN BRAIN

The physics of representation

Russell A. Poldrack[1]

Received: 15 February 2020 / Accepted: 10 July 2020 / Published online: 30 July 2020

Abstract
The concept of "representation" is used broadly and uncontroversially throughout neuroscience, in contrast to its highly controversial status within the philosophy of mind and cognitive science. In this paper I first discuss the way that the term is used within neuroscience, in particular describing the strategies by which representations are characterized empirically. I then relate the concept of representation within neuroscience to one that has developed within the field of machine learning (in particular through recent work in deep learning or "representation learning"). I argue that the recent success of artificial neural networks on certain tasks such as visual object recognition reflects the degree to which those systems (like biological brains) exhibit inherent inductive biases that reflect the structure of the physical world. I further argue that any system that is going to behave intelligently in the world must contain representations that reflect the structure of the world; otherwise, the system must perform unconstrained function approximation which is destined to fail due to the curse of dimensionality, in which the number of possible states of the world grows exponentially with the number of dimensions in the space of possible inputs. An analysis of these concepts in light of philosophical debates regarding the ontological status of representations suggests that the representations identified within both biological and artificial neural networks qualify as legitimate representations in the philosophical sense.

Keywords Representation · Neural networks · Inductive bias · Machine learning

1 Introduction

The ontological status and epistemic utility of mental representations are topics of enduring debate within the philosophy of mind. Neuroscientists have forged ahead largely unaware of these debates, using the term widely to describe the systematic

✉ Russell A. Poldrack
russpold@stanford.edu

[1] Stanford University, Stanford, CA, USA

empirical relationships that are often found to exist between neural activity and features of the external world (Vilarroya 2017). Recent work in computer science has also begun to use the term in the context of structured relationships within a machine learning system's input and/or output features. My goal in this paper is to argue that this empirical and computational work provides important insights regarding philosophical questions about the ontological status of representations in the brain.

I will first lay out how the concept of representation is used within neuroscience, highlighting the degree to which the term is applied across multiple scales of brain activity. I will then argue that we can gain substantial traction in understanding representations from work in machine learning that has focused on the learning of representations (also known as "deep learning"), using the particular example of the recognition of visual objects. I will focus on the fact that these networks are able to learn very difficult problems with relatively few parameters compared to the complexity of the input, and argue that this reflects an inductive bias that arises from the particular architecture of the networks, which was inspired by biological visual systems. I argue that the representations observed in both natural and artificial neural networks do real representational work, meeting Ramsey's "job description" and qualifying them as legitimate examples of subpersonal representations under at least some philosophical schemes. Based on a set of additional insights from machine learning theory, I will argue that any system that behaves adaptively in the world *must* contain internal states that bear systematic causal, informational, and structural relationships with aspects of the world. Any system that does not contain such representations is destined to fail due to the fact that there are simply too many possible combinations of features to effectively memorize, known in computer science as the "curse of dimensionality".

Representation in neuroscience

The use of the term "representation" in neuroscience generally refers to a systematic relationship between features of the natural world and the activity of neurons in the brain. The presence of neural responses that are organized in a way that is structurally isomorphic with the external world has been known for more than a century, since the first discoveries in the nineteenth century that the somatosensory cortex (which responds to physical stimulation of the skin) is arranged in the brain in a way that maintains some of the spatial organization of the physical body (Wilson and Moore 2015). By "structurally isomorphic" I mean here that there is a systematic relationship between the activity of neurons and the structural (usually spatiotemporal) features of the world, such that the larger-scale organization of neural activity maps onto the structure of the world at the relevant scale. In particular, one of the foundational findings in neuroscience is that most of the brain's primary cortices are organized in a topographic manner representing the spatial (visual, somatosensory, motor) or temporal (auditory) structure of the physical body or external world. The discovery of systematically organized patterns of activity that are isomorphic with the external world has also extended well beyond primary cortices. Well-known examples include the spatial organization of cortical columns in area MT (Albright et al. 1984), in which adjacent columns respond to visual motion in systematically related directions;

the “salience maps” of the parietal cortex that topographically reflect the intention to make an eye movement to a particular location in space (Goldberg et al. 2006); and the delay-sensitive neurons of the prefrontal cortex that reflect the intention to make a future eye movement in a particular direction (Goldman-Rakic 1995). Perhaps the most famous example is the “place cell” of the hippocampus (O’Keefe and Dostrovsky 1871), each of which responds most strongly when the animal is in a particular location in extrapersonal space.

In each of these examples, the use of “representation” is commissioned by the fact that there is a systematic relationship between the structure of the world and the pattern of neural activity. In some cases the macroscopic spatial organization of neural responses is isomporphic with the structure the world, such as the retinotopic organization of neural responses in the visual cortex, the tonotopic organization of the auditory cortex, and the somatotopic organization of the somatosensory cortex; in these systems, adjacent neurons respond to adjacent portions of the relevant stimulus space. In other cases, such as place cells in the hippocampus, the macroscopic spatial organization of neuronal responses is not isomorphic with the external world (i.e. nearby place cells do not necessarily code for nearby regions of space), but the firing patterns of neurons nonetheless exhibit a systematic relationship with the organism’s experiences in the spatial world. Regardless of the lack of macroscopic isomorphism, most neuroscientists are comfortable stating that these neurons “represent” the particular features of the external world to which those neurons respond.

The content of neural representation is often identified with a neuron’s “receptive field”, a concept generalized from the study of early sensory cortices that refers to the specific features that cause a neuron to fire most vigorously and selectively. The work that establishes this kind of knowledge assumes that the degree to which a neuron represents a feature of the world can be inferred from its pattern of activity, such that higher levels of firing occur when the stimulus is closer in the relevant feature space to that neuron’s ideal stimulus. Further definition of function is sought by identifying the boundary conditions for increased activity, seeking to establish selectivity for particular features. However, there is some degree of functional indeterminacy inherent in any ascription of function to a neural system based on this approach, directly analogous to that noted in the philosophical literature on representations (e.g. in the context of teleosemantic theories that posit selection for a particular function; Neander 1995). For example, presenting an image of a natural scene on computer screen causes strong activation in primary visual cortical area V1. Knowing only this, it would be impossible to determine the fundamental function of V1: Is it an “images on computer screens detector”, a “scene detector”, a “complex visual stimulus” detector, or a detector of simple features such as edges or color patches? The classic approach to this problem is to systematically decompose the stimulus (based on a priori knowledge of the stimulus domain), through a cycle of abductive and hypothetico-deductive reasoning. For example, the early understanding of the receptive fields of the visual system by Hubel and Wiesel was initially spurred by an accidental finding that led to experiments that reduced the stimulus to its most basic elements (lines of various widths and orientations, moving in various directions) and examined the responses of individual neurons to those elemental stimuli.

A more recent example comes from the study of high-level visual perception, in which a debate has raged regarding the functional description of a particular region in the fusiform gyrus that runs along the bottom of the temporal lobe. Early research had demonstrated that this area was active in response to human faces, and in 1997 Kanwisher et al. proposed that the region was *specialized* for the processing of faces, naming it the "fusiform face area" (FFA). This ascription was made based on the relatively strong response of this area to faces compared to other classes of objects; note, however, that the region is also responsive to those other object classes, simply more so to faces. This strong claim of functional specialization was subsequently criticized from two distinct directions. One set of studies claimed that the region was not in fact specialized for face processing, but instead for the expert processing of stimuli at a subordinate level of categorization, which encompasses faces but could also encompass other non-face stimuli for which the individual has expertise. Gauthier et al. (1999), for example, showed that this area was activated when car experts or bird-watchers viewed objects from their area of expertise, but not other objects, leading to an alternative description of the area as the "flexible fusiform area" (since the "FFA" moniker had become common parlance following the original Kanwisher paper). Another critique was based on the use of pattern analysis methods (to be described in more detail in the following section), which showed that objects from non-face categories (such as cats, scissors, and chairs) could be distinguished with high accuracy based on fMRI signals in the supposedly "face-selective" region (Haxby et al. 2001), suggesting that at the regional level there may be a blending of these functions.

A related approach uses statistical models known as "encoding models" to more directly assess the relationship between stimulus features and neural activity. In this case, a statistical model is developed that includes factors that describe each feature of interest, allowing the researcher to determine statistically the degree to which any particular feature is uniquely associated with neural activity. This approach helps address a challenge with many cases of functional ascription, which is that there are often multiple possible explanatory features that are are correlated with one another; for example, humans are inherently expert at the recognition of human faces, such that faceness and expertise are necessarily correlated. In order to disentangle multiple possible explanatory features, one must identify the degree to which they uniquely explain the signal of interest. The encoding model approach takes advantage of the fact that when multiple correlated features are combined within a standard linear statistical model, the model will identify the degree to which each feature accounts for unique variability within the signal, thus providing statistical evidence for or against a particular functional ascription. As an example, Engelhard et al. (2019) recorded the activity of dopamine neurons in mice while the animals navigated a virtual-reality maze. The investigators developed a statistical model that included many different aspects of the mouse's experience, from visual features to motor activity to expected and received reward, and fit this model to the activity of each individual dopamine neuron. The advantage of this approach is that the linear model estimates the *unique* contribution of each variable to the neuron's behavior, effectively discounting correlated contributions and thus resolving functional indeterminacy that would arise if the contributions were being assessed independently. In this particular instance, the results from the encoding model demonstrated that dopamine neurons represent a much broader set of features than had been expected based on previous theories. Importantly, the scope of inference from this approach is limited to features that are included in the model; it is always possible that the true functional ascription relates to some correlated feature that is not included in the model.

Representational spaces

In recent years, neuroscientists have moved towards a conceptualization of "representational spaces", driven in particular by work that has used pattern analysis methods to understand the larger-scale organization of neuronal responses (e.g. Davis and Poldrack 2013). In this approach, high-dimensional patterns of neural activity (such as functional magnetic resonance imaging [fMRI] data or simultaneous recordings from large numbers of neurons) are examined across a range of stimuli, and the similarity of the neural responses is used to identify a low-dimensional projection of the responses across stimuli that is often referred to as a "representational space". A classic example of this approach comes from Kriegeskorte et al. (2008), who examined the response of the inferior temporal cortex to a diverse set of visual images in both humans (using fMRI) and macaque monkeys (using electrophysiological recordings). In each case, a clustering algorithm was applied to the response to a large number of simultaneously recorded neurons/voxels in order to visualize a low-dimensional embedding of neural responses to specific images. Across species there was a highly similar clustering of visual images, with a distinction between living and non-living objects at the top level, followed by subcategories within those larger categories (such as faces versus body parts, or natural scenes versus objects). In fact, analysis procedures of this sort are now commonly referred to by the general name of "representational similarity analysis", which implies that the patterns of activity being measured in the brain (and whose similarity is being assessed) *are* the representations.

One challenge with the analysis of similarity spaces is that they are fundamentally indeterminate: There is no single "correct" similarity space within which to compare patterns of activity, just as in general there is no single "correct" decomposition of a dataset into lower dimensionality. As just one example, one can perform a matrix decomposition that either allows all values or only non-negative values, each of which is perfectly legitimate. Further, when dimensionality reduction methods are applied to neural data where the underlying representational space is known (such as area MT, which is known to represent the direction and speed of visual motion), the resulting decompositions do not always provide a clear picture of the (known) representational structure (Goddard et al. 2018).

The most common uses of representational similarity analysis attempt to sidestep this issue, by comparing similarity spaces computed according to a common similarity metric. For example, we might compute the similarity between patterns of neural activity for different stimuli using a particular similarity measure, and then further compare the second-order similarity of these similarity patterns across species or experimental conditions. This does not absolve the approach of indeterminacy; it simply pushes that indeterminacy down to a level below the inferences that are being made. However, neuroscientists are generally comfortable endorsing claims about the similarity of representational spaces, despite the fact that there is no unique underlying space in which they can be defined.

Representation in computer science

Classically, the concept of "representation" has been used in computer science to refer to the format of stored information; for example, in a digital computer integers are "represented" using a binary system. However, the term is increasingly used to refer to structured informational systems that bear substantial resemblance to the representations discussed in neuroscience, particularly within the study of deep artificial neural networks, known as "deep learning" (LeCun et al. 2015). Few readers outside of that field are probably familiar with the fact that the field of deep learning also goes by the name of "representation learning", to which an entire conference is dedicated (the International Conference on Learning Representations). Here too the concept refers to the format of the information, but the goal in representation learning is not a lossless transformation (for example between the decimal 8 and its binary representation 1000). Instead, the structure of the information is changed in service of some task; as noted by David Marr in his famous explication of the nature of visual representation, "how information is represented can greatly affect how easy it is to do different things with it" (Marr 1982). For example, given some grayscale image of an object (like my cat in Fig. 1), the most natural format for the image is a two-dimensional matrix of gray levels. For some operations this format might be useful (for example, if we wanted to compute the average brightness of the image), but for other operations there are likely to be different formats that could be more useful. For example, if we want to find different objects in the image, it might be useful to know where the edges are located in the image, in order to find the boundaries of the different objects. On the other hand, if we specifically want to know whether there is a cat in the image, it would be useful to directly apply a "cat detector", returning an image that provides the probability of a cat being present at each location in the image. The goal of representation learning in this context is to find ways to transform the original features that will result in optimal performance on a some particular task, such as classification or prediction.

Many in the philosophy literature will be familiar with the "connectionist" neural network models that became popular in the 1980s. These models were the source of substantial discussion and controversy regarding the degree to which they could instantiate symbolic representations (Fodor and Pylyshyn 1988; Pinker and Prince 1988; Smolensky 1988). The deep learning approach borrows fundamentally from these models, but takes advantage of several advances that have allowed these models to greatly outperform classical connectionist models, sometimes achieving human-level performance on difficult tasks such as object recognition or natural language processing. The underlying structure of deep learning models is very similar to those original models, comprising units loosely inspired by neurons, with learning occurring through changes in the strength of connections between units. The primary advance is evident in the name: these new models are much *deeper*, in terms of the number of hidden layers, compared to classical connectionist models, which often only had a single hidden layer. The ability to effectively train deep neural networks has been

Fig. 1 A grayscale image of my cat Coco wearing a hat. This image contains 1434 X 1076 ~ 1.5 million pixels. With 256 possible gray levels in each pixel, there are $256^{1542984}$ possible distinct images of this size that could be generated—far more than the estimated number of atoms in the universe. Creating each of these images, one per nanosecond, would take many orders of magnitude longer than the time elapsed since the Big Bang

afforded by a number of theoretical advances as well as improvements in computing hardware and the availability of massive training datasets such as ImageNet (Deng et al. 2009). This depth allows current models to learn much more complex and hierarchical representations than earlier models; this appears to be particularly important to allow the network to learn to rely on task-relevant signals rather than nuisance variation (Buckner 2018).

An artificial neural network is characterized by three features (Richards et al. 2019): Its *architecture*, an *objective function* that it aims to maximize (i.e. a formal measure of accuracy of the task to be performed), and a *learning rule* by which it is trained to maximize that objective. Here I will focus particularly on a class of network architectures that has gained prominence for the identification of visual images, the hierarchical convolutional neural network (which I will abbreviate as *HCNN*), a schematic example of which is shown in Fig. 2. This class of network architectures is characterized by a set of *layers* that can perform various simple mathematical functions on inputs

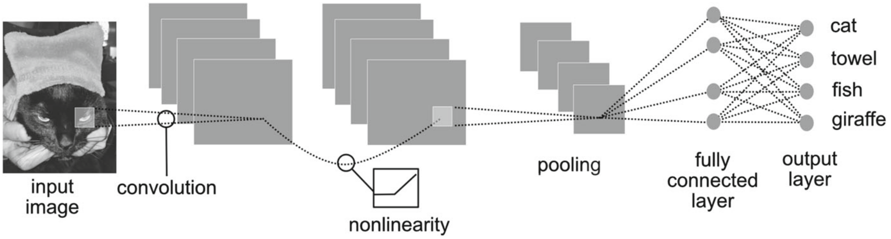

Fig. 2 Example of a convolutional neural network architecture. Individual layers perform simple computations, such as convolution with a kernel (effectively a weighted average), imposing a nonlinearity (e.g., by setting all values below some threshold to zero), and pooling across sets of lower units. Other layers are fully connected to one another, as in a standard connectionist network. A deep HCNN would stack many sets of these layers between input and output

(either from the world, or from another layer); what makes it "deep" is the presence of a large number of such layers. The "convolutional" aspect of the HCNN refers to a particular aspect of its architecture, in which units at one layer perform a weighted average (or "convolution") over a subset of units from an earlier layer; this particular feature (along with a number of others described in Fig. 2) has been strongly inspired by the structure of biological visual systems. An HCNN is trained to perform a particular task (which is embodied in the objective function) through a learning algorithm that aims to optimize this objective by changing the parameters of the model. The objective function quantifies the degree to which the network's output matches the goal state of the task; for example, if the task is to identify the category of an object out of 1000 possible categories, then the objective might be some function of the distance between the actual probability of each class for a particular image (which will be 1 for its true category and zero for all others) and the estimated probabilities for each category generated by the model. Each layer has parameters that are learned, which include the kernels used for convolution as well as the weights on the connections between layers. These parameters are modified on the basis of the error for each image during training using an error-correcting learning mechanism such as backpropagation, in which the errors at the output layer of the network are propagated backward and used to to adjust weights at each lower layer.

Over the last decade there has been rapid progress in the ability of HCNNs to perform object recognition tasks. For example, when the image of my cat from Fig. 1 is presented to the VGG19 neural network (which was the state of the art in 2014), the model estimates a 19% probability that the picture is a Siamese cat, and a 10% probability that it is a bath towel. The state of the art in 2019, implemented in the Clarifai image recognition tool[1] classifies the image as a "cat" with 99.5% probability, even though the system has almost certainly never seen an example of a cat wearing a hat before. Deep learning approaches have made similar progress in other domains, including language understanding and action planning. It must be noted, however, these networks still have important limitations in their ability to capture human generalization abilities (Sinz et al. 2019).

Further insight into the nature of the representations that are learned by HCNNs can be obtained using an *in silico* analogue to the neurophysiological recordings long used by neuroscientists to understand the receptive fields of neurons—essentially identifying the patterns that are learned by units at various levels of the network. Panel A of Fig. 3 shows examples of the representations learned by the units in an HCNN that are particularly responsive to different portions of a natural image (Olah et al. 2018). This particular network (GoogLeNet) was trained to recognize objects (including dogs and cats) based on a large set of training examples, and from that training it appears to have developed units that serve as "detectors" for features such as dog faces, cat faces, floppy ears, and furry legs—though these natural language descriptions necessarily fail to capture the fundamental function of those detectors, which are ultimately defined by the numeric values of the parameters in the network. It is nonetheless difficult to resist the natural tendency to view these putative representations as having interpretable

[1] https://www.clarifai.com/demo.

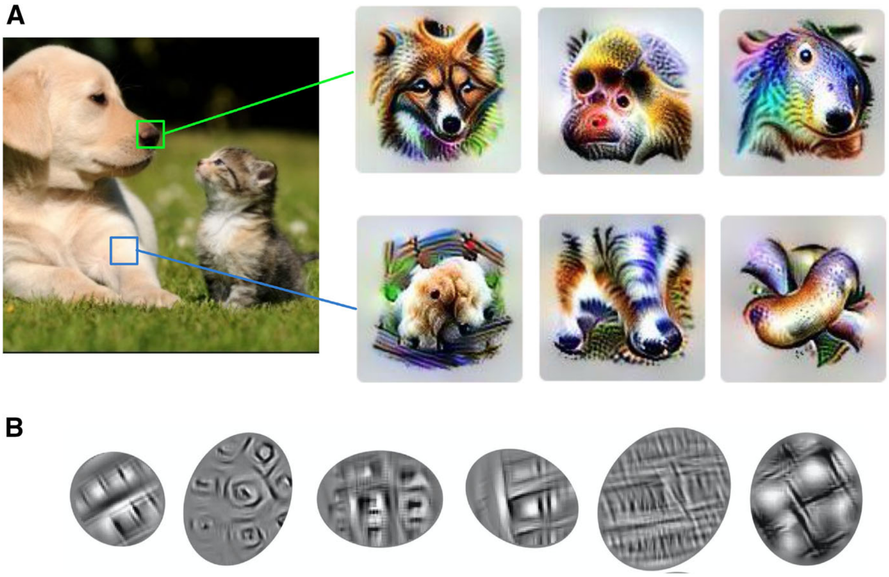

Fig. 3 Panel A: Examples of images that maximally stimulate units within a HCNN, reprinted under CC-BY from (Olah et al. 2018). Panel B: Examples of images that maximally stimulate individual temporal lobe neurons, generated by the deep image synthesis approach of (Bashivan et al. 2019). Reprinted with permission from AAAS

semantics falling at levels intermediate between the raw input and the final object category decision.

A striking aspect of the success of deep learning is how these systems are able to perform so well with so relatively few parameters. Imagine a very simple learning system that would simply record the correct category for any particular image in a lookup table; this would require at least as many parameters as there are possible images. Current deep learning systems have tens of millions of free parameters; this may seem quite high, but it is immensely smaller than the dimensionality of the space of possible images that are being recognized (cf. Fig. 1). To understand how this is possible, it is useful to view the job of a learning machine (such as a neural network or a biological organism) in terms of *function approximation*—that is, approximating the function that relates the state of the world to the most appropriate actions in a way that maximizes some objective (such as the long-run value of the outcomes of those actions). For example, the appropriate function approximation for the image of my cat in Fig. 1 given the task of describing the image might be: *f(image) = speak("That's Coco the cat, wearing a hat.")*. Artificial neural networks are known to be able to approximate any continuous function (under a set of assumptions and network features)(Hornik et al. 1989), though it is not guaranteed that an effective approximation can be learned in reasonable time. How is it that HCNN's are apparently so successful at learning effective approximations in such a relatively short amount of time?

An emerging answer to this question is that there is a set of inductive biases that are built into the architecture, which are well-matched to the underlying function that relates visual images of objects to their category membership. As discussed in Fig. 1, the space of possible images is very large, even when constraining the image to a relatively small pixel array represented as gray scale values. However, the space of possible images that could be produced by the world is exceedingly smaller than the space of all possible images, a large majority of which would simply look like “noise” to the human observer. This limitation arises from the physical laws of our universe. Lin et al. (2017) outlined a set of fundamental features of the laws that govern the evolution of physical systems in our world, which are low-dimensional, local, and invariant to various transforms (i.e. exhibiting physical symmetries). They argued that these constraints on the effective dimensionality of the observed world are the reason that both brains and HCNNs are able to effectively operate in the world despite having many fewer parameters than the dimensionality of the space of possible inputs. For example, the use of convolution over small image patches in artificial neural networks, and the relative small spatial receptive fields of the early visual cortex in primates, both leverage the locality of the physical world. They further argued that the hierarchical structure of the world (reflecting the compositional nature of the multiscale physical processes that give rise to the observed world) provides insight into the particular effectiveness of hierarchical neural networks (both natural and artificial) for solving these problems. Because the neural network composes a large number of relatively simple and encapsulated operations, the number of parameters in the model grows relatively slowly; as Lin et al. put it, this structure enables a “computational swindle” in which the number of required parameters grows linearly, rather than exponentially, with the number of relevant dimensions in the world.

Linking artificial and natural representations

As deep learning systems have become increasingly powerful, neuroscientists have begun to investigate the parallels between those models and the biological brains. This work has very clearly established the similarity in functional organization between artificial and natural neural networks; here I will focus on a set of studies by Dan Yamins and James DiCarlo on the primate visual system (Yamins and DiCarlo 2016). In their work, a large number of HCNNs (varying in the specific details of their architecture such as the number and organization of layers) are trained to perform an image recognition task that involves labeling a large number of images according to their object categories, and the best-performing model is selected; this is roughly akin to an evolutionary process, selecting for the architecture that best performs the task. The HCNN class of model architectures is loosely inspired by the structure of the primate visual system (LeCun et al. 2015), but no neural data are used in training the network on the task. Independently, a monkey is presented with the same images while activity is recorded from a large number of neurons in its visual system. The researchers tested whether the organization of the task-optimized neural network model is similar to neural activity in the monkey, defined as the degree to which activity of neurons in each brain region can be predicted using a simple (linear) readout of the activity patterns within each layer of the HCNN. The hierarchy of the visual system was replicated within the neural network, such that earlier visual regions were best predicted by earlier layers in the network, and later regions were best predicted by later layers in the network. The investigators further examined the degree to which the similarity structure of patterns of activity across the different images (i.e. their "representational similarity") was similar between the the monkey neural activity and the model's activity. In fact, the similarity structure of the neural network model was able to explain nearly all of the explainable variance in neural responses across stimuli in the monkey, demonstrating a remarkable match between the representations in these two systems. Similar results have been obtained using neuroimaging in humans, in both the visual and auditory systems (Kell et al. 2018; Khaligh-Razavi and Kriegeskorte 2014).

Above I discussed the nature of the internal representations that are learned by HCNNs, which were obtained through a process akin to *in silico* electrophysiology. Recently this approach for understanding the representations within an artificial neural network has been extended to better understand the representations within the primate brain. Bashivan et al. (2019) developed an approach called "deep image synthesis", which uses a closed-loop system involving a HCNN to generate patterns of visual stimulation that maximally activate individual neurons in the monkey's visual area V4. The model was able to generate stimuli (see Panel B of Fig. 3) that resulted in rates of neural firing in these neurons that were far greater than those evoked by any available natural image stimulus. Thus, the intermediate representations learned by an HCNN not only allow it to perform well at object recognition tasks, but are likely to be very similar to the representations present in biological neurons; otherwise it is highly unlikely that the images generated by maximizing the activity of those units would so strongly activate individual neurons in the primate brain.

Do neural representations meet the job description?

Despite the profligacy with which both neuroscientists and machine learning researchers use the term "representation", their widespread usage of the term does not necessarily legitimize its use in the philosophical sense. Rather, we must demonstrate that these posited representations fulfill the "job description" for doing real representational work (Ramsey 2007). There are, of course, nearly as many different conceptions of this job description as there are philosophers of mind. My approach here follows the outline of Shea's (2013) proposed pluralistic approach to the naturalization of representational content, which offers a relatively clear set of criteria for representational explanation:

> The strategy I am advocating is to examine a variety of representational explanations, and for each to identify:
>
> (a) An explanandum concerning how the system operates or behaves in relation to its environment.

(b) A putative explanation of (a) that relies in part on attributing representational properties to the system (e.g. keeping track of p, aiming at q, etc.).
(c) An account of how the explanation in (b) succeeds (remaining open to there being no such account).
(d) If there is a positive answer to (c), a characterisation of the kind of properties the representational properties of the system would have to be for the explanation in (b) to succeed in explaining (a) in accordance with the account (c).

Another way of phrasing Shea's criterion (d) was expressed by Ramsey (2007).

> Are there mindless systems in which an internal element is performing a role that is most naturally (or intuitively, or justifiably, or beneficially) viewed as representational in nature?

Neither of the these definitions is as precise as one might like, but they nonetheless provide a starting point for an assessment of the representations identified in both natural and artificial neural networks.

As an example, we can apply these criteria to the primate inferior temporal cortex in the context of its role in visual object recognition. The explanandum in this case is the behavioral ability to recognize and name a particular object given a particular pattern of visual stimulation (such as the image in Fig. 1). Theories of visual processing (which are implemented within the HCNN models described above) propose that the visual cortex builds increasingly complex representations of the visual world, with early regions representing relatively simple features such as edges, and later regions representing more complex features. Thus, the representational work performed by these regions is the hierarchical decomposition of the visual world in a way that allows the kind of flexible object recognition behavior observed in primates. As argued above, this decomposition is successful by virtue of its match to the structure of the world; in this case, the hierarchical convolutional architecture of the primate visual system that is mimicked in HCNN models reflects the compositional and hierarchical structure of the macroscopic physical world that gives rise to visual images. Evidence for the success of this account comes from the empirical success of HCNN models, both at performing object recognition tasks and at predicting the activity of neurons in the primate visual cortex, as well as the primate's ability to perform the very same task using similar representations. The state-of-the-art HCNN models can now approach human performance on a number of visual tasks. Perhaps, more surprisingly, HCNN models (which were trained to perform an object recognition task) can predict the activity of neurons in the primate visual cortex as well as, and sometimes *better than*, models that were trained on neural data (Cadena et al. 2019). HCNNs can also be used predictively to generate synthetic visual images that can drive activity in neurons that is greater than the activity evoked by any of a large number of natural visual images (Bashivan et al. 2019). These results demonstrate a significant degree of empirical success of the explanation of object recognition in terms of a hierarchy of representation.

Shea's criterion (d) encapsulates the "job description" question: What kinds of properties would our representations have to have in order for the theory to succeed?

First, these representations must be reliably triggered when the the relevant feature is present in the visual scene, reflecting a causal relation between the relevant feature in the world and its representation within the system. However, causal dependency is not sufficient to warrant representational status, as there are cases of causal dependency that at least some readers would not take as being truly representational. For example, beta-cells in the pancreas respond reliably to changes in the level of blood sugar, with electrophysiological responses that are striking similar to neurons in the brain, but few philosophers would accept the claim that they are "representing" the blood sugar level in the same way that brains represent the contents of thought; they are simply "receptors".

A stronger argument is based on the structural relationships between the content of representations in visual cortex and the structure of the visual world. The specific hierarchical architecture of both artificial HCNNs and the primate visual system gives rise to a set of representations that reflect the hierarchical and compositional structure of the visual world. Mathematically, one can view the work of these representations as projecting the high-dimensional visual input onto a low-dimensional manifold in which natural images live; there are many versions of the image in Fig. 1 that could be created by adding particular forms of noise, but which would be nonetheless perceived as the same object, because of this projection to a common location in the manifold. It is *only* through the lens of representation that the function of these systems, and their fundamental functional isomorphisms, make any sense at all.

Another important aspect of neural representations is the ability for them to perform their representational role in the absence of the triggering stimulus ("decouplability": Chemero 2009; Clark 1997). It has long been known that stimulation of specific sets of neurons could change an animal's behavior, and the recent advent of optogenetic technologies allowing the precise control of neural responses has provided striking demonstrations of the ability of patterns of neural activity to trigger relevant behaviors in the absence of the usual triggering stimulus. A particularly compelling demonstration of this comes from recent work that has used optogenetic stimulation to "implant" a visual percept in the visual cortex of a mouse (Marshel et al. 2019), causing the animal to behave as if the percept existed even in the face of no actual visual stimulation. Similarly, electrical stimulation of face-sensitive brain areas in humans can result in the apparent perception of illusory faces superimposed on the visual world (Schalk et al. 2017). Other work has shown the ability to reactivate specific memories by optogenetically activating specific sets of cells in the mouse hippocampus, evoking fear-related behaviors in a new context (Liu et al. 2012) It is difficult to explain this kind of behavior without reference to representations that are triggered by the exogenous neural stimulation, thus evoking the processes of visual perception or memory in the absence of direct causal influence via the natural pathway. One might question whether these are truly examples of decoupling, because they involve exogenous stimulation rather than endogenous processes. However, the fact that these externally imposed patterns of activity can result in appropriate behavioral outputs clearly satisfies the definition of decoupling outlined by Chemero (2009): "A representation R is decouplable just in case it can at least sometimes perform its function in a system when it is not in causal contact with its target T.“ The present discussion suggests a broader view of decoupling than is often discussed in the philosophical literature (though see

Thomson and Piccinini (2018)), extending to both endogenous (mentally generated) or exogenous (externally induced) patterns of activity that nonetheless serve to play a particular functional role in the absence of their natural cause.

An additional requirement often placed on representations is the ability to misrepresent (Dretske 1986). An interesting case of misrepresentation in artificial neural networks occurs in response to "adversarial examples", in which an image of a particular object is modified in a way that is imperceptible to humans but causes the network to reliably misclassify the object as a member of a different (incorrect) category (Szegedy et al. 2013); in the most common example, an imperceptible change to an image of a panda causes the network to misclassify the modified image as a gibbon. The presence of adversarial examples provides potentially useful insight into a question that arises in the discussion of misrepresentation—namely, the "disjunction problem" (Fodor 1987): is the content of the "gibbon" representation gibbons (in which case, the adversarial example causes a false tokening of the "gibbon" representation), or is it the disjunction "gibbons, or adversarial examples that cause the network to respond gibbon", in which case the network is not making (and by definition cannot make) an error. Our understanding of neural networks provides a computational view of misrepresentation: The error of incorrectly tokening the "gibbon" representation reflects cases in which an image of something other than a gibbon becomes embedded in the portion of the manifold associated with gibbons in the training data; as DiCarlo et al. (2012) put it, the manifolds become "tangled". We can sidestep the usual questions about functional indeterminacy that arise in the context of misrepresentation (e.g. Neander 1995), because in the case of the neural network we know exactly what function it was trained to perform—the objective function on which the network was trained, which in this case is minimizing classification error for images of visual objects. Thus, we can state with certainty that with respect to the trained objective of the model, tokening the "gibbon" representation to an image of a panda is clearly an error rather than a disjunction. This viewpoint provides a potentially useful mechanistic approach to questions of mispreresentation in the context of object recognition; further work will be necessary to determine whether this approach scales to more abstract mental concepts.

The necessity of representation for effective behavior

The insights gained from computational analysis of machine learning systems allow one to go beyond the claim that representations *do* exist, to argue that representations *must* exist in order for any organism (natural or artificial) to behave intelligently in the world. To establish this argument, I return to the notion of function approximation as the goal of an adaptive organism or system. A generic way to learn such an approximation would be through trial-and-error combined with memorization. The organism would simply try various possible actions in various states and update the value of each state-action pair according to the results, in essence performing exhaustive search and memorizing the most appropriate action for any particular state. For example, in the case of naming an object in an image, the system could simply try possible names for each particular image until it found the correct name, or remember the name that it was directly taught. The problem for any such *model-free* learner is that the space of state-action combinations grows exponentially with the number of variables that are involved. This is known as the "curse of dimensionality", first noted by Richard Bellman in the context of optimization for controlling a dynamical system. As the number of dimensions grows, the number of samples needed to cover the entire space of possible inputs grows exponentially. In the case of our photo, for a single pixel we would need 256 samples to entirely cover the space of possible values. For two pixels, we need 256^2 (or more than 65,000) samples; and so on. For toy problems like learning which of several slot machines to play in a casino, it is possible for a model-free machine learning system to accurately perform the task—but only if it is provided with the correct representation of the inputs and outputs by the researcher. For any problem in the real world, the number of dimensions of the state and action variables will ensure that one never has enough experience to accurately determine any more than a miniscule number of possible state-action values. The lack of any knowledge about the relationships between states also prevents the organism from generalizing from one state to another based on their similarity.[2]

A solution to the curse of dimensionality is to impose some structure on the function approximation, allowing the system to learn a smaller set of parameters than the exponentially-growing space of possible state-action values. This imposes an inductive bias on the system, such that system will generalize well to the degree that the bias imposed by the model structure accords with the true structure of the world. For example, take the image shown in Fig. 1. There is a large number of versions of this image that could be generated (e.g. by rotating the image 3 degrees and shifting it three pixels down) that would still be recognizable as my cat Coco wearing a hat; it is only by making some assumptions about the visual world (in this case regarding symmetries of visual images) that the primate visual system is able to easily recognize them all as the same cat. Recognizing the cat in this photo is also no problem for state-of-the-art image recognition tools that are based on artificial neural networks. The only way that an artificial neural network with millions of parameters can so effectively generalize its behavior to a new image with such a huge possible dimensionality is that it has learned a projection of the high-dimensional data in a lower-dimensional space in which natural images reside; for example, learning that object identity is invariant to transformations such as small rotations or translations in the image.

One might argue that the notion of dimensionality is relative—that is, that there is no "natural" scheme under which to quantify the dimensionality of a particular stimulus. While this is true in principle, in reality any sensory scheme that is going to be successful in the world must necessarily have inputs of high enough dimensionality to invoke concerns about the curse of dimensionality. For example, the human retina has well over 100 million photoreceptors; but even for the lowly fruit fly *Drosophila melanogaster*, which has about 5600 photoreceptors (Hardie 1985), the dimensionality of the resulting signal space requires a structured representation of the visual world in order to successfully skirt the curse of dimensionality. It is also important to note that this representation could arise either through evolutionary effects on the nature

[2] This is not to say that organisms don't sometimes use model-free learning. In fact, there is reason to think that some habits may rely on exactly this kind of learning (Dolan and Dayan 2013).

of the neural architecture (as seen in the structure of the primate visual system) or from plasticity mechanisms that adapt to the world. For example, animals deprived of particular forms of visual input (such as horizontal lines) will fail to develop neural representations of those features (Blakemore and Cooper 1970); whereas evolution has provided the architecture and learning mechanisms that give rise to structured representations, those representations develop by virtue of an interaction between these mechanisms and the organism's experience.

The foregoing arguments are meant to establish the metaphysical status of representations; their epistemic utility is another question (cf. Chemero 2009). Within the visual system, the representational approach has been highly successful (as argued above), whereas in the study of the motor system, dynamical systems approaches have led to greater insights (Shenoy et al. 2013). Increasingly the representationalist and dynamical systems approaches are being brought to bear in conjunction. A prime example was seen in a study by Mante et al. (2013), which examined how the prefrontal cortex switches its decision based on context. Monkeys were trained to perform either a motion detection or color detection task, based on a contextual cue, while responses from individual neurons in the prefrontal cortex were recorded. The responses of individual neurons in the prefrontal cortex reflected many different aspects of the task, but an analysis of the low-dimensional dynamics of the entire population of neurons provided substantially greater insight into how the region coded information about the task, with different dimensions in the state space coding for different aspects of the task. The authors conclude: "In light of our results, these mixtures of signals can be interpreted as separable representations at the level of the neural population. A fundamental function of PFC may be to generate such separable representations, and to flexibly link them through appropriate recurrent dynamics to generate the desired behavioural outputs." This fluid combination of representationalist and dynamicist thinking highlights the degree to which conceptual dichotomies may be useful for advancing the careers of theorists but are generally abandoned in favor of synthetic approaches in practice.

Conclusion

The concept of representation is fundamental to the explanatory practices of neuroscientists, and increasingly important to machine learning researchers as well. I have argued here that a particular type of representation, referring to patterns of activity that bear a systematic relationship to the structure of the external world and play a causal role in behavior, is fundamentally necessary for any intelligent organism. These representations provide the organism with an inductive bias that is matched to the structure of the world, without which the organism will be quickly overcome by the curse of dimensionality. More generally, the arguments here provide a proof of concept for the utility of both empirical and computational results from neuroscience in the analysis of fundamental philosophical questions.

Acknowledgements Thanks to Tyler Bonnen, Rosa Cao, Felipe De Brigard, Dawn Finzi, Inman Harvey, Grace Huckins, Jay McClelland, Gualtiero Piccinini, David Sussillo, Jessey Wright, Dan Yamins, the

Stanford Philosophy of Neuroscience Journal Club, and the Neural Mechanisms Online community for helpful comments and discussion.

References

Albright, T. D., Desimone, R., & Gross, C. G. (1984). Columnar organization of directionally selective cells in visual area mt of the macaque. *Journal of Neurophysiology*, *51*(1), 16–31. https://doi.org/10.1152/jn.1984.51.1.16.

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*,. https://doi.org/10.1126/science.aav9436.

Blakemore, C., & Cooper, G. F. (1970). Development of the brain depends on the visual environment. *Nature*, *228*(5270), 477–8. https://doi.org/10.1038/228477a0.

Buckner, C. (2018). Empiricism without magic: Transformational abstraction in deep convolutional neural networks. *Synthese*,. https://doi.org/10.1007/s11229-018-01949-1.

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., et al. (2019). Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS Computational Biology*, *15*(4), e1006897. https://doi.org/10.1371/journal.pcbi.1006897.

Chemero, A. (2009). *Radical embodied cognitive science*. Cambridge: MIT Press.

Clark, A. (1997). *Being there: Putting brain, body, and world together again*. Cambridge: MIT Press.

Davis, T., & Poldrack, R. A. (2013). Measuring neural representations with fmri: Practices and pitfalls. *Annals of the New York Academy of Sciences*, *1296*, 108–34. https://doi.org/10.1111/nyas.12156.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *CVPR09*.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–34. https://doi.org/10.1016/j.neuron.2012.01.010.

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–25. https://doi.org/10.1016/j.neuron.2013.09.007.

Dretske, F. (1986). Misrepresentation. In R. Bogdan (Ed.), *Belief: Form, content, and function* (pp. 17–36). Oxford: Oxford University Press.

Engelhard, B., Finkelstein, J., Cox, J., Fleming, W., Jang, H. J., Ornelas, S., et al. (2019). Specialized coding of sensory, motor and cognitive variables in vta dopamine neurons. *Nature*, *570*(7762), 509–513. https://doi.org/10.1038/s41586-019-1261-9.

Fodor, J. A. (1987). *Psychosemantics*. Cambridge: MIT Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1–2), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5.

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., & Gore, J. C. (1999). Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nature Neuroscience*, *2*(6), 568–73. https://doi.org/10.1038/9224.

Goddard, E., Klein, C., Solomon, S. G., Hogendoorn, H., & Carlson, T. A. (2018). Interpreting the dimensions of neural feature representations revealed by dimensionality reduction. *Neuroimage*, *180*(Pt A), 41–67. https://doi.org/10.1016/j.neuroimage.2017.06.068.

Goldberg, M. E., Bisley, J. W., Powell, K. D., & Gottlieb, J. (2006). Saccades, salience and attention: The role of the lateral intraparietal area in visual behavior. *Progress in Brain Research*, *155*, 157–75. https://doi.org/10.1016/S0079-6123(06)55010-1.

Goldman-Rakic, P. S. (1995). Cellular basis of working memory. *Neuron*, *14*(3), 477–85. https://doi.org/10.1016/0896-6273(95)90304-6.

Hardie, R. C. (1985). Functional organization of the fly retina. In D. Ottoson (Ed.), *Progress in sensory physiology* (Vol. 5, pp. 1–79). Berlin: Springer.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425–30. https://doi.org/10.1126/science.1063736.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, *17*(11), 4302–11.

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, *98*(3), 630–644.e16. https://doi.org/10.1016/j.neuron.2018.03.044.

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology*, *10*(11), e1003915. https://doi.org/10.1371/journal.pcbi.1003915.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., et al. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–41. https://doi.org/10.1016/j.neuron.2008.10.043.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–44. https://doi.org/10.1038/nature14539.

Lin, H. W., Tegmark, M., & Rolnick, D. (2017). Why does deep and cheap learning work so well? *Journal of Statistical Physics*, *168*(6), 1223–1247. https://doi.org/10.1007/s10955-017-1836-5.

Liu, X., Ramirez, S., Pang, P. T., Puryear, C. B., Govindarajan, A., Deisseroth, K., et al. (2012). Optogenetic stimulation of a hippocampal engram activates fear memory recall. *Nature*, *484*(7394), 381–5. https://doi.org/10.1038/nature11028.

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, *503*(7474), 78–84. https://doi.org/10.1038/nature12742.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York: Henry Holt and Co. Inc.

Marshel, J. H., Kim, Y. S., Machado, T. A., Quirin, S., Benson, B., Kadmon, J., et al. (2019). Cortical layer-specific critical dynamics triggering perception. *Science*,. https://doi.org/10.1126/science.aaw5202.

Neander, K. (1995). Misrepresenting and malfunctioning. *Philosophical Studies*, *79*(2), 109–41. https://doi.org/10.1007/BF00989706.

O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, *34*(1), 171–175. https://doi.org/10.1016/0006-8993(71)90358-1.

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., et al. (2018). The building blocks of interpretability. *Distill*,. https://doi.org/10.23915/distill.00010.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1–2), 73–193. https://doi.org/10.1016/0010-0277(88)90032-7.

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

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., et al. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*(11), 1761–1770. https://doi.org/10.1038/s41593-019-0520-2.

Schalk, G., Kapeller, C., Guger, C., Ogawa, H., Hiroshima, S., Lafer-Sousa, R., et al. (2017). Facephenes and rainbows: Causal evidence for functional and anatomical specificity of face and color processing in the human brain. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(46), 12285–12290. https://doi.org/10.1073/pnas.1713447114.

Shea, N. (2013). Naturalising representational content. *Philosophy Compass*, *8*(5), 496–509. https://doi.org/10.1111/phc3.12033.

Shenoy, K. V., Sahani, M., & Churchland, M. M. (2013). Cortical control of arm movements: A dynamical systems perspective. *Annual Review of Neuroscience*, *36*, 337–59. https://doi.org/10.1146/annurev-neuro-062111-150509.

Sinz, F. H., Pitkow, X., Reimer, J., Bethge, M., & Tolias, A. S. (2019). Engineering a less artificial intelligence. *Neuron*, *103*(6), 967–979. https://doi.org/10.1016/j.neuron.2019.08.034.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, *11*(1), 1–23. https://doi.org/10.1017/S0140525X00052432.

Szegedy, .C, Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R. (2013). Intriguing properties of neural networks. arXiv e-prints arXiv:1312.6199.

Thomson, E., & Piccinini, G. (2018). Neural representations observed. *Minds and Machines*, *28*(1), 191–235. https://doi.org/10.1007/s11023-018-9459-4.

Vilarroya, O. (2017). Neural representation: A survey-based analysis of the notion. *Frontiers in Psychology*, *8*, 1458. https://doi.org/10.3389/fpsyg.2017.01458.

Wilson, S., & Moore, C. (2015). S1 somatotopic maps. *Scholarpedia*, *10*(4), 8574.

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–65. https://doi.org/10.1038/nn.4244.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.