Black Boxes and Theory Deserts:
Deep Networks and Epistemic Opacity in the Cognitive Sciences

Frank Faries[1] and Vicente Raja[2*]

[1]Department of Philosophy, University of Cincinnati (USA)
[2]Rotman Institute of Philosophy, Western University (Canada)

[*]Corresponding Author: Vicente Raja – vgalian@uwo.ca
Postdoctoral Fellow at Rotman Institute of Philosophy
Western University
1151 Richmond Street North
London, Ontario, Canada, N6A 5B7

Abstract

Cognitive scientists deal with technology in a very particular way: they use technology to understand perception, action, and cognition. This particular form of human-machine interaction (HMI) is very well illustrated by the use cognitive scientists make of artificial neural networks as models of cognitive systems and, more concretely, of the brain. However, the activity of cognitive scientists in this context suffers from the shortcoming of epistemic opacity: artificial neural networks are too difficult to interpret and understand, so in many cases they remain black boxes for researchers. In this paper, we provide a diagnostic for such epistemic opacity based on dominant cognitive science's lack of theoretical resources to account for the activity of artificial neural networks when taken as models of the brain. Then, we offer the guidelines of a solution founded on the notion of information developed in ecological psychology.

Keywords: ecological information; human-machine interactions; machine learning; cognitive science

1. Introduction

Human-machine interactions (HMI) in cognitive science can be studied from two different points of view. Cognitive scientists can study HMI themselves. How do people engage with new technologies and the different opportunities offered by them? These studies encompass, among many other topics, the way we engage with virtual and augmented reality devices (Raja & Calvo, 2017; Schettler et al., 2019), the use of sensory substitution devices (Favela, Riley, Shockley, & Chemero, 2018; Lobo et al., 2018), or the undesirable socio-political effects of using biased algorithms in machine learning and deep learning (Birhane & Guest, 2021).

A different way to look at HMI in cognitive science is to focus on how cognitive scientists engage with technologies to gather new knowledge and new understanding of

human and animal psychology. A contemporary example of this form of HMI has to do with neural networks. Several developments in machine learning and deep learning (Goodfellow et al., 2016; LeCun, et al., 2015) have made (deep) neural networks a relevant branch of research in the cognitive sciences. These neural networks are able to either reach or even surpass human performance in tasks such as image classification (He et al., 2015), translation (Wu et al., 2016), and various games (Mnih et al., 2015). However successful neural networks are, several voices have raised in the last decade regarding their epistemic opacity (see Humphreys, 2004, 2009; Parker, 2013; Winsberg, 2001, 2010; Stuart & Nersessian, 2019). The problem, so the story goes, is that although cognitive scientists are able to build up and train them, deep neural networks eventually become black boxes that provide neither explanation nor understanding due to the number of parameters and degrees of freedom they have (Hasson et al. 2020). This issue stands by itself as a problem different from others previously described in the research on neural networks—e.g., the famous systematicity challenge Jerry Fodor and Zenon Pylyshyn posited to connectionism—and is the motivation for the move toward so-called Explainable AI (XAI) and the issue of trust in artificial intelligence (cf. Doshi-Velez & Kim, 2017; Duran & Formanek, 2018; Lipton, 2018; Murdoch et al., 2019). This is the kind of HMI we will explore in this paper.

In the following sections, we are going to evaluate the reasons for the epistemic opacity of deep neural networks and will try to provide a plausible solution to it. In section 2, our thesis is that HMIs involving deep neural networks in cognitive science are epistemically opaque not because those networks intrinsically are black boxes but because they either (i) work under the wrong theory or (ii) are used just as a substitute for the theoretical needs of cognitive science. In both cases, the lack of a proper theoretical structure is the reason why deep neural networks appear as black boxes when trying to explain perception, action, and cognition. In section 3, after the proposed diagnostic, we will provide several examples of different efforts made by researchers in machine/deep learning to make this technology less opaque. Finally, in section 4, we propose a radical departure from the dominant theoretical framework in the cognitive sciences in order to find a theoretical environment that helps the interpretability, explainability, and understanding of these networks and their role as models of cognitive systems. Such a new theoretical environment is based on ecological psychology and pivots on the notion of ecological information.

2. A Theory Desert for Many Black Boxes

The problem with the epistemic opacity of deep neural networks has to do with the inability faced by cognitive scientists to use those networks to gather new information, knowledge, or understanding of the cognitive processes they are supposed to be modelling. The source of this inability has been described from the point of view of both the complexity of deep neural networks as models (Castelvecchi, 2016) and the lack of connection between these networks and real-world phenomena (Sullivan, 2020). We think that these criticisms, although plural, fall within a broader category: the fact that deep neural networks have evolved within a theoretical environment that lacks the proper resources to make them useful explanatory tools for cognitive science. Such a theoretical environment is twofold. On the one hand, deep neural networks have evolved in an engineering environment that does not need to be directly related to or concerned with perception, action, and cognition. In this

sense, deep neural networks are understood like tools that can be used to solve different engineering problems (e.g., image recognition). As Goodfellow et al. (2016) put it:

> While it is true that deep learning researchers are more likely to cite the brain as an influence than researchers working in other machine learning fields, such as kernel machines or Bayesian statistics, one should not view deep learning as an attempt to simulate the brain. [...] While some deep learning researchers cite neuroscience as an important source of inspiration, others are not concerned with neuroscience at all. (p. 15-16).

On the other hand, among those researchers of deep neural networks that are indeed concerned with neuroscience, the theoretical environment is almost exclusively based on the computational-cum-representational theory of cognition (Fodor, 1975, 1981; Pylyshyn, 1984; for a recent treatment, see Kriegeskorte & Douglas, 2018). Such a theoretical environment requires specific computations and concrete representations to explain cognition. In this context, deep neural networks become a form of explanatory black box as their features (e.g., dimensions, complexity, etc.) make researchers unable to put them in clear relationship to those needed computations and representations. Let's take a closer look at this issue.

It is important to realize that deep neural networks used in deep learning systems are *just* artificial neural networks. Artificial neural networks are *deep* when they have many layers between their input and their output. Such depth became common currency in the late 2000s when the renaissance of artificial neural networks in such a deep fashion started leading to dramatic improvements in machine learning (Hinton et al., 2006; Hinton & Salakhutdinov, 2006). However, the neural network architecture on which it is built was devised over half a century ago (McCulloch & Pitts, 1943) and the main algorithms (e.g., backpropagation and stochastic gradient descent) were applied in the field during the 1980s (Werbos, 1982; LeCun, 1988; Rumelhart, Hinton, & Williams, 1986). Indeed, although the basic technology has improved, it has not substantially changed much since then and the recent advances in deep learning and, more generally, in machine learning can be attributed to data availability and computational brute force (Goodfellow et al., 2016).

Generally speaking, artificial neural networks are composed of nodes and edges. The nodes are the computing components of the network and they are vaguely inspired by neurons. The edges are the connections between nodes. Artificial neural networks are input-output systems in which information usually runs from an input layer of nodes and towards an output layer of nodes. Between input and output layers usually we find one or more *hidden* layers of nodes. The nodes of each layer are connected to the nodes of the other layers through edges. The strength of these connections—i.e., the influence the activity of a node has in the activity of the nodes of the next layer—depend on the *weights* of those edges: the greater the weight of the edge, the stronger the connection between nodes (see Figure 1).

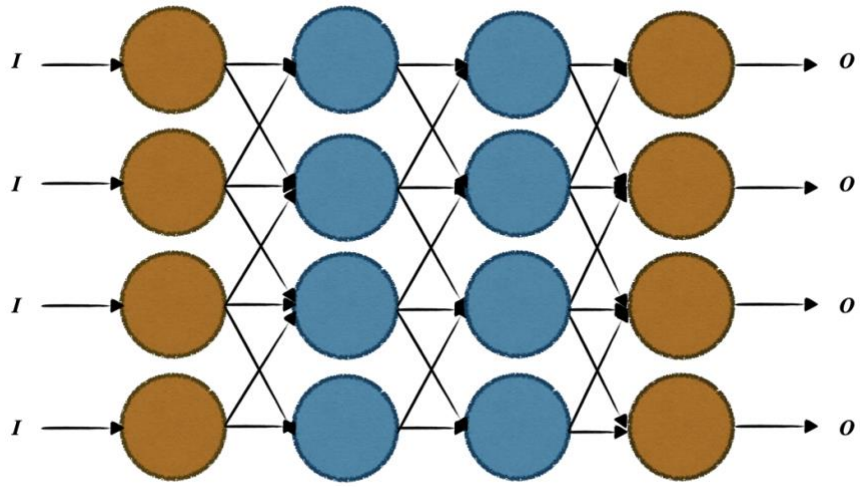

Figure 1. Schema of an artificial neural network. Colored circles are *nodes* and black arrows are *edges* connecting nodes. Information flows from left to right (the direction of the arrows) or, in other words, from the input (***I***) to the output (***O***)—this is the reason why these networks are sometimes called feed-forward neural networks. The set of four brown nodes to the left is the *input layer*. The two central sets of four blue nodes are two *hidden layers*. The set of four brown nodes to the right is the *output layer*.

In artificial neural networks, desired input-output relationships could be probed by adjusting these weights like so many knobs and dials. Of note is the view of information processing implicit in this model. As we have already noted, information flows from inputs to outputs (left to right in Figure 1). Inputs are usually understood as discrete features represented as a range of values—e.g., the features of a living thing (e.g., that it has two legs, walks upright, has no tail, and so on)—and they can come either from outside the networks, as in the case of the input layer, or from the previous layers of the network, as in the case of the hidden and the output layers. In the latter case, it becomes more and more difficult to say which features or combinations of features are represented in each layer. Usually, processing in each node occurs by taking the weighted sum of the values of those inputs, and giving an output based on that sum. Then either this output is an input for the next layer or the activation value of that node belongs to the output layer. For instance, suppose an artificial neural network gets a picture as an input and has to determine whether there is a human being in the picture. To do so, one possible structure of the input, as we have just noted, is that each node of the input layer gets one feature of a human being. Then, each node sends that feature as output to the nodes of the next layer where individual features can, for instance, be combined. As the processing goes on layer by layer, the eventual result is having a pattern of activation in the output layer that reflects the fact that the input was actually a picture with a human being in it (or not, depending on the input).

In this simple example, we have offered a rather shallow understanding of what is going on in artificial neural networks when they are presented with a task such as recognizing a human being in a picture. Providing the actual details of this process is outside of the scope of this paper, but we want to point out some subtleties. Broadly speaking,

artificial neural networks are no more and no less than universal function approximators. And it can be shown that, granting a few restrictions, artificial neural networks can approximate any computable function. In the case of our example, the artificial neural network just needs to approximate a function that takes pictures as inputs and then delivers labels like "human being" or "not human being" as outputs. We have chosen to depict this process in terms of feature detection and feature combination, but that is just one of the possible ways to understand the activity of the network. Other possibilities are, for instance, to take the artificial neural networks to be making some kind of inference or to be learning some representation—distributed or not—of the process that generated the inputs or both. These and other possibilities, along with different algorithms to implement them, can be found in the literature (for an exhaustive review of the field, see Goodfellow et al., 2016). That said, a general understanding of the architecture of artificial neural networks and their activity as function approximators is enough to keep with the rest of the paper.

This succinct review of artificial neural networks promptly highlights that the symbiosis between the prevailing science of the brain, as composed of neurons and their connections, and logic/mathematics, both in the application of functions and the (early!) application of network theory, was made apparent from the start.[1] Subsequent synergies with computer science have developed along these lines such that improved storage and processing power now make feasible achievements foregone in the periods of reduced funding and interest in AI during the 1970s, '80s, and '90s, known as the "AI winter" (McDermott et al., 1985; Russell & Norvig, 2003, p. 22; Howe, 1994; see also Lighthill, 1973). Current deep neural networks now have many more nodes, and, crucially, additional hidden layers, but the very foundations of the system remain pretty much the same since its beginnings.

With the addition of depth—i.e., with the addition of hidden layers—came profound computational benefits. The notion of a hidden layer was first introduced by Minsky and Papert (1969), who, despite having shown that an additional layer made computing logical relationships like the exclusive or (XOR) possible, pessimistically concluded "We believe that [networks with hidden layers] can do little more than can a low order perceptron" (Minsky & Papert, 1972, p. 38). In fact, it is this depth, and its subsequent benefits, that is the hallmark of deep neural networks.

The "deep" in deep learning refers to additional processing layers of deep neural networks. As noted by Cameron Buckner in a recent article, "deeper networks can solve certain types of classification and decision problems exponentially more efficiently than shallower networks" (Buckner, 2019, p. 1). He goes on to illustrate this using sum-product networks as an example. As Delallau and Bengio (2011) have formally proven, in a network composed of only two kinds of nodes—one which returns a weighted sum of its inputs, and the other which returns a product—used to solve polynomial functions, deeper architectures capitalize on factorization to increase computational efficiency:

[1] The background of both developers of the first models of artificial neural networks speaks to this fact. Warren McCulloch himself was a neurophysiologist, and Walter Pitts a mathematician. That their combined insights yielded a logical electrical circuit is emblematic of the intertwining of neurobiology with the engineering and computational technologies used to study neurobiology that was the foundation of cybernetics, and later cognitive science.

> Specifically, the number of times an input product composed at an earlier layer can be reused in more complex products built by later layers increases exponentially with the network's depth…[Thus,] functions that can be efficiently represented as redeploying simpler computations to hierarchically compose more complex computations—can be represented and computed exponentially more efficiently in a deep architecture than a shallow one (Buckner, 2019, p. 4).

This depth, in combination with the exponential increase in data availability and some improvements in the array of algorithms at different levels of the network, accounts for the computational power of deep networks, and the diversity of domains in which these models find application. Put plainly, if a question can be posed in the form of a solvable function, and the factors required for solving it can be measured or recorded as data, then a deep network with access to that data can find the solution, often very quickly, and increasingly free of human direct intervention. This is the reason why deep neural networks are said to be used in relatively theory-neutral or atheoretical ways. In cases like disaster prevention or marketing, any system that can quickly provide accurate predictions, say, of the likelihood of forest fires (Safi & Bouromi, 2013) or the designation of a potential customer (Bahari & Elayidom, 2015), will be valued at the expense of any kind of theoretical transparency. In these contexts, if the deep neural network is able to provide a good result, whether we understand the way the network has arrived at that result or not is completely secondary, if not irrelevant.

By contrast, other fields where deep neural networks are applied, and scientific inquiry in a more general sense, seem to require a theoretically-loaded endeavor. And this is the case in cognitive science, even though much of the theoretical-neutrality rhetoric from what we might call "metaphysically unfettered" efforts seems to be making its way into it. As noted in a recent article by Perconti & Plebe (2020), more and more deep neural networks are used as research tools in the cognitive sciences. We find that, at the same time, more and more complaints are emerging regarding their epistemic opacity (e.g., Stuart & Nersessian, 2019; Sullivan, 2020). Put simply, the problem is that, unlike in some other fields, if deep neural networks are used in cognitive science as models of brains or cognitive systems in general, the way the networks arrive at their results is an important part of the *explanandum* researchers are pursuing to explain. This *explanandum* is indeed difficult to explain in the case of deep neural networks. These networks just have too many interacting nodes, edges, weights, parameters, inputs, etc. To provide a comprehensive explanation of all these components and interactions is not feasible with the typical tools researchers in the cognitive sciences have at their disposal. Concretely, it seems impossible to find the proper relationships between these various components and interactions and the central notions of the dominant theory in cognitive science: *computation* and *representation*. What kind of computation is instantiated by these networks if any? Are different groups of nodes instantiating different steps of the computational algorithm? If so, which group of nodes is doing what? What is the network representing if anything? These are kinds of questions that seem to have no possible answer when deep neural networks are used in the cognitive sciences as models of cognitive systems. For this reason, deep neural networks become *black boxes*: elements that play a role in the explanation of a given cognitive ability or effect but that remain themselves unexplained.

One important aspect of the role of deep neural networks as black boxes in cognitive science is that they not only become black boxes for the lack of theoretical tools to explain them, but that they are actually *embraced* as black boxes by many cognitive scientists. The result in these cases, as we see it, is a *theory desert*: researchers in cognitive science are forced to choose between embracing a powerful engineering device with no regard for a theory backing it up, or attempting to apply a computational-cum-representaional theory that has no tools to make the device epistemically transparent. In other words, researchers find themselves in a desert in which they have no theory at all to fight against the opacity of deep networks. However, we think there is no reason to assume that these networks are intrinsically inert for explanatory/understanding purposes. On the contrary, it is our contention that the problem is that the current dominant computational-cum-representational theory of cognition has no resources to make them useful. Weighed down by finding herself in a desert with inappropriate or outdated theoretical machinery, or finding herself with no theory at all, the deep learning modeler may be tempted to abandon cognitive science, perhaps taking their talents to another more fruitful area. Rather than suffer the mass exodus of talented researchers to other fields, a potential solution to this issue is to explore the theoretical space for suitable alternatives. As we will go on to show, a viable theoretical alternative already exists which is better suited to meet the task of overcoming epistemic opacity in deep neural networks.

3. Removing Black Boxes?

There are at least two different points of view from which one may address the issue that deep neural networks have effectively become an epistemologically opaque black box in scientific explanations. The first of these points of view is quite general and has to do with different ways in which deep learning and machine learning modellers—not only those with ties with the cognitive sciences but all of them—try to make the networks they work with amenable to a better understanding. This point of view is based on techniques that allow for *interpreting* the algorithms used in deep neural networks and their results in a deeper way. The second point of view is specific to the cognitive sciences and has to do with finding relationships between deep neural network activities and brain activities in order to shed light on both of them.[2] The hope of this approach is that both systems are sufficiently similar such that insights about one system can be applied to the other system. Before we put this approach into question, let's say a few more words regarding both points of view.

Within the general research on deep learning, one example of the efforts made to better understand the activities of deep neural networks is the growing literature on their interpretability (see Chakraborty et al., 2017). Although achieving adequate task performance is a central aim of deep learning systems, it is far from being the only one of a handful of important concerns regarding their activity. Other important concerns might include safety (Pereira & Thomas, 2020), nondiscrimination (Birhane & Guest, 2021), or privacy (De Cristofaro, 2020). With respect to these other criteria, some of them difficult to quantify, interpretability commonly acts as a proxy. As Doshi-Velez and Kim (2017) put it, "if the system can *explain* its reasoning, we then can verify whether that reasoning is sound

[2] We say this is a point of view specific to the cognitive sciences because that's our focus. In principle, the same approach can be taken from any scientific field.

with respect to these auxiliary criteria" (p. 1; emphasis in the original). Interpretable deep neural networks, on this view, provide human-readable justification for their outputs or, at least, researchers have some (hopefully) reliable resources, models, and meta-models to interpret those deep neural networks (for a review, see Linardatos et al., 2020). In other words, interpretable deep neural networks are those whose activity can be understood by researchers under some circumstances. The information researchers can gather from these deep neural networks is, both theoretically and methodologically, the key for their interpretability.

In many cases, researchers explicitly use the technical notion of information from information theory. For instance, many of the cost functions—i.e., the functions that measure the performance—of unsupervised deep learning systems make explicit use of Shannon information (Shannon, 1948) to estimate the maximum likelihood of the parameters of the network. This use of information theory provides some insights regarding the activities deep learning systems are performing: generally speaking, deep neural networks are approximating a desired function by minimizing some other function (the cost function) and that minimization is cashed out in terms related to uncertainty and Shannon information. Such an interpretation of the activities of deep neural networks is, however, indirect. The researchers only have access to a proxy of the performance to interpret the activities of the network. They know it is doing better or worse in the task of interest, but they don't know *how* it is doing it. In other words, having access to a performance index based on Shannon information does not tell researchers anything regarding how the deep neural network is actually carrying out the task. In this sense, this technical use of information theory does not allow us to see into the black boxes of deep neural networks.

In other cases, however, researchers can gain interpretational grip on deep neural networks by different means. Some of these means consist of explicit interpretability tools, as for instance, DeepExplain (Zeiler & Fergus, 2014), Grad-CAM (Selvaraju et al., 2017), or InterpretML (Ribeiro, Singh, & Guestrin, 2016). All of them are interesting ways and toolboxes with which researchers target diverse interpretability issues of deep neural networks by developing tailored algorithms and visualization methods. However, there are more fundamental approaches to these issues. These approaches do not consist of ad-hoc interpretability methods but take a more basic route associated with the architectural and functional properties of deep neural networks. One of these approaches is, for instance, to constrain the features and activities of deep neural networks to get informative outcomes—in this case, the notion of information is used in a non-technical, common language way. One of these constraints is known as *sparsity*. Signals of interest are, generally speaking, highly structured. Some deep learning and other machine learning models exploit this structure to find a low-dimensional model of the data; that is, a model that captures the different aspects of the data with just a set of few parameters that can be cast in terms of rules in a classifier (Friedman & Propescu, 2008; Letham, et al., 2015), the pairwise interactions in a matrix (Lou, et al, 2013; Caruana, et al., 2015), or the number of coefficients in an algorithm (Ribiero, et al., 2016; Kindermans, et al., 2018), for instance. In other words, a sparse model compresses a high-dimensional feature space of a dataset of inputs into a few predictively/performatively adequate outputs.

The sparsity of some deep neural networks permits improving their interpretability insofar as the high-dimensionality of their input is reduced in their output to an extent that can be understood by researchers. For instance, if a data set of million pictures is encoded

by a deep neural network in a handful of factors that account for the variability of all the features of those pictures, then that handful of factors can be used to generate new instances of the dataset—this is what variational autoencoders, for instance, can do (Kingma & Welling, 2014, 2019). But then, researchers have more information about that generative process: they know that the deep neural network is generating new instances of the data set from a specific set of factors and not another one. In this sense, some information about the activity of the deep neural network is gathered by the sparsity constraint. However, the concrete steps of the encoding or the generating activities performed by the deep neural network, although more constrained in their possibilities, remain completely unknown. Therefore, the network remains a black box regarding its concrete activities for the most part.

Finally, deep learning systems are sometimes interpreted through their application, either in real world cases or sufficiently similar simpler instances. Sometimes this involves expert assessment of the system's outputs. For instance, an algorithm for correcting segmentations in microscopy data could be evaluated by expert review of the same target image task (Suissa-Peleg et al., 2016). In other cases, system outputs can be checked against competing explanations for quality of explanation type (Kim, et al. 2015). As in the case of the previous efforts on the interpretability of the activities of deep neural networks, using this kind of post-hoc assessment regarding the performance of the network is not providing any further information regarding its concrete activities to achieve that performance level. Once again, deep neural networks remain epistemically opaque.

The instances of interpretability just reviewed apply to deep neural networks writ large. Namely, these are strategies researchers can follow irrespectively of their interest in neuroscience. Beyond them, however, there are some efforts within neuroscience to remove the black boxes of their explanations that make use of these networks. These efforts have to do more directly with theory building and theory refinement. This is the case, for instance, of recent attempts to link the features of artificial and natural neural networks (Bashivan et al., 2019; Buckner, 2018; Poldrack, 2020; Yamins & DiCarlo, 2016). These views take, on the one hand, the deep structure of successful artificial neural networks, and, on the other, the organization of neurons in the cellular networks of the brain as combined evidence that successful implementations of the former stand as accurate cognitive models of the latter. For instance, Cameron Buckner (2018) concludes, "convolutional and pooling nodes correspond to simple and complex cells in the mammalian neocortex, which are organized in hierarchical layers as depicted in the layers of a DCNN [Deep Convolutional Neural Network]. Thus DCNNs provide a mechanistic model of abstract categorization and perceptual similarity judgments in mammals" (p. 5364). In other words, at least according to some researchers in the field, the activities and outputs of a black-box neural network can be interpreted through the application of mechanistic theory to its features.

This is a straightforward application of the model-to-mechanism mapping (3M) requirement held by various neo-mechanist philosophers (see Kaplan & Craver, 2011; also Craver & Kaplan, 2020; Craver, 2007). This requirement states that the features in a model—the neural network in this case—are explanatory to the extent that they can be mapped to real entities, activities, or organizational features of a mechanism. This mapping occurs by comparing a model's structure against the structure of the target system. Using an example strangely beloved by philosophers of science, a model of a toilet will be explanatory to the extent that the model identifies the component parts of the toilet (e.g., handle, valve, tank,

etc.) and their orchestrated activities (e.g., pulling, lifting, filling, etc.). Likewise, a model of similarity judgments in mammals will be explanatory, not by how often the model's judgment of similarity matches a mammal's, but to the extent that the model's components can be mapped to the entities and activities, properly organized, which actually account for similarity judgments in mammals.

There is an interesting aspect of this way of proceeding in the case of deep neural networks and their status as black boxes. The mentioned use of DCNNs in the cognitive sciences becomes an application of mechanistic theory of cognition. In this sense, properties of DCNNs—convolution or max-pooling, for instance—rather than simply being opaque mathematical procedures when they are at play in a system as complex as a deep neural network, become *concrete* ways of modeling the activities of neurons in the brain. In other words, the very abstract activities of the DCNNs are mapped to specific functions of biological neurons and neural networks. It is by this process, for instance, DCNNs are claimed to be able to implement a hierarchical processing which Buckner (2018) dubs "transformational abstraction". The process, so the story goes, "iteratively converts sensory-based representations of category exemplars into new formats that are increasingly tolerant to 'nuisance variation' in input" (Buckner, 2018, p. 5339). It is this process and its underlying architecture, Buckner claims, that accounts for visual similarity judgements in the mammalian brain. However, it is important to note again that this use of the mechanistic theory of cognition just maps aspects of a mechanistic description of the mammalian brain in a given task to abstract aspects of DCNNs. In this sense, the DCNNs is *interpreted*: we have more information about what cognitive activity the DCNN is carrying out. However, the DCNNs status as black boxes remains the same. The *concrete* activities of the DCNN remain completely opaque to the understanding of the researchers. The specific details of a given convolution of two given layers of the deep neural network are unknown. Of course, researchers know what a convolution is, but they do not know the details on how deep neural networks exactly carry it out to achieve the proper performance in visual similarity judgements (e.g., the specific features of the input combined in the convolution step and how they map onto the features combined in natural neural networks; see Poldrack, 2020). The information gathered by the application of the mechanistic theory of cognition, therefore, does not remove the black box but assumes it as such in the mapping with the description of the biological target system.[3]

However, it must be noted that the described mapping between the DCNN and the biological target system is *only* based on performance. Namely, other than the fact that DCNNs perform well in the visual similarity judgement task, there is no other reason to assume that the DCNNs can be mapped in concrete aspects of the biological mechanism. The DCNN remains opaque, thus the mapping can only be done from the biological mechanism to the DCNN and not the other way around: researchers much choose which concrete aspect of the biological mechanism is modelled by the DCNN as there is no intrinsic information of the network to decide on the issue. Again, at all effects, although the DCNN is interpreted, it remains a black box for the neuroscientist.

[3] It should be noted that this strategy is by no means unique to mechanists. A similar tactic has recently been employed by Ofner and Stober (2018), exploiting the affinity between variational auto-encorders (Kingma & Welling, 2014; Rezende, et al, 2014) and the free energy principle as it relates to the predictive activity in the brain (Friston, 2010; Friston & Stephan, 2017)—see also Raja et al. (2021).

The reviewed attempts to interpret deep neural networks have helped researchers gain understanding in limited domains. However, it is clear that, both in the general case and in the concrete case of neuroscience, deep neural networks remain for the most part an epistemically opaque black box. As we have tried to show in the previous two sections, this situation is not a product of the nature of deep neural networks but a product of its theoretical environment. Either because of a general disinterest in neuroscience or because of the theoretical commitments which interpret deep neural networks without removing black boxes, these architectures, although tremendously powerful, remain very limited tools when used to understand cognitive systems. Our proposal is that if we have a theory that better characterizes and constrains the limits of the cognitive systems and their environment (e.g., the kind of things cognitive systems can do and the kind of things environments can offer; see Raja 2020a), we will have better chances to understand what is going on in deep neural networks. Namely, by providing the proper theoretical framework, deep neural networks may be interpreted. This might not solve all the issues but would help us to open some black boxes. In the next section, we provide some guidelines on how this can be done.

4. Deep Neural Networks in the Theory Eden

The previous sections illustrate that at least part of the problem of epistemic opacity in deep neural networks has *quantitative* origins and that at least part of the possible solutions to that problem have to do with some form of quantitative reduction. Researchers are not able to properly explain and understand the activities of deep neural networks because of the many degrees of freedom both in the data used and in the network itself. Deep neural networks are composed of many layers of many nodes with many connections between them. And all these many components have one or more associated parameters. Because of this, most contemporary deep neural networks are defined in terms of incredibly large amounts of parameters—sometimes of the order of tens of millions (Hasson et al., 2020)—that make the activities of the networks impossible to represent and conceptualize by researchers. What is the network defined by that amount of degrees of freedom actually doing? How can we even represent a system of these characteristics? How is each parameter related to the activity of the network? The epistemic opacity of deep neural networks partly depends on the inability to answer these and similar questions on how to address this concrete form of HMI. This is the reason why many strategies for deep neural network interpretability are based on somehow reducing the amount of dimensions in which those parameters define the activities of the networks (i.e., their degrees of freedom) to be able to understand what's going on in them.

The problems of the inherent structure of deep neural networks are exacerbated by the data they deal with. When data is understood as *input*, it provides an extra layer of degrees of freedom. The pictures used to train a network, for instance, can vary in many different ways, can portray different kinds of objects, can include different levels of noise, etc. In other words, information that can be extracted from that input is huge. And the computation-cum-representational paradigm does not provide researchers with an understanding of information compatible with this issue—for a review of the problems with the notion of information and the application of information theory to cognitive science and neuroscience, see, e.g., de-Wit et al. (2016), Nizami (2019), Gallistel (2020). Additionally, the aspects of the *input* deep neural networks use to perform whatever function they are

designed to perform (e.g., classification, recognition, denoising…) are usually different from those cognitive scientists and neuroscientists would enumerate if asked about the way cognitive systems perform the same function. Again, for instance, the aspects of pictures upon which deep neural networks rely in order to accomplish their "visual" function seem to have nothing to do with edges, contours, shades, and all the other features typically used in the cognitive science/neuroscience of vision (see, e.g., Poldrack, 2020). This problem is very clear in the case of adversarial examples; namely, those individual inputs that seem to belong to the same data pool but that make deep neural networks fail in their performance (Buckner, 2019; Ilyas et al., 2019). Instances of these adversarial examples are pictures with a little bit of noise added. Although they have the same properties other pictures have (e.g., same edges, contours, and so on), networks often misclassify them. As in the case of the huge amount of information in the input, researchers have no theoretical resources to deal with adversarial examples insofar as they are not able to map the fundamental concepts used in the cognitive sciences to the activity of deep neural networks, providing a specific instance of *link uncertainty* as described by Sullivan (2020). Our proposal gains traction in facing these two theoretical limitations of the computation-cum-representaitonal paradigm.

In order to improve the human-machine interactions in the context of the sciences of the mind, we propose to look outside its dominant theoretical paradigm to find resources that allow for (i) a better characterization of the information available in the inputs of deep neural networks and (ii) a better characterization of the very activities of those deep neural networks when considered models of the cognitive activities of the brain. More concretely, it is our contention that the notion of *ecological information* from ecological psychology is able to qualitatively and quantitatively constrain the possible activities of deep neural networks (when used as models of the brain) and, therefore, it can work as a theoretical aid to fight against their epistemic opacity. In other words, by framing the explanation of the activities of deep neural networks as models of the brain within the framework of ecological psychology, researchers may find themselves outside of the theoretical desert of the computation-cum-representational paradigm, and in the lush oasis of a theory eden.

Ecological psychology was founded by J. J. Gibson (1966, 1979) and E. J. Gibson (1969) as a general approach to perception and perceptual learning. Later, it became a comprehensive alternative theory for cognitive science in general (Chemero 2009) and its principles and main concepts have been applied fields like social psychology (Heft, 2021), developmental psychology (Adolph & Hoch, 2019), or neuroscience (de Wit & Withagen, 2019; Raja, 2018, 2019, 2020b; Raja & Anderson, 2019). Probably, the most famous ecological concept is the concept of *affordance*, which refers to the opportunities for action organisms find in their environment (Chemero, 2003; Heras-Escribano, 2019). Affordances are, for instance, the *grab-ability* of a mug or the *climb-ability* of a stair. The key aspect of affordances we are interested in is that, according to ecological psychology, they are perceptually available to organisms because they are specified by ecological information. In this sense, we will not offer anymore analysis on affordances themselves but we will focus on what ecological information is, what does it mean that ecological information specifies them, and how it is possible.

Within ecological psychology, ecological information is constituted by patterns of stimulation that surround organisms and that can be used by them to perceptually guide their behavior (Segundo-Ortin, et al., 2019; Warren, 2021). These patterns of stimulation are, put simply, structures in the energies to which the different sensory modalities are sensitive.

For instance, the structures in light in the case of vision, the structure in the vibrating air in the case of audition, or the structures in the dissipation of chemicals in the case of olfaction. In the case of vision, the layout of the surfaces of the objects that surround a given organism reflects the light that comes from the light sources in that environment and structures that light in a particular way—for example, there is likely more illumination over your desk than under your desk due to the layout of the elements of your room/office with regard to the light sources in it. This entails that at any position of the environment where the visual system of an organism can be, a slightly different structure of light is available to it. And when the organism moves around, an ever-changing structure is available to its visual system. This ever-changing structure is called *optic flow* and some *invariant* structures in this flow are considered ecological information for vision in ecological psychology.

But what are these invariant structures that constitute ecological information? It is pretty clear that the typical basic properties in which dominant cognitive science describe visual features, such as color, contours, shapers, and so on, are not invariant in the optic flow. Your desktop looks like a rectangle, a square, or a trapezoid depending on the place in the environment you are looking from, and its color changes depending on whether there is sunlight or artificial light and depending on the intensity of that light. These properties cannot be ecological information insofar as they are not invariant. However, there are other properties of the optic flow that are invariant. Think, for instance, about the optic flow generated by the forward locomotion of an organism: the direction of the heading is *invariably* in the center (origin) of the flow and the flow itself is *invariably* centrifugal from it—i.e., any point of the flow that is not in the origin follows a trajectory that would eventually leave it outside of the visual field. Backward locomotion, on the contrary, generates an *invariably* centripetal flow—i.e., any point of the flow that is not in the origin (now a vanishing point) follows a trajectory that would eventually make it collapse in the vanishing point. These two different kinds of flow are invariant with regard to other features of the environment (e.g., color, shapes, illumination) and lawfully specify events of locomotion on it. In this sense, these two flows are invariant structures that can constitute ecological information.

There are many other invariant structures in the patterns of optic flow that, like the just described centrifugal and centripetal flows, can constitute ecological information and many of them have been already described and formalized in the ecological literature. A chief example of these is *tau* or the time-to-contact invariant (for a review, see Lee, 2009). Put simply, anything that approaches a visual system occupies more and more space in the visual field of the observer as the approach goes on—you can test that just by making your hand approach your eyes and notice how it covers more and more parts of the background as you do it. In other words, the approaching thing expands in the visual field while the non-approaching background remains the same. Such an expanding pattern occurs irrespectively of the color, size, shape, etc., of whatever is approaching the visual system. In this sense, the pattern is invariant in all but in one way that may be referred to as the "speed" of expansion: the faster the approaching thing expands in the visual field, the sooner it will make contact (i.e., will hit/will occupy the same point in space) with the visual system. This is the reason why this invariant structure named *tau* can be used by the observer to know the time-to-contact to an approaching object and, therefore, can be considered as an instance of

ecological information.[4] Indeed, that observers use *tau* to control their behavior with respect to approaching things has been found in a great number of studies both in experimental psychology (e.g., Craig & Lee, 1999; Craig, et al., 2000; Lee, 2005; Lee & Reddish, 1981) and neuroscience (e.g., Sun & Frost, 1998; van der Weel & van der Meer, 2009; van der Weel, et al., 2019).

Given this succinct presentation of ecological information as a central concept within ecological psychology, we want to point out one of the fundamental consequences when it is used to describe the input of a cognitive system like the brain: that ecological information greatly constraints the degrees of freedom of *relevant* information in the input. In other words, there are many ways in which the input may vary (e.g., color, contrast, shape, contours, etc.), but these are just irrelevant for the cognitive system. The only variable the cognitive system needs to detect to properly perform the activity is the variable of ecological information. In the case of approaching objects, for instance, the cognitive system or brain just needs to detect the expansion of the approaching thing (i.e., tau) in the visual field, regardless of the other many properties of the input. As well as other variables of ecological information in other tasks, tau is invariant with respect to the approaching object. Namely, the color of the object, or its shape, or its edges, vary in the visual field with respect to the position of the observer, the illumination conditions, etc. However, the expansion of the object in the visual field on which tau is based *remains invariant*—i.e., remains the same—with regard to all these changes, so the latter are cast out as irrelevant for the system: brains can ignore all the degrees of freedom that do not have to do with tau as only tau is needed to perform the task. In this sense, the informational possibilities of the input are constrained to one kind of invariant event. Thus, just by reframing the theoretical space of cognitive science in terms of ecological psychology and, more concretely, in terms of ecological information, researchers in the field can constrain the degrees of freedom of the input and make it easier to understand when explaining the activities of the cognitive system. And the same can apply to deep neural networks. Deploying the language of ecological information is, in this sense, just another form of *dimensionality reduction*.

Trying to understand deep neural networks from the theoretical space of ecological psychology begins from a somewhat counterintuitive place. The way the ecological framework is able to help solving the *quantitative* issues that underlie the epistemic opacity of deep neural networks when they are used as models of cognitive systems begins by looking at the way what we know about cognitive systems and the environment they face can constrain some of the degrees of freedom of the input. Instead of directly looking at the activities of deep neural networks themselves and trying to limit or minimize their own degrees of freedom, we should start by reducing the degrees of freedom in our data set. The ecological approach forces us to curate our data set in order to provide the network with the relevant ecological information for the task at hand—e.g., a data set in which tau is present for time-to-contact judgements. By doing so, we are sure that, if the deep neural network is a model of brain activity, this activity *must* be dealing with one and only one variable, the one that constitutes ecological information, and we can put all the other sources of variability in the data set aside. This is a straightforward way to reduce the degrees of freedom of the deep

[4] The actual formalization of *tau* is more complicated and involves the inverse of the relative dilatation of the optic angle of the approaching object. We do not need these details to grasp the general idea of ecological information. For further details, see Lee (2009).

neural network and an aid to a better understanding of it. However, it may be unclear why the deep neural network *must* be dealing with ecological information and not other aspects of the input. The answer to this concern is, again, theoretical. Ecological psychology describes (at least perceptual) activities of cognitive systems in terms of detecting ecological information. So, if we accept this theoretical framework and we take deep neural networks to be models of the cognitive systems, they must model this detection of ecological information. The quantitative reduction needed to fight against the epistemic opacity of these networks is therefore acquired by the constraints in the possible explanations posited by the theoretical framework of ecological psychology. The computational-cum-representational framework does not posit these constraints. Indeed, it virtually posits no constraints in what the input can be, in what the computation executed by the deep neural network can be, or in what the deep neural network can be representing. Such a lack of theoretical constraint is what directly leads to the epistemic opacity issue. On the contrary, the much more constrained space allowed by the ecological theory enables understanding. Within the ecological framework, researchers know what's the relevant informational aspect of the input and that the activity of the network—e.g., in a loose sense, the computation executed by it—is detecting that specific information. In this sense, the constraints posited by ecological psychology in our explanations of cognitive systems using deep neural network models can be considered a form of *enabling constraint* (Anderson, 2015; Raja & Anderson, 2021): limitations that enabling some functionality (explainability/understability in this case) by virtue of that very limitation.[5] Summing up, what ecological psychology provides is a theoretical framework that provides tools to overcome the epistemic opacity of deep neural networks when used as models of cognitive systems by virtue of using ecological information as a limitation of the relevant aspects of their inputs and their activities.

It is important to point out that, although our proposal is strictly theoretical so far, some recent work in the field of computer vision, and concretely on the topic of object segmentation and tracking, may be seen as a proof of concept of the theoretical framework presented in here. Tsao & Tsao (2021) have recently provided a mathematical proof of the plausibility of ecological information for visual perception and of its computational implementation. Concretely, they prove that:

> [O]bject surface information is redundantly represented by the field of ambient optic arrays through two of its topological structures: the pseudogroup of *stereo diffeomorphisms* and the set of *infinitesimal accretion borders*. Formulated in terms of ecological optics, vision is a fully constrained, well-posed problem. Complete information for perception of objects as discrete, persistent units is contained in the visual environment itself within the field of ambient optic arrays. (p. 3; emphasis is ours).

[5] A simple example of an enabling constraint is the limitation of the movability of a well-functioning human knee. Well-functioning human knees can bend in very limited ways—pretty much only backwards—and, along with other features of the human body, this limitation is what makes it possible that we walk the way we walk, we run the way we run, or we control our posture the way we do it. In this sense, the constraint in the bending of well-functioning human knees is enabling and, therefore, it is an enabling constraint.

Technicalities aside, Tsao & Tsao (2021) show how two invariants of the optic flow (stereo diffeomorphisms and infinitesimal accretion border) indeed exist in that flow and constitute information adequate to identify objects as individual units (object segmentation) and to track them in the visual field. Namely, these invariants and no other properties of the optic flow (e.g., colors, shapes, contrasts, and so on) are the ones needed for successful object segmentation and tracking. In this sense, any system devoted to implementing object segmentation and tracking, and that has sets of frames of the optic flows as inputs, may use those invariants to be successful. Actually, Tsao & Tsao (2021) show that these functions can be (relatively) easily implemented in a computational system by means of a general algorithm with a few parameters. Insofar as deep neural networks are universal function approximators, the functions should be in principle possible to implement by them—Tsao & Tsao discuss this extent (2021, p. 12-14). Thus, this work can be taken as a concrete proof of concept of our general proposal although we acknowledge more work needs to be done in pursuing its generality.

Beyond the plausibility of ecological information as a constraint for the research with deep neural networks, our proposal still leaves one fundamental question open: how do deep neural networks detect ecological information? In other words, how can we know that the network is actually detecting ecological information and the way it does it? These questions resemble the ones we asked before while analyzing the computational-cum-representational framework. In that moment, we pointed out that the specific details of concrete activities of deep neural networks when they successfully perform a task (i.e., the specific way a face recognition algorithm is implemented by the parameters of the network) remains completely opaque to the understanding of the researchers. Similarly, we can say that, even if we know a deep neural network is detecting ecological information, we do not really know *how* such a detection function is implemented by the network or *what* is actually being represented on it.

To date, the resources the ecological approach provides to answer this kind of question are not as powerful as ecological information in constraining the dimensionality of the data set. However, there are some interesting aspects of the approach that point out a number of possible solutions. First, if deep neural networks are used as models of the brain within the ecological theoretical framework, the function implemented by the network is clear: detecting the relevant ecological information for the task at hand. There is only one variable of ecological information that can be detected, so there is no question on what features of the input are used and, therefore, there is no question on whether the network is implementing an unknown function. This is, again, an example of theoretical constraint that aids explainability and understanding of deep neural networks. And second, ecological psychology provides us with the notion of *resonance* (Raja 2018, 2020b). Put simply, cognitive systems (e.g., brains) detect ecological information by resonating to it, where "resonating" means to be coupled to the dynamics of the situation through ecological information. For instance, suppose a situation in which the relevant ecological information for the task at hand is *tau*. We know that the brain of an observer is resonating to tau if we find that the value of tau of the whole situation and the value of tau of brain dynamics—measured by with an EEG, for instance—are coupled. This kind of situation may be named as *tau*-coupling and has been empirically tested in a variety of studies (e.g., van der Meer et al., 2019, and references therein). The consequence is that if a deep neural network is modeling a cognitive system in this kind of situation, the prediction under the ecological

framework is that the network will be *tau*-coupled to the input in terms of its *global* dynamics for whatever measurement of these dynamics we select. Therefore, notice that the relevant aspect of the activity of the deep neural network in this case has to do with its global dynamics and not with local aspects regarding particular layers or parameters, as it is the case within the computational-cum-representational framework and its feature-based approach, for instance. In this sense, although resonance itself does not fully account for the *how* question regarding the activity of deep neural networks, it does open a different look on its activity—i.e., a global look instead of a local look—that may be productive on its own or in combination with other strategies as the ones suggested in the previous section.

A related consequence of this ecological view of deep neural networks concerns the way in which deep neural network modeling is done in cognitive science. This view suggests not only that we reevaluate *what* it is we are modeling in applying deep neural networks to cognitive systems (i.e., ecological information), constraining the inputs to the system, but also *how* that modeling is done. On the model of ecological psychology, organisms perceive affordances, and thus ecological information, by interacting with features of their environment to uncover invariant features in what we might think of as "dynamic information processing" (cf. Faries & Chemero, 2019). This contrasts with the standard computational-cum-representational view in which computations are made over static representations. By parity of reasoning, this ecological view of deep neural networks suggests that information gleaned from deep neural network models is not best determined by post-hoc quantitative reductions on single instances of a network model. Instead, researchers uncover ecological information about deep neural networks by interacting with them: perhaps by perturbing the system by incremental changes to parameters over multiple runs, or attempting to "break" them by sending inputs engineered to test the system's responses. This is the spirit of work in XAI on iterative random forests (Basu, et al., 2018; Kumbler, et al., 2018), and adversarial networks (Buckner, 2019; Ilyas, et al., 2019), respectively. On an ecological view, these and related efforts are well-deserving of increased focus and attention.

We think these are examples enough of the different ways the theoretical constraints posited by an ecological framework in the understanding of the activities of deep neural networks when they are used as models of cognitive systems may help with the issue of their epistemic opacity. It is worth noting though, that this kind of solution only applies to a specific kind of HMI, that is, to the interaction between cognitive scientist/neuroscientists with the technology of machine and deep learning. Interactions with this kind of technology in other contexts is outside the scope of this paper as the application of the ecological theory to them is not granted without further justification that we will not pursue here.

5. Conclusion

In this paper we have shown that by using a specific notion of information, ecological information, we can improve a specific kind of Human-Machine Interaction (HMI): the one between cognitive scientists and machine/deep learning technologies. We have shown that when deep neural networks are used in the contemporary sciences of the mind, researchers find themselves in a theory desert: they do not have the theoretical resources to explain/understand the activities of deep neural networks and their places within the overall explanatory enterprise of cognitive science and neuroscience. In this sense, deep neural

networks are explanatory black boxes in the field due to what has been labelled as their epistemic opacity. Then, we have claimed that a change in the underlying theoretical assumptions from which deep neural networks are understood in the science of the mind could help against such epistemic opacity. Concretely, we have claimed that understanding deep neural networks under an ecological framework puts them within a theory eden. The notion of ecological information—and, to a lesser extent, the notion of resonance—provides adequate theoretical constraints to aid the explanatory enterprises of cognitive scientists when they use this kind of technology. The success or failure of this theoretical move is, of course, an empirical question and will only be decided after such empirical work is deployed. However, given our own exposition of reasons and the proofs of concept discussed in it, we think we have provided enough theoretical reasons to make the whole enterprise palatable both to philosophers and cognitive scientists.

References

Adolph, K. E., & Hoch, J. E. (2019). Motor development: Embodied, embedded, enculturated, and enabling. *Annual review of psychology,* 70, 141-164.

Anderson, M. L. (2015). Beyond componential constitution in the brain: Starburst Amacrine Cells and enabling constraints. In T. Metzinger & J. M. Windt (Eds.), *Open MIND: 1(T)*. Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958570429.

Bahari, T. F., & Elayidom, M. S. (2015). An efficient CRM-data mining framework for the prediction of customer behaviour. *Procedia computer science, 46*, 725-731.

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science,* 364, 6439.

Basu, S., Kumbier, K., Brown, J. B., & Yu, B. (2018). Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences U.S.A.* 115: 1943–1948.

Birhane, A. & Guest, O. (2021). Towards decolonising computational sciences. *Women, Gender & Research*, 30(1), 60-73.

Buckner, C. (2018). Empiricism without magic: Transformational abstraction in deep convolutional neural networks. *Synthese,* 195(12), 5339-5372.

Buckner, C. (2019). Deep learning: A philosophical introduction. *Philosophy compass,* 14(10), e12625.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings*

of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1721–1730). Sydney, Australia.

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature News,* 538(7623), 20.

Chakraborty S., et al. (2017). Interpretability of deep learning models: A survey of results. *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI).* (pp. 1-6). http://doi.org/10.1109/UIC-ATC.2017.8397411

Chemero, A. (2003). An outline of a theory of affordances. E*cological Psychology* ,15(2), 81–195.

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

Craig, C. M., & Lee, D. N. (1999). Neonatal control of nutritive sucking pressure: Evidence for an intrinsic tau-guide. *Experimental brain research,* 124, 371–382.

Craig, C. M., Delay, D., Grealy, M. A., & Lee, D. N. (2000). Guiding the swing in golf putting. *Nature,* 405, 295–296.

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. New York: Oxford University Press.

Craver, C. F., & Kaplan, D. M. (2020). Are more details better? On the norms of completeness for mechanistic explanations. *The British journal for the philosophy of science*, 71(1):,287-319.

De Cristofaro, E. (2020). An overview of privacy in machine learning. *arXiv*, 2005.08679. https://arxiv.org/abs/2005.08679

de Wit, M. M., & Withagen, R. (2019). What should a “Gibsonian neuroscience” look like? *Ecological psychology,* 31(3), 147-151. http://doi.org/10.1080/10407413.2019.1615203

de-Wit, L., Alexander, D., Ekroll, V., & Wagemans, J. (2016). Is neuroimaging measuring information in the brain? *Psychonomic bulletin & review,* 23, 1415-1428.

Delalleau, O., & Bengio, Y. (2011). Shallow vs. deep sum-product networks. In *NIPS'11: Proceedings of the 24th International Conference on Neural Information Processing Systems* (pp. 666–674). New York: Curran Associates Inc.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv,* 1702.08608.

Durán, J., & Formanek, N. (2018). Grounds for trust: Essential epistemic opacity and computational reliabilism. *Minds and machines*, 28, 645–666.

Faries, F., & Chemero, A. (2019). Dynamic information processing. In M. Sprevak & M. Colombo, (Eds.), *The Routledge handbook of the computational mind* (pp. 134-148). London: Routledge.

Favela, L. H., Riley, M. A., Shockley, K., & Chemero, A. (2018). Perceptually equivalent judgments made visually and via haptic sensory-substitution devices. *Ecological Psychology*, *30*(4), 326-345.

Fodor, J. (1975). *The Language of Thought*. New York: Thomas Y. Crowell.

Fodor, J. (1981). *Representations*. Cambridge: MIT Press.

Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *Annals of Applied Statistics, 2*, 916–954.

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

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

Gallistel, C. R. (2020). Where meanings arise and how: Building on Shannon's foundations. *Mind & language*, 35, 390–401. https://doi.org/10.1111/mila.12289

Gibson, E. J. 1969. *Perceptual learning and development*. New York: Appleton-Century-Crofts.

Gibson, J. J. (1966). *The senses considered as perceptual systems*. Boston: Houghton Mifflin.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, MA: MIT Press.

Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct fit to nature: An evolutionary perspective on biological and artificial neural networks. *Neuron*, 105(3), 416-434.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *arXiv*, 1502.01852.

Heft, H. (2021). Ecological psychology as social psychology? *Theory & psychology,* 30(6), 813-826.

Heras-Escribano, M. (2019). *The philosophy of affordances*. London: Palgrave Macmillan.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science, 313*(5786), 504-507.

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural computation, 18*(7), 1527-1554.

Howe, J. (1994). Artificial Intelligence at Edinburgh University : A perspective. Archived from the original on 17 August 2007. Retrieved 30 August 2007.

Humphreys, P. (2004). *Extending ourselves: Computational science, empiricism, and scientific method*. New York: Oxford University Press.

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese,* 169, 615–626.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). Adversarial examples are not bugs, they are features. *arXiv,* 1905.02175. http://arxiv.org/abs/1905.02175.

Kaplan, D. M., & Craver, C. F. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy of science* ,78(4), 601-627.

Kim, B., Chacha, C., & Shah, J. (2015). Inferring robot task plans from human team meetings: A generative modeling approach. *Journal of Artificial Intelligence research* ,52:,361-398.

Kindermans, P.J., Schütt, K.T., Alber, M., Muller, K, Erhan, D., Kim, B., & Dähne, S. (2018). Learning how to explain neural networks: Patternnet and Patternattribution. *arXiv*, 1705.05598 https://arxiv.org/abs/1705.05598

Kingma, D. P., and Welling, M. (2014). Auto-encoding variational Bayes. *arXiv*, 1312.6114. https://arxiv.org/abs/1312.6114.

Kingma, D. P., and Welling, M. (2019). An introduction to variational autoencoders. *arXiv*, 1906.02691. https://arxiv.org/abs/1906.02691.

Kriegesgorte, K., & Douglas, P. (2018). Cognitive computational neuroscience. *Nature Neuroscience,* 21, 1148–1160.

Kumbier, K., Basu, S., Brown, J. B., Celniker, S., & Yu, B. (2018). Refining interaction search through signed iterative random forests. arXiv:1810.07287

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

LeCun, Y., Touresky, D., Hinton, G., & Sejnowski, T. (1988). A theoretical framework for back-propagation. In *Proceedings of the 1988 connectionist models summer school* (Vol. 1, pp. 21-28). Burlington, MA: Morgan Kaufmann.

Lee, D. N. (2005). Tau in action in development. In J. J. Rieser, J. J. Lockman, & C. A. Nelson (Eds.), *Action as an organizer of learning and development* (pp. 3–49). Hillsdale: Lawrence Erlbaum Associates.

Lee, D. N. (2009). General tau theory: Evolution to date. *Perception,* 38, 837–858.

Lee, D. N., & Reddish, P. E. (1981). Plummeting gannets: A paradigm of ecological optics. *Nature,* 293, 293–294.

Letham, B., Rudin, C., McCormick, T. H., Madigan, D. (2015). Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics, 9*, 1350–1371.

Lighthill, J. (1973). Artificial intelligence: A general survey. *Artificial Intelligence: A paper symposium*. Science Research Council.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy, 23*, 18. https://dx.doi.org/10.3390/e23010018

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue,* 16(3), 31-57.

Lobo, L., Nordbeck, P. C., Raja, V., Chemero, A., Riley, M., Travieso, D., and Jacobs, D. (2019). Route selection and obstacle avoidance with a short-range haptic sensory substitution device. *International Journal of Human–Computer Studies, 132,* 25–33.

Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013). Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 623–631). Gold Coast, Australia.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics* ,*5*(4):,115-133.

McDermott, D., Waldrop, M. M., Chandrasekaran, B., McDermott, J., & Schank, R. (1985). The Dark ages of AI: A panel discussion at AAAI-84. *AI Magazine, 6*(3), 122. https://doi.org/10.1609/aimag.v6i3.494

Minsky, M., & Papert, S. A. (1969). *Perceptrons: An introduction to computational geometry.* Cambridge: MIT Press.

Minsky, M., & Papert, S. A. (1972). *Artificial intelligence progress report*. AI Memo 252, Massachusetts Institute of Technology.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature,* 518(7540), 529-533.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences of the United States of America, 116(44), 22071–22080. https://doi.org/10.1073/pnas.1900654116

Nizami, L. (2019). Information theory is abused in neuroscience. *Cybernetics and human knowing,* 26(4), 47-97.

Ofner, A., & Stober, S. (2018). Towards bridging human and artificial cognition: Hybrid variational predictive coding of the physical world, the body and the brain. *Advances in neural information processing systems* - Proceedings of NeurIPS 2018.

Parker, W. (2013). Computer simulation. In S. Psillos and M. Curd (Eds.), *The Routledge Companion to Philosophy of Science, 2nd Edition* (pp. 135–145). Oxford , UK: Routledge.

Perconti, P., & Plebe, A. (2020). Deep learning and cognitive science. *Cognition,* 203, 104365.

Pereira, A., & Thomas, C. (2020). Challenges of machine learning applied to safety-critical cyber-physical systems. *Machine Learning & Knowledge Extraction,* 2, 579-602.

Poldrack, R. A. (2020). The physics of representation. *Synthese*. https://doi.org/10.1007/s11229-020-02793-y

Pylyshyn, Z. (1984). *Computation and Cognition*. Cambridge, MA: MIT Press.

Raja, V. (2018). A theory of resonance: Towards an ecological cognitive architecture. *Minds and machines,* 28(1), 29–51.

Raja, V. (2019). From metaphor to theory: The role of resonance in perceptual learning. A*daptive behavior,* 27(6), 405–421.

Raja, V. (2020a). Embodiment and cognitive neuroscience: The forgotten tales. *Phenomenology and the cognitive sciences*. https://doi.org/10.1007/s11097-020-09711-0

Raja, V. (2020b). Resonance and radical embodiment. *Synthese*. https://doi.org/10.1007/s11229-020-02610-6

Raja, V., & Anderson, M. L. (2019). Radical embodied cognitive neuroscience. *Ecological psychology,* 31(2), 166–181.

Raja, V., & Anderson, M. L. (20201). Behavior considered as an enabling constraint. In F. Calzavarini and M. Viola (Eds.), *Neural Mechanisms* (pp. 209-232). Cham, Switzerland: Springer.

Raja, V., & Calvo, P. (2017). Augmented Reality: An ecological blend. *Cognitive Systems Research*, 42, 58-72.

Raja, V., Valluri, D., Baggs, E., Chemero, A., & Anderson, M. L. (2021). The Markov blanket trick: On the scope of the free energy principle and active inference. [Under review. Preprint available at: http://philsci-archive.pitt.edu/18843/].

Rezende, D. J., Mohamed, S., & Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. *arXiv*, 1401.4082. https://arxiv.org/abs/1401.4082

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *arXiv*, 1602.04938. https://arxiv.org/abs/1602.04938

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D.E. Rumelhart, J.L. McClelland (Eds.)*, Parallel distributed processing,* Vol. 1 (pp. 318-362). Cambridge, MA: MIT Press.

Russell, S. J., & Norvig, P. (2003). *Artificial intelligence: A modern approach* (2nd ed.), Upper Saddle River, New Jersey: Prentice Hall.

Safi, Y., & Bouroumi, A. (2013). Prediction of forest fires using artificial neural networks. *Applied Mathematical Sciences, 7*(6), 271-286.

Schettler, A., Raja, V., & Anderson, M. L. (2019). The Embodiment of Objects: Review, Analysis, and Future Directions. *Frontiers in Neuroscience*, 13, 1332. http://doi.org/10.3389/fnins.2019.01332

Segundo-Ortín, M., Heras-Escribano, M., & Raja, V. (2019). Ecological psychology is radical enough: A reply to radical enactivists. *Philosophical psychology,* 32(7), 1001–1023.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *arXiv*, 1610.02391. https://arxiv.org/abs/1610.02391

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System technical journal 27*, 379-423, 623-656.

Stuart, M., & Nersessian, N. (2019). Peeking inside the black box: A new kind of scientific visualization. *Minds and machines,* 29, 87–107.

Suissa-Peleg, A., Haehn, D., Knowles-Barley, S., Kaynig, V., Jones, T. R., Wilson, A., Schalek, R., Lichtman, J. W., & Pfister, H. (2016). Automatic neural reconstruction from petavoxel of electron microscopy data. *Microscopy and microanalysis,* 22(S3), 582-583.

Sullivan, E. (2020). Understanding from machine learning models. *The British journal for the philosophy of science*. Forthcoming. Preprint at: http://philsci-archive.pitt.edu/16276

Sun, H., & Frost, B. J. (1998). Computation of different optical variables of looming objects in pigeon nucleus rotundus neurons. *Nature neuroscience,* 1(4), 296–303.

Tsao, T., & Tsao, D. Y. (2021). A topological solution to object segmentation and tracking. *arXiv*, 2107.02036. https://arxiv.org/abs/2107.02036

van der Weel, F. R., & van der Meer, A. L. H. (2009). Seeing it coming: infants' brain responses to looming danger. *Naturwissenschaften,* 96(12), 1385–1391.

van der Weel, F. R., Agyei, S. B., & van der Meer, A. L. H. (2019). Infants' brain responses to looming danger: Degeneracy of neural connectivity patterns. *Ecological psychology,* 31(3),182–197.

Warren, W. H. (2021). Information is where you find it: Perception as an ecologically well-posed problem. *i-Perception,* 12(2), 1-24.

Werbos, P. J. (1982). Applications of advances in nonlinear sensitivity analysis. In *System modeling and optimization* (pp. 762-770). Springer: Heidelberg.

Winsberg, E. (2001). Simulations, models, and theories: Complex physical systems and their representation. *Philosophy of science*, 68, S442–S454.

Winsberg, E. (2010). *Science in the age of computer simulation*. Chicago, USA: University of Chicago Press.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv*, 1609.08144. https://arxiv.org/abs/1609.08144

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience,* 19(3), 356-365.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European Conference on Computer Vision.* Springer: Berlin/Heidelberg, Germany (pp. 818–833). https://doi.org/10.1007/978-3-319-10590-1_53