Debt-free intelligence: Ecological information in minds and machines

Tyeson Davies-Barton,[1] Vicente Raja,[2] Edward Baggs,[3,4] and Michael L. Anderson[2,5,6]

[1]Department of Philosophy, University of British Columbia, Main Mall, V6T 1Z1, Vancouver, British Columbia, Canada.
[2]Rotman Institute of Philosophy, Western University, Richmond Street, N6A 5B7, London, Ontario, Canada.
[3]Department of Language and Communication, University of Southern Denmark, Campusvej, 5230, Odense, Denmark.
[4]Danish Institute for Advance Study, University of Southern Denmark, Fioniavej, 5230, Odense, Denmark.
[5]Brain and Mind Institute, Western University, Richmond Street, N6A 5B7, London, Ontario, Canada.
[6]Department of Philosophy, Western University, Richmond Street, N6A 5B8, London, Ontario, Canada.

Corresponding author's E-mail: tyesondaviesbarton@gmail.com
Contributing authors: vgalian@uwo.ca; ebag@sdu.dk; mljanderson@gmail.com

Abstract Cognitive scientists and neuroscientists typically understand the brain as a complex information-processing system. A limitation of this information-processing metaphor is that it requires that the brain has access to a finite set of possible informational messages—a neural code—and it is unclear how this can be accounted for without appealing to a priori knowledge. For this reason, Dennett once argued that the information-processing metaphor requires cognitive neuroscience to take out a non-repayable loan of intelligence. However, recent advances in machine learning have resulted in the development of a family of algorithms, including the class of algorithms known as autoencoders, that seem capable of evading the problem of non-repayable loans of intelligence. We evaluate whether autoencoders are indeed resilient against the loans of intelligence problem. We agree that they can be so characterized. We argue, however, that autoencoders can more usefully be understood not in terms of Shannon information but instead as a proof of concept of how neural networks can attune to ecological or Gibsonian information. We thus propose that autoencoders belong to a class of algorithms for modeling the brain that have recently been dubbed direct fit algorithms.

Keywords Cognitive neuroscience · Information theory · Ecological psychology · Autoencoders · Machine learning

1 Introduction

Cognitive scientists and neuroscientists have traditionally thought of the brain as an extremely complex communication system. This way of thinking about the brain can be traced back to Helmholtz's telegraph metaphor in 1863. According to the metaphor, the brain—and the nervous system as a whole—can be thought of as being composed of different structures or communication centers that send coded messages to one another. For example, the human retina is said to encode a visual input signal that is sent to the visual processing areas of the cortex, where it is then decoded or processed. An advantage of thinking of the brain as a communication system is that it enables cognitive neuroscientists to describe the brain in terms of information theory, the formal mathematical framework developed by Claude Shannon (1948). Shannon's theory provides a formal definition of information, understood in the form of messages that can be sent using a finite pre-existing code, or alphabet. This alphabet is often referred to as the communication system's lookup table.

Describing the brain in terms of information theory has historically proved extremely fruitful. However, many have argued that there are also substantial limitations that come along with the decision to describe the brain as a communication system. In this paper we focus on an objection raised by Dennett (1981) known as the loans of intelligence

problem. The problem is as follows. A communication system can only operate if it already has access to a pre-existing finite code, or alphabet, of possible messages. If the brain is literally a communication system, then at least two questions immediately arise: What is the brain's alphabet? And where does the brain acquire this alphabet from? Cognitive neuroscientists began to think of the brain as a communication system long before they were able to answer these two crucial questions. Dennett's accusation was that these neuroscientists were thereby taking out a loan of intelligence.

One strategy that cognitive neuroscientists can use to try to evade the loans of intelligence problem is the appeal to learning. Perhaps the brain fills in the content of its lookup table purely through learning, for example by applying some appropriate learning algorithm to the input that it receives. We examine one class of unsupervised learning algorithms, known as autoencoders, that may be able to learn in this way. Autoencoders are a fairly simple form of neural network designed to take some input signal and copy it into an identical or near-identical output. What is interesting about autoencoders is that they are able to achieve this despite the fact that the hidden layers of the network are of lower dimensionality than either the input or the output layer. In order to copy the input, autoencoders cannot transmit the entire structure of the input in uncompressed form, but must extract hidden regularities of the input that it can use to compress the message. This extraction of regularities could perhaps be thought of as a process whereby the network constructs its own lookup table. If this is correct, then autoencoders would constitute a proof, in principle, that it is possible to evade the loans of intelligence problem while maintaining that the brain is a communication system based on Shannon information.

However, there are ways of thinking about information different from Shannon's. One alternative is Gibsonian information (Gibson 1950, 1966, 1979). Shannon conceived of information as a finite set of messages that are transmitted in the form of a code. Gibson, by contrast, thought of information as the structure that is available in the ambient energy that surrounds an observer. For Gibson, information consists in patterns that exist in the light, sound, chemical distributions, and so on, all of which—in the case of terrestrial animals—can be detected in the air.

We will suggest that autoencoders are better understood not in the framework of Shannon information, but in the framework of Gibsonian information. Autoencoders indeed constitute a proof of concept for a solution to the loans of intelligence problem. But the solution is not that the brain fills in a lookup table by making use of an unsupervised learning algorithm. The solution is to reject the idea that the brain requires a lookup table in the first place. We suggest that the brain does not require a lookup table because the brain is not a communication system. Towards the end of the paper we claim that autoencoders are best thought of as belonging to a family of modeling algorithms that Hasson et al. (2020) recently dubbed direct fit models. In our opinion, such direct fit models provide a promising direction for future brain modeling research.

2 Information theory and the problem of non-repayable loans of intelligence

Information theory (Shannon 1948) is the framework for how information is understood in mainstream cognitive neuroscience (Gallistel & King 2010; Marr 1982; Rieke et al. 1997; Stone 2012). Regarding perception, the story goes as follows: sensory inputs (e.g., aspects of the retinal image) are conceptualized as *messages* containing information about properties of the environment (e.g., surface properties of objects) in that the former encode the latter. Neurological processes also encode or reencode sensory input in accordance with a neural code,[1] and pass the messages along, until eventually a part of the brain decodes the messages resulting in a perceptual representation of the objective world (Gallistel & King 2010; Stone 2012). According to standard Shannon information theory, such a communication system only works if there is a limited and already known set of possible messages that could be communicated, as is the case for instance with Morse Code. One of the most significant objections to this information-theoretic framework for perception is its inability to explain how the brain acquires or already possesses the knowledge of the range of possible environmental messages (i.e., its lookup table) (Dennett 1978; Turvey 2019; Turvey et al. 1981). Following Dennett (1981), we call this the problem of non-repayable loans of intelligence. In this section, we outline the information-theoretic framework for perception, and explain why all models based on it are limited due to this problem of non-repayable loans.

2.1 The brain as a communication system

For decades, mainstream cognitive neuroscience has used information theory as its organizing framework[2], explicitly treating the brain as a communication system (Gallistel & King 2010; Marr 1982; Rieke et al. 1997; Stone 2012). For instance, in an early review of the field, Perkel and Bullock (1968) portray "the nervous system [as] a communication machine," while in a recent commentary, Gallistel (2019) claims that information theory is vital for understanding "world-brain communication" (p. 26).

While this model is applied to the brain at numerous levels of analysis,[3] the central hypothesis for perception is that sensory stimulations and neural activities function as

[1] The exact form of such neural codes is a debated topic. Main proposals include variations of rate coding, wherein information is transmitted via firing rates (e.g., Rieke et al. 1997), and temporal coding, wherein information is transmitted via the timings of action potentials, (e.g., Foffani et al. 2009).

[2] For some of the first applications of information theory in cognitive science and neuroscience, see Attneave (1954), MacKay & McCulloch (1952), and Rapoport & Horvath (1960).

[3] In a recent survey of information theory's applications in cognitive neuroscience, Nizami (2019) demonstrates that there are at least *hundreds* of different ways in which the model has been applied to the brain, and that there is no consensus on which level of analysis is the most valid.

information about environmental properties insofar as the brain can decode them to produce perceptual representations of the world, and appropriate motor responses (Gallistel & King 2010; Stone 2012). For instance, regarding sensory stimulations, Gallistel and King (2010) claim that "the physical processes in the world that convert source information (for example, the reflectance of a surface) to proximal stimuli (the amount of light from that surface impinging on the retina) encode the source information" (p. 23). And regarding neural activities, Brette (2019) describes the received view as follows:

> In what sense is the neural code 'information' about objective properties of the world? According to the technical sense of coding, it is information in the sense that these properties can be inferred from neural activity. (2019, p. 6)

This information-theoretic understanding of neural activity is expressed in common claims such as that a "fly's visual system carries information about motion in the timing of spikes down to sub-millisecond resolution" (Nemenman et al. 2008, p. 4), or that "many cortical neurons encode variables in the external world via bell-shaped tuning curves" (Series et al. 2004, p. 1129).

The choice to use information theory to describe brain processes places constraints on the kinds of representations that neuroscientists say they are investigating. For instance, Kravitz et al. (2011) claim that activity in the parahippocampal place area (PPA) functions as information for the spatial, but not the semantic properties of a visual scene because, as they write, "the response of PPA could not be used to decode the high-level semantic category of scenes even when spatial factors were held constant, nor could category be decoded across different distances," while spatial properties could be decoded from such activity (p. 7322).

These constraints arise from the way that Shannon information is defined. According to Shannon (1948), a communication system is composed of five parts: first, an information source (which produces the message); second, a transmitter (which encodes the message and transmits the encoded message as a signal); third, a channel (the medium through which the signal is transmitted); fourth, a receiver (which decodes the signal so as to retrieve the original message); and fifth, a destination (the individual for whom the message is intended). Such systems function in accordance with a set of rules by which the messages are encoded and decoded that thereby specifies which signals correspond with which possible messages. Obviously, both the sender and the receiver must agree upon the distribution of possible messages and the encoding algorithm for the former to be able to successfully communicate to the latter.

Shannon information is a measurement of successful communication—of the amount of information that is transmitted from a communication system's source to its receiver. More formally, it is a measurement of how much *the receiver's* uncertainty about the probability distribution of the set of possible messages coming from a particular source is

reduced when the receiver decodes either a single message or a sequence of messages from that distribution (Shannon 1948). For our purposes, the key point here is that Shannon information is neither an objective property of the message nor the signal but is rather relative to the communication system. As Gallistel and King (2010) put it, "the information communicated from a source to a receiver by a signal is an inherently subjective concept" (p. 10).

Shannon information is subjective for two reasons. First, whether something qualifies as information depends on whether it can be obtained by a receiver via a decoding process. As de-Wit et al. (2016) explain:

> Good encryption algorithms will make the target information appear as noise to an observer or receiver who does not have the correct decryption key. When the observer has the correct decryption key, the information in the message is interpretable. (p. 1417)

For instance, a single dot (a signal) can be considered information rather than noise for a receiver that knows Morse code (i.e., the look-up table in this case), and can thereby retrieve the letter 'E' (the message) by decoding the single dot.

Second, Shannon information is receiver-dependent because, as mentioned earlier, it is measured in terms of the amount of uncertainty that it reduces in a receiver. As such, it is calculated with reference to the *receiver's beliefs* about the relevant probability distribution, which can be divided into two categories: first, beliefs about the values in the probability distribution (the messages that could be sent); and second, beliefs about the probability of each value's occurrence (the likelihood of each particular message being sent) (Gallistel & King 2010).

To illustrate, consider the following: "two messages, one of which is heavily loaded with meaning and the other of which is pure nonsense, can be exactly equivalent, from the present viewpoint, as regards information" (Weaver 1953, p. 265). Specifically, if the receiver believes that the two messages come from the same amount of possible messages, and have the same likelihood of occurring, then the two messages will contain the same amount of information, regardless of their meanings. This is formally illustrated by Shannon's equation for calculating the amount of information conveyed in a single message:

$$h(i) = log_2 \frac{1}{P(i)}$$

In this equation, h is the standard notation for Shannon information, i is a single message from the probability distribution, and P(i) is the probability of that message's occurrence according to the receiver. Evidently, in order to have a belief concerning P(i), the receiver must also have beliefs concerning the other messages it could receive as well as their

probabilities of occurrence. For another example, consider Shannon's equation for calculating the average Shannon information (also known as *entropy*) of the value of a random discrete variable from a probability distribution:

$$H(X) = \sum_{i=1}^{n} P(x_i) log_2 \frac{1}{P(x_i)}$$

In this equation, *H(X)* is the average Shannon information (the entropy) of the value of a random discrete variable *X* from the probability distribution, *n* is the number of possible messages in the probability distribution, and P(*xi*) is the probability of a random message in the probability distribution. Clearly, in order to calculate H(X), beliefs concerning *n* and P(*xi*) must be supplied by the receiver.

2.2 The problem of non-repayable loans of intelligence

Many have expressed caution about using Shannon's information theory as a framework for studying human beings, including Shannon himself (Shannon 1956). One of the most significant objections is that applying information theory to the brain involves taking out a non-repayable loan of intelligence (Dennett 1981). As Dennett (1981) explains:

> Any time a theory builder proposes to call any event, state, structure, etc., in any system (say the brain of an organism) a *signal* or *message* or *command* or otherwise endows it with content, he *takes out a loan* of intelligence. He implicitly posits along with his signals, messages, or commands, something that can serve as a signal-*reader*, message-*understander*, or *commander*, else his 'signals' will be for naught, will decay unreceived, uncomprehended. (p. 12)

The problem is that information-theoretic models of perception take out loans of intelligence that cannot be repaid within this framework.

Let us illustrate this problem by looking at the efficient coding hypothesis (Stone 2012). This theory posits that the brain retrieves properties of distal objects by decoding proximal stimuli in accordance with coding algorithms. By viewing the brain's perceptual access to the external world as a process of "world-brain communication" (Gallistel 2019, p. 26), the efficient coding hypothesis takes out no less than four loans of intelligence. First, it endows the brain with innate knowledge of the set of distal properties that proximal stimuli and neural codes allegedly represent (i.e., the set of possible messages that the brain can receive). As Brette (2019) notes, "neural codes carry information only by reference to things with known meaning" (p. 2). Second, it endows the brain with innate beliefs about the likelihood of each distal property being transmitted by proximal stimuli (i.e., the probability of each message being sent). Gallistel and King (2010) explicitly

acknowledge these two loans of intelligence when claiming that "the information communicated by a signal depends on the receiver's (the subject's) prior knowledge of the possibilities and their probabilities" (p. 9). Third, it endows the brain with innate knowledge of the correspondences that allegedly exist between distal properties and proximal stimuli, since such knowledge is necessary for the brain's decoding process. And fourth, it endows the brain with innate knowledge of the code by which distal properties are encoded by proximal stimuli. Again, as Gallistel and King (2010) affirm, "no agreement about code between sender and receiver, no communication" (p. 7). To decode proximal stimuli, the brain must know, for instance, the physical laws by which the reflectance levels of objects are encoded into luminance levels in the retinal image (Gallistel & King 2010, p. 23). As Warren (2005) explains, such decoding requires that the brain innately knows, among other things, "that natural surfaces are regularly textured, that terrestrial objects obey the law of gravitation," and "that light comes from above" (p. 357-358).

These four loans of intelligence, however, are ultimately non-repayable within this form of explanation. In order for an organism to *acquire* knowledge about its environment —concerning, for instance, distal properties, the laws of gravitation, or the common direction of light—it must *already* be epistemically open to its environment. Therefore, such knowledge cannot be used to explain *how* organisms gain this epistemic access without vicious circularity. In other words, if prior knowledge about the environment in the form of a lookup table is necessary *for* perception, then such knowledge cannot be acquired *through* perception. But this framework cannot explain the source of this non-perceptual knowledge (Dennett 1981; Chemero 2009; Turvey et al. 1981; Turvey 2019).[4] As Dennett (1981) summarizes, such a form of explanation "will have among its elements unanalyzed man-analogues endowed with enough intelligence to read the signals, etc., and thus the theory will postpone answering the major question: what makes for intelligence?" (p. 12). But more importantly for our purposes, such forms of explanation will postpone answering *how* organisms perceive their environments.

3 Autoencoders as loan-free communication systems?

We take the argument concerning non-repayable loans of intelligence to apply to all proposals that characterize cognitive systems as engaging in some form of coding-and-decoding or inference to gather epistemological contact with their environments. This includes all proposals that make use of information theory as a metaphor for cognitive systems. If this is correct, then it would seem to warrant the strong conclusion that information theory as developed by Shannon and others is a non-starter for understanding cognitive systems (see Warren 2021).

[4] Notably, the problem of non-repayable loans of intelligence is not unique to contemporary explanations of perception; it has been a longstanding issue for all accounts that treat perception as a process involving coding and decoding or inferences (Turvey 2019).

However, recent advances in machine learning, and more concretely in representational learning, have generated systems that appear to be engaging in such coding-and-decoding or inferential activity, yet without requiring non-repayable loans of intelligence. These systems are *autoencoders* (Baldi 2012). The workings of autoencoders are often illustrated in terms of communication systems (e.g., Hinton & Zemel 1993; Hinton & van Camp 1993) and, more generally, the tools of information theory are used to describe them (see Goodfellow et al. 2017). In this sense, autoencoders can be seen as instances of communication systems, i.e., systems that engage in effective coding-and-decoding information in an input-output fashion. And their robustness against the argument of non-repayable loans of intelligence may be seen as a vindication of the application of information theory to the brain. So the story goes: if autoencoders can be taken to be a model of the brain and they need no loans of intelligence to represent their input, information theory may be adequate for cognitive science and neuroscience. In this section, we critically engage with this way of understanding autoencoders, their status as a model of the brain, and their relationship to information theory.

3.1 What is an autoencoder?

Autoencoders are the archetypical example of an unsupervised representation learning algorithm (Goodfellow et al. 2017).[5] Usually built up as feed-forward neural networks, autoencoders are trained to instantiate a function that copies their input to their output. In the process of doing so, autoencoders first *encode* the input in the form of a low-dimensional representation of its relevant features and then *decode* this representation back to the original form of the input. Importantly, autoencoders seem to be able to do so without supervision and, therefore, without requiring non-repayable loans of intelligence to guide their learning.

As just noted, an autoencoder is usually a kind of feed-forward neural network. Feed-forward neural networks are mathematical objects that can be represented as having components (nodes) and connections (edges) and that may be used to approximate mathematical functions. In this sense, feed-forward neural networks are able to provide outputs $\boldsymbol{y}$ for inputs $\boldsymbol{x}$, such that $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$ being $\boldsymbol{f}$ the function we want to approximate. Autoencoders are specific kinds of feed-forward neural networks that aim to approximate a function that provides an output $\hat{\boldsymbol{x}}$ for an input $\boldsymbol{x}$ such that $\hat{\boldsymbol{x}} = \boldsymbol{x}$ or, at least, $\hat{\boldsymbol{x}} \approx \boldsymbol{x}$. In simpler terms, autoencoders aim to copy their input to their output. Of course, a trivial and

[5] We provide a high-level, conceptual understanding of autoencoders. To do so, first, we avoid most technical details. For those details, we refer the reader to the abundant literature on the topic (e.g., Goodfellow et al. 2017; Baldi 2012). And second, we accept the linguistic conventions in the machine learning literature without further argument—e.g., we will use "representation learning" as it is commonly understood in the field although, for instance, there might be tensions between this notion of representation and the one used in the cognitive sciences (see Anderson & Champion, in press). We will only qualify these linguistic conventions if it is strictly necessary for our own argument.

uninteresting way to do so is learning the identity function. However, autoencoders are interesting because by adding some constraints to the system, they encode a lower-dimensional *representation* $\hat{\boldsymbol{h}}$ of the relevant factors $\boldsymbol{h}$ that account for the variability in the input $\boldsymbol{x}$ (see Figure 1).[6] And they do so without the need for explicit guidance about the correct values of the output $\hat{\boldsymbol{x}}$, but just by minimizing a *loss function* that captures the difference between $\hat{\boldsymbol{x}}$ and $\boldsymbol{x}$.[7] In this sense, autoencoders are able to learn $\hat{\boldsymbol{h}}$ without using labeled training sets and, therefore, autoencoders are an instance of unsupervised learning.

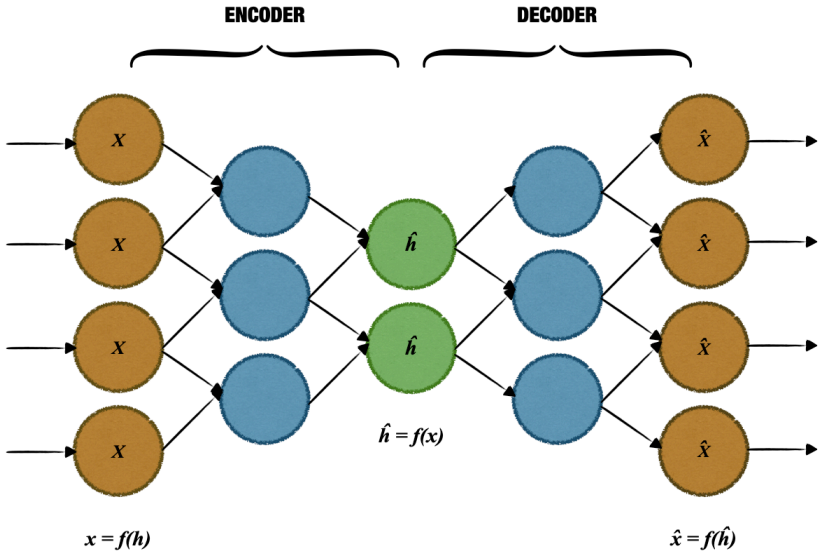

Fig. 1 General schema of an autoencoder. A feed forward neural network composed of nodes (circles) and edges (arrows). It has an input layer (left brown) and an output layer (right brown). The flow of information goes from the input layer to the output layer through the hidden layers (green and blue). The input of the autoencoder is $\boldsymbol{x}$, which is generated by $\boldsymbol{h}$ such that $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{h})$. The aim of the autoencoder is to give an output such that $\hat{\boldsymbol{x}} = \boldsymbol{x}$ or, at least, $\hat{\boldsymbol{x}} \approx \boldsymbol{x}$. To do so, the autoencoder learns a representation $\hat{\boldsymbol{h}}$ such that $\hat{\boldsymbol{h}} = \boldsymbol{f}(\boldsymbol{x})$ and, then, generates $\hat{\boldsymbol{x}}$ from $\hat{\boldsymbol{h}}$ mimicking the generative process form $\boldsymbol{h}$ to $\boldsymbol{x}$. The autoencoder can be further distinguished as an *encoder*, that provides as an outcome $\hat{\boldsymbol{h}}$ from the $\boldsymbol{x}$ input (from left brown to green in the figure), and

[6] These constraints may be of very different kinds: forcing $\hat{\boldsymbol{h}}$ to be lower-dimensional than $\boldsymbol{x}$ (undercomplete autoencoders), or forcing the network to have less edges between nodes (sparse autoencoders), or using a corrupted input (denoising autoencoders), among others (for a review, see Goodfellow et al. 2017).

[7] There are many different loss functions, from mean squared error to evidence lower bound (ELBO)/negative free energy, for instance. The fundamental idea is that, by minimizing the loss function, the autoencoder ends up instantianting a good $\hat{\boldsymbol{h}} \approx \boldsymbol{h}$ given the training data.

a *decoder*, that provides as an outcome $\hat{\boldsymbol{x}}$ from the $\hat{\boldsymbol{h}}$ input (from green to right brown in the figure). The autoencoder learns the representation $\hat{\boldsymbol{h}}$ in an unsupervised fashion—i.e., without the need for labeled data.

3.2 Autoencoders and the loans of intelligence problem

Autoencoders are regarded as archetypal examples of representational learning insofar as $\hat{\boldsymbol{h}}$ is understood as a representation of $\boldsymbol{x}$ and, more concretely, as a representation of the set of factors $\boldsymbol{h}$ that give rise to the data $\boldsymbol{x}$. The concrete relationship between $\boldsymbol{h}$ and $\boldsymbol{x}$ can be described in different ways, but $\boldsymbol{h}$ is usually taken to be the set of *causal factors* that give rise to the whole probability distribution of data from where $\boldsymbol{x}$ is selected. In other words, $\boldsymbol{h}$ are the causal factors of the *generative process* that gives rise to $\boldsymbol{x}$. For instance, $\boldsymbol{h}$ would be the set of environmental factors that cause the set of retinal inputs $\boldsymbol{x}$; and $\hat{\boldsymbol{h}}$ could be understood as a model of the environmental factors that cause those retinal inputs. This fact has two consequences.

First, the general consequence is that autoencoders may be understood as two connected feed-forward networks: the *encoder* and the *decoder*. In Figure 1 (above), we can split the whole network by the middle green layer. If we do so, the left half of the network would be the encoder and would have $\boldsymbol{x}$ as input and $\hat{\boldsymbol{h}}$ as output. Conversely, the right half of the network would be the decoder and would have $\hat{\boldsymbol{h}}$ as input and $\hat{\boldsymbol{x}}$ as output. There are two compatible ways to characterize the relationship between encoder and decoder. First, the decoder is just an inversion of the encoder. In order to achieve $\hat{\boldsymbol{x}}$, the decoder only has to implement the inversion of the coding function implemented by the encoder. More concretely, if the encoder implements $\hat{\boldsymbol{h}} = \boldsymbol{f}(\boldsymbol{x})$, the decoder has to implement $\hat{\boldsymbol{x}} = \boldsymbol{f}(\hat{\boldsymbol{h}})$. But notice that the function implemented by the decoder, $\hat{\boldsymbol{x}} = \boldsymbol{f}(\hat{\boldsymbol{h}})$, is analogous to the function of the generative process, $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{h})$. This leads to the second way to characterize encoders and decoders in which the decoder is also referred to as a *generative model*: $\hat{\boldsymbol{h}}$ is a model of the causal factors of the probability distribution from where $\boldsymbol{x}$ is selected and, therefore, it allows for the decoder to generate more (and new) instances of $\boldsymbol{x}$. An important part of the contemporary efforts in machine learning is based on different uses of this notion of generative model (Hasson et al. 2020; Jaakkola & Haussler 1999; Salakhutdinov 2015).

The other consequence of understanding $\hat{\boldsymbol{h}}$ as a representation of the causal factors that generate $\boldsymbol{x}$ is that, in terms of communication systems, $\hat{\boldsymbol{h}}$ can be straightforwardly interpreted as the lookup table of the system. As we have noted, one main feature of autoencoders is that the function they implement is one that copies the input in the output. In this sense, they resemble communication systems in the sense that the latter aim for the receiver to get something like a copy of the message encoded by the sender, with compression, or noise, or sparse communication channels, which is why these systems are useful. To do so, communication systems need a lookup table that both sender and receiver know in advance. However, autoencoders learn that lookup table from the input! If we take

$\boldsymbol{x}$ to be what the sender puts in the communication system and $\hat{\boldsymbol{x}}$ to be what the receiver gets, the only thing needed for $\hat{\boldsymbol{x}} = \boldsymbol{x}$ or, at least, $\hat{\boldsymbol{x}} \approx \boldsymbol{x}$ is a proper $\hat{\boldsymbol{h}}$. Thus, $\hat{\boldsymbol{h}}$ is effectively the lookup table of the communication system. But, as autoencoders learn $\hat{\boldsymbol{h}}$ in an unsupervised fashion, it seems fair to say that they avoid Dennett's problem of the non-repayable loans of intelligence: autoencoders learn the lookup table ($\hat{\boldsymbol{h}}$) directly from the data ($\boldsymbol{x}$), so they do not need to ask for any loan of intelligence.

But do autoencoders truly avoid the problem of non-repayable loans of intelligence? Recall that the problem was taken to apply to all attempts to use Shannon's communication framework for understanding brains and neural networks. The loans of intelligence problem asks: where does the lookup table come from, in such systems, such that it is able to decode messages? For living systems, a popular suggestion is that the lookup table is provided by natural selection. On this view brains come pre-assembled with the necessary look-up tables: "The genetic program underlying brain development does the stipulations necessary to get the brain's many representational systems up and running" (Gallistel 2020, p. 393). Whatever the merits of such a view (and we suspect there aren't many; see, e.g. Anderson & Finlay 2014)[8], what we have just reviewed would appear to be an example of a system that can successfully capture and communicate arbitrary messages, while utterly lacking any evolutionary history.

This presents us with something of a trilemma: we can take autoencoders to be a counter-example to the loans of intelligence argument, paving the way toward a vindication of the Shannon framework for understanding the brain as well. Alternatively, if we wish to explain the success of autoencoders within the Shannon framework, but accept the validity of the loans of intelligence objection, we would need to identify *what* the loans of intelligence are, and how they are repaid. Finally, we can explore *other* frameworks to help us understand the success of autoencoders.

We will not pursue the first option here. We take the argument to be sound, and as we have seen above, even proponents of the Shannon framework for understanding the brain seem to accept their obligation to say whence comes the prior knowledge required to operate a neural communications system (Gallistel 2020; Gallistel & King 2010).

The second option is worth some attention, however, which we will provide immediately below. It is well-known that in many forms of machine learning the decisions and actions of the system designer can be vital to the success or failure of the system; data labeling, feature selection, restricting the solution space, hand-coding priors, etc., are all explicit loans of the designer's intelligence to the system. Perhaps there are some hidden loans in the case of autoencoders, as well.

---

[8] In fact, this is essentially a transcendental argument. Brains are communication systems. Communications systems need a lookup table to work. Brains work. Therefore they by necessity have a lookup table. It is a condition of the possibility of working brains. QED. The only respectably naturalized Deus ex Machina available to provide this lookup table is natural selection.

As we will shortly see, although any system designer inevitably encodes some of their intelligence in the system, the simplicity and generality of autoencoders make it unlikely that those loans are sufficient to explain their success. This leaves us with the third option, exploring alternate frameworks. We will discuss the third option in the next section.

3.3 Autoencoders: Debt-free communication systems?

Autoencoders are *unsupervised* learning systems. The notion of unsupervised learning has a very specific meaning in machine learning. Put simply, it means that the process of learning does not require specific examples of correct outputs to scaffold learning. A system is said to engage in unsupervised learning when it requires no *labeled* examples in the training set. Imagine a situation in which we train a feed-forward neural network with a training set of pictures and we expect the network to classify those pictures in two different categories: "cat" and "dog". In this context, when we provide an input $\boldsymbol{x}$ to the feed-forward neural network, we expect an output $\boldsymbol{y}$ that is either "cat" or "dog". A *supervised* learning strategy in this situation would be one in which, for (at least) some of the inputs $\boldsymbol{x}$, the trainer of the feed-forward neural networks provides explicit feedback on what the output $\boldsymbol{y}$ should be. Thus, for at least some inputs, the feed-forward neural network would get not just an input picture, but also explicit guidance on whether the picture falls into the category "cat" or the category "dog". In this sense, the picture is labeled with the right category.

In the context of machine learning, unsupervised learning strategies are those that do not follow the described supervised strategy. Therefore, unsupervised learning does not entail that a given feed-forward network engages in fully autonomous learning, but just that the learning strategy does not make use of labeled examples in the training set. Labels are indeed good examples of a form of agreed prior knowledge between sender and receiver for the receiver to know that it is getting the right message. The labels constitute the intelligence loaned to the system that would allow it to to learn its own lookup table, because the labels provide the otherwise inaccessible knowledge of the correspondence between signal and message. In a supervised learning system, we could identify the input as a form of sender, the feed-forward neural network as a form of channel, and the output as a form of receiver. It is clear that labeled examples would play the role of prior knowledge used by the receiver (output) to learn the right mapping between the signal sent by the sender (input) and the message. In this sense, an unsupervised learning system as an autoencoder that does not use labeled examples might be thought of as a successful communication system without a lookup table. But does having no need for a pre-specified lookup table mean having no need to take out loans of intelligence? To what extent does this consequence follow from unsupervised learning? Is it true that autoencoders need no prior knowledge to learn?

As we have noted in the previous section, the unsupervised learning algorithm of autoencoders rests on the minimization of a loss function. Namely, autoencoders learn to

copy the input $\boldsymbol{x}$ into the output $\hat{\boldsymbol{x}}$ by minimizing the difference between them (i.e., by losing as little as possible of $\boldsymbol{x}$ in $\hat{\boldsymbol{x}}$). In its common interpretation, this minimization of the loss function ensures autoencoders encode a representation $\hat{\boldsymbol{h}}$ that captures the set of causal factors $\boldsymbol{h}$ that generate $\boldsymbol{x}$.[9] And with this representation $\hat{\boldsymbol{h}}$ the autoencoder is able to generate $\hat{\boldsymbol{x}}$ such that $\hat{\boldsymbol{x}} = \boldsymbol{x}$ or, at least, $\hat{\boldsymbol{x}} \approx \boldsymbol{x}$. The question, then, is whether autoencoders need prior knowledge in order to perform this minimization of the loss function. And the answer seems to be affirmative: although not requiring explicit knowledge in the sense of labeled examples in the training set, *many* autoencoders need some amount of prior knowledge and *all* autoencoders rest on some assumptions that can be understood as prior knowledge about the real distribution of the data.[10]

An example of autoencoders that require some amount of prior knowledge to minimize their loss function are *variational autoencoders* (VAEs; Kingma & Welling 2019). VAEs are autoencoders that engage in a form of approximate Bayesian inference to make $\hat{\boldsymbol{h}}$ similar to $\boldsymbol{h}$.[11] To do so, VAEs minimize a loss function based on a quantity known as evidence lower bound (ELBO) or negative variational free energy. To harness the minimization of this loss function, however, VAEs are required to rely on prior knowledge; concretely, in knowledge about the joint probability between $\boldsymbol{h}$ and $\boldsymbol{x}$. This joint probability can be learnt by VAEs and, actually, VAEs may be thought as devoted to learning it.[12] However, the joint probability is needed from the very beginning of the learning process and, therefore, even in some random fashion (i.e., even making it as assumption-less as possible), some assumptions must be made regarding its metrics, its structure, and its relevant variables. These assumptions effectively work as prior knowledge regarding the kind of input $\boldsymbol{x}$ and generative process, $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{h})$, the autoencoder must discover. In this sense, although not in the fully explicit fashion of labeled data, the assumptions VAEs make about the joint probability of $\boldsymbol{h}$ and $\boldsymbol{x}$ work as some form of prior knowledge.

These assumptions in the case of VAEs are on top of general assumptions that are made in the case of all autoencoders and, in wider terms, in the case of all representational

---

[9] We do not need too many details on this process but, conceptually speaking, by minimizing the loss function, autoencoders find the set of parameters $\boldsymbol{\theta}$ that make $\hat{\boldsymbol{h}}$ most similar to $\boldsymbol{h}$, such that $\hat{\boldsymbol{h}}(\boldsymbol{x}; \boldsymbol{\theta}) \approx \boldsymbol{h}$.

[10] In this argument we are ignoring the obvious fact that these systems do not learn their own learning algorithm. The loans of intelligence objection deals with need for pre-existing knowledge, not preexisting functional structure.

[11] As Bayesian system, $\hat{\boldsymbol{h}}$ in VAEs stands for a probability distribution known as the *recognition density* and $\boldsymbol{h}$ stands for the *true posterior* of the Bayes theorem that relates $\boldsymbol{h}$ and $\boldsymbol{x}$. We do not need these details for our current purposes; for the sake of consistency we have stuck with the notation that we have used above. Also, VAEs are not the only Bayesian autoencoders but serve as an illustration of all of them.

[12] Statistically speaking, the joint probability $p(\boldsymbol{h}, \boldsymbol{x})$ is known as a *generative model* insofar as all possible factors of variance $\boldsymbol{h}$ of all possible $\boldsymbol{x}$ are included in them. In a sense, when a VAE learns a proper $\hat{\boldsymbol{h}}$, its decoder effectively encodes a joint probability $p(\hat{\boldsymbol{h}}, \hat{\boldsymbol{x}})$ and, therefore, the decoder of the VAE becomes a generative model in the sense advanced in section 3.1.

learning systems. Bengio et al. (2013) provide a non-exhaustive list of these general assumptions. Some of them are, for instance, that the changes in the function learnt by these systems are *smooth*, that the relationship between (at least some of these) variables is *linear*, that these variables are *independent*, that the causal factors explaining data are *sparse*, or that most relevant causal factors exhibit a *slow temporal change*. As Goodfellow et al. (2017) claim, these assumptions are hints about the underlying factors of the unlabeled data and "take the form of implicit *prior beliefs* that we, the designers of the learning algorithm, impose in order to guide the learner." (p. 544-545—emphasis added).[13]

Clearly, autoencoders are not truly blank slates with no debts owed. But could *any* learning system be like that? We doubt it. The question here is whether the kinds of prior knowledge identified above are of the sort to which the loans of intelligence objection applies. If so, then far from being a counterexample to the loans of intelligence objection, autoencoders would be a working system in which the loans have been identified and repaid, and this objection to treating them (and by extension, cognitive systems more generally) within the Shannon framework would have been cleared. But if not, we still have to choose between denying the validity of the loans of intelligence objection, and finding an alternate way of understanding these systems.

A first clue about the right path to choose comes from considering the persisting disanalogy between lookup tables and the sorts of assumptions identified above. A lookup table is an explicit, stored, data structure representing the range of options (message) in a given context. If anything qualifies as domain knowledge, a lookup table would. Now, turn to the sorts of assumptions listed above. Some seem to be more naturally understood as parameter settings that adjust the operation of the learning processes to best handle the data it is exposed to. Although that setting is knowledge for the system designer, it is less clear that it is knowledge for the system. It's just part of its operation, never becoming part of the content of any message that the system might be passing around. Similarly, although it is common to call the assumptions underlying particular learning mechanisms "implicit beliefs", there is an important disanalogy to be noted: beliefs are representations *for the system*, used directly by the system, and operated on by the system. In contrast, assumptions of the sort noted are simply characteristics of the (data) environment within which the system operates optimally, and outside of which it will start to fail. As Warren (2005) notes in a related context:

> Perceptual systems become attuned to informational regularities in the same manner that other systems adapt to other sorts of environmental regularities (such as a food source): possessing the relevant bit of physiological plumbing (whether an enzyme or a neural circuit) to exploit a regularity confers a

[13] The importance of these hints cannot be overstated. For instance, these hints define a model family for $\hat{\boldsymbol{h}}$. If the distribution of $\boldsymbol{x}$s from where the training set is selected does not lie in the same model family, it cannot be properly estimated (see Goodfellow et al. 2017, p. 131).

> selective advantage upon the organism. Since the water beetle larva's prey floats on the surface of the pond and illumination regularly comes from above, possession of an eye spot and a phototropic circuit can enhance survival and reproductive success. But if illumination were ambiguous and prior knowledge were required to infer the direction of the prey, it is not clear how such a visual mechanism would get off the ground. Natural selection converges on specific information that supports efficacious action.
>
> What the [traditional] view treats as assumptions imputed to the perceiver can thus be understood as *ecological constraints* under which the perceptual system evolved. The perceptual system need not internally represent an assumption that natural surfaces are regularly textured, that terrestrial objects obey the law of gravitation, or that light comes from above. Rather, these are facts of nature that are responsible for the informational regularities to which perceptual systems adapt, such as texture gradients, declination angles, and illumination gradients. They need not be internally represented as assumptions because the perceptual system need not perform the inverse inferences that require them as premises. The perceptual system simply becomes attuned to information that, within its niche, reliably specifies the environmental situation and enables the organism to act effectively. (p. 357-8)

What this suggests is that autoencoders are in fact *not* taking out loans of intelligence, or at least it is not clear whether that's the right way to describe the situation. But if they are not, then what explains their success? We think they do not require loans of intelligence because they are not in fact communication systems of the sort Shannon describes (although they could of course be deployed in a communication system setting) and that Gallistel & King (2010), among many others, extend to the brain. Instead, we believe that they may belong to a family of "direct fit" algorithms that can capture regularities in input and that we do not need to describe them as building generative models of those data, nor as having pre-specified knowledge about what messages those data might contain. If that's right, it would align autoencoders more closely with the ecological understanding of information detailed by J. J. Gibson (1979, 1966) than with Shannon's notion of information outlined above.

4 An ecological take on autoencoders

Cognitive neuroscientists have assumed that the information available to animals, their senses, and their brains is best analyzed in the terms Shannon (1948) developed for understanding (and building) communication systems. There is, however, an alternative tradition that is nearly as old, originating in Gibson (1950) and revised and elaborated in later work. Gibson rejected the signal processing view of perception (and with it the

analogy between brains and computers) in favor of a theory of direct perception; that is, perception unmediated by sensations and sense data, and not requiring the construction of world models. Instead, Gibson postulated that animals could directly detect and adaptively respond to information available in what he called the ambient optic array. That information wasn't to be understood as signals needing to be decoded, as messages for the brain to decipher, but rather as structure in light that specifies the properties of the environment. He writes:

> There are currently two radically different usages of the word "information" in psychology. One I will call afferent input information and the other optic-array information. The former is familiar; it is information conceived as impulses in the fibers of the optic nerve. Information is assumed to consist of signals, and to be transmitted from receptors to the brain. Perception is a process that is supposed to occur in the brain, and the only information for perception must therefore consist of neural inputs to the brain.
>
> Optic-array information is something entirely different. It is information in light, not in nervous impulses. It involves geometrical projection to a point of observation, not transmission between a sender and a receiver. It is outside the observer and available to him, not inside his head. In my theory, perception is not supposed to occur in the brain but to arise in the retino-neuro-muscular system as an activity of the whole system. The information does not consist of signals to be interpreted but of structural invariants which need only be attended to. (Gibson 1972, p. 79)

Gibson refers to structures in energy arrays which he called invariants, i.e., patterns in light, sound, etc. that an organism can, in principle, detect, and that are potentially informative for the organism about some structure in the world. An example of an invariant is *tau*, which is related to the perceptible rate of optical expansion of an approaching object, and directly specifies time-to-contact. Another example is the horizon-ratio relation: the ratio of the amount of an object that appears to be above the cut of the horizon to that which is below the horizon. Because the horizon is always exactly at the observer's eye-level, the horizon-ratio relation specifies the height of the object in terms of the height of the observer's eyeballs above the ground (Bootsma & Oudejans 1993; Sedgewick 1973). The notion is that there exist many such regularities in the ambient array, and animals learn to use them to perceive. Animals' attunement to such invariant regularities enables or constitutes veridical perception of the world (Segdwick 2021; Turvey 2019). This is an important contrast with the traditional view. As Gibson (1972) notes:

> It has long been assumed by empiricists that the only information for perception was "sensory" information. But this assumption can mean different things. If it

> means that the information for perception must come through the senses and not through extrasensory intuition, this is the doctrine of John Locke, and I agree with it, as most of us would agree with it. But the assumption might mean (and has been taken to mean) that the information for perception must come over the sensory nerves. This is a different doctrine, that of Johannes Müller, and with this we need not agree. To assume that visual information comes through the visual sense is not to assume that it comes over the optic nerve, for a sense may be considered as an active system with a capacity to extract information from obtained stimulation. The visual system in fact does this. Retinal inputs lead to ocular adjustments, and then to altered retinal inputs, and so on. It is an exploratory, circular process, not a one way delivery of messages to the brain. (p. 80)

One important aspect of Gibson's position that is worth bringing to the fore is its endorsement of the claim that there are no *epistemic* mediators in perception, no content-carrying packets in need of decoding, interpretation, or association, be they sensations, sense impressions, sense data, or any of the panoply of ontological posits that Merleau-Ponty collectively dismissed as the imaginary "pointillistic impacts" of the world. There are, of course, myriad *causal* mediators for perception; it is only our immersion in a world of causes that makes perception possible. But the causal process of perception does not involve reconstructing the outside world from its momentary impacts on the sense organs, but the active sampling of and adjusting to or coupling with the structure in energy arrays —light, sound, chemical gradients, and the like. It is this causal coupling, which Gibson called "resonance", that underlies our direct perception of the world. He writes:

> Instead of supposing that the brain constructs or computes the objective information from a kaleidoscopic inflow of sensations, we may suppose the orienting of the organs of perception is governed by the brain so that the whole system of input and output resonates to the external information. (Gibson 1966, p. 5)

What is it to resonate to information? Raja (2019, 2021) analyzes the concept in terms of the lawful fit between neural activity and the information contained in the optic (or other) array. More specifically, Raja posits that there is a dynamic coupling between the central nervous system and the environment such that properties of the neural activity reflect properties of the environment, allowing for adaptive behavior. This coupling constitutes the "fit" between the brain and the world. Put differently, what perceptual systems do is not reconstruct the world from irritations at the exteroceptive surfaces, but seek out and adjust to structure in energy arrays. This structure is the ecological information to which animals attune.

This conception of ecological information as environmentally accessible structure to which an organism can adjust or attune is an important alternative to Shannon information, and potentially opens up new ways of understanding not just brains (Anderson 2014; Raja & Anderson 2019) but also autoencoders and other artificial neural networks. Such alternate approaches to understanding these systems is especially important if you agree with such authors as Brette (2019) and Nizami (2019) that the coding metaphor faces significant conceptual difficulties in the neurosciences (and the cognitive sciences more broadly). Stated up front, the idea is this: what autoencoders are able to do so well is adapt to real structure in the input data. Autoencoders achieve this by adjusting the connection weights in their network so as to capture the structure in the input.

Importantly, capturing real structure does *not* need to mean extracting a function or building a generative model (much less recovering a lookup table). Instead, all that is required is a brute-force "direct fit" to the high-dimensional structure in the inputs. Hasson et al. (2020) develop this thought in great detail. Classically, neural networks are understood as universal function approximators, able to learn any arbitrary mapping between input and output with enough time, data and connections (e.g., Hornik et al. 1989; Zhou 2020). While not questioning the truth of the universal approximation theorem, Hasson et al. (2020) question whether function learning is the most productive analogy for understanding these systems:

> We argue that neural computation is grounded in brute-force direct fitting, which relies on over-parameterized optimization algorithms to increase predictive power (generalization) without explicitly modeling the underlying generative structure of the world. (p. 418)

The key word in the quote is "explicitly". They do not deny that one can *model* the action of the network as implementing a generative function; they deny this is the only way to understand what it is doing. What makes the function learning perspective attractive, they argue, is the textbook view of machine learning in which overparameterized models are associated with overfitting, and thus poor generalization. This can be true in contexts in which training data offers only narrow coverage of the domain space (say, faces); there successful generalization to novel faces depends on successful extrapolation, and overfit models do not extrapolate well. However, if there is sufficient coverage of the domain, interpolation—which over-parameterized models do perfectly well—can render extrapolation unnecessary for generalization. They write:

> We contend that [the] textbook view should be revised to account for the fact that in a data-rich setting, over-parameterized models can provide a mindless yet powerful form of generalization. Any model is designed to solve a particular type of problem, and the problem to be solved changes drastically when we shift

> from preferentially sampling a limited parameter space in a controlled experimental setting to densely sampling a wide parameter space using big data in a performance-oriented real-life setting. (Hasson et al. 2020, p. 418)

It has long been a tenet of the ecological framework that organisms richly sample their environments, wherein they encounter sufficient information to specify the objects and properties of the world without the need for building complex inner models (Blau & Wagman 2022; Lobo et al. 2018; Warren 2021). This is in stark contrast to the view commonly pushed in more conventional cognitive science that the information available to the organism is sparse and impoverished. Where the available data *are* sparse, as in the early days of machine learning, it is natural to explain successful generalization in terms of generative models. In rich environments, however, this explanation is unnecessary.

> Dense sampling of the problem space can flip the problem of prediction on its head, turning an extrapolation-based problem into an interpolation-based problem… interpolation uses local computations to situate novel observations within the context of past observations; it does not rely on explicit modeling of the over-arching generative principles. Unlike extrapolation, interpolation was thought to provide a weak form of generalization because it can only predict new data points within the context of past observations… But this problem only arises if the scope of the training space is small or impoverished. (Hasson et al. 2020, p. 418-19)

Insofar as this is true, it seems most natural to understand the capacities of both real brains and autoencoders in terms of their iterative adjustment to the available structure in their environments. Neither real brains nor autoencoders require loans of intelligence because they are not dealing with Shannon information, but are instead tracking and adjusting to ecological information. In a word, they resonate to it.

It would take much more work than we can put in here to establish that the ecological framework offers a better way to understand real brains and artificial ones, but that effort is well underway (Anderson 2014; Bruineberg & Rietveld 2019; deWit & Withagen 2019; Fultolt et al. 2019; Raja 2018, 2019, 2021; Raja & Anderson 2019; Segundo-Ortin & Hutto 2021; van Dijk & Myin 2019). Here we will instead point to some future directions for this overarching project.

5 Open questions and future directions

Testing the *ecological hypothesis* for autoencoders—and for real and artificial brains more generally—involves at least two different challenges. The first challenge concerns the notion of ecological information itself. If both brains and artificial neural networks adjust or

resonate to this information, a first level of analysis must necessarily address the adequate description of the information in the structure of the ambient energy arrays surrounding an organism or in the structure of the data used to train an artificial neural network. This is part of the work on perception and behavior developed by ecological psychologists during the last decades. Variables of ecological information (i.e., invariants of the ambient energy arrays) have been described in different behavioral situations. Famous illustrations of this are the study of the already mentioned ecological variable *tau* in looming situations (Lee 2009), the investigation of different invariants of the optic flow to perceptually guide navigation both in sparse and crowded environments (Warren et al. 2001; Warren 2018), or the description of many other informative aspects of different ambient energy fields in general (Turvey 2019; Warren 2021). Advances regarding our understanding of ecological information provide the necessary background to fully capture the processes of resonance both in real and in artificial brains.

More concretely, a proper understanding of the ecological information in the training datasets is crucial to gather evidence regarding the possible resonant activity of autoencoders. Such an understanding will minimally involve a mathematical description of the ecological information in the dataset and a way to relate the activity of the autoencoder to it. Detailed mathematical descriptions of ecological information are usual in the ecological literature. Beyond the informational invariants just mentioned, for instance, Tsao & Tsao (2021) have used differential topology to offer a general mathematical account of *ecological optics*—i.e., the description of the ambient optic array in terms of its ecological structure—and the invariants of object segmentation and ocludding edges. They have also provided an algorithm to detect these invariants. The mathematical description offered by Tsao & Tsao (2021) might be used to characterize different objects and object-relations of different training datasets used on autoencoders aimed at object identification, for example. This way researchers would know which invariants are in each dataset and, therefore, which invariants a given autoencoder should be adjusting to. Interestingly, this logic could be reverted and autoencoders might be used to explore the space of possibilities in the structure of training datasets. If the datasets are naturalistic enough, the results may be used to identify possible variables of ecological information that could potentially be tested in behavioral studies at a later stage of research.

Once the ecological information is well described, the second challenge for the development of the ecological hypothesis is the accurate characterization of the very activity of autoencoders. If we follow Hasson et al. (2020), above, we will be in a direct fit environment with a rich dataset. In this context, the activity of the autoencoder is described in terms of interpolation by local adjustments and not extrapolation by learning a generative model of the dataset. In other words, the autoencoder would not be (described as) learning a function of the data but resonating to the structure available in the dataset. A further step would be to pursue the concrete characterization of this framework in terms compatible with ecological information. Many of the invariants described in the ecological

literature entail some form of transformation—i.e., they literally are invariant in the mathematical sense of being "invariant under a transformation". In this sense, detecting invariants involves the training dataset must be dynamic: each data sample must have a temporal dimension (e.g., a video instead of a snapshot). This kind of temporal-input approach is already being developed in the machine learning literature (e.g., Zhu et al. 2017) and it is for sure what is needed to detect invariants like *tau*, for instance. The question now would be how to relate an informational invariant like *tau* with the idea of interpolation. The relationship between interpolation and categorization, for example, seems to be very straightforward, but in the case of informational invariants it is potentially more obscure—e.g., what are we interpolating in a looming situation? Variables of ecological information? Maybe appropriate actions? Perhaps both? These are open questions that require further work under the ecological hypothesis for autoencoders. We are not in the position to answer them right now, but only to raise them. However, they are concrete enough to allow for relatively straightforward implementation given the state-of-the-art in machine learning and, therefore, to allow for future directions of research.

References

Anderson, M. L. (2014). *After phrenology: Neural reuse and the interactive brain*. Cambridge, MA: MIT Press.

Anderson, M. L., & Champion, H. (2021). Some dilemmas for an account of neural representation: A reply to Poldrack. [Preprint] http://philsci-archive.pitt.edu/id/eprint/20003

Anderson, M. L. & Finlay, B. L. (2014). Allocating structure to function: The strong links between neuroplasticity and natural selection. *Frontiers in Human Neuroscience*, *7*, 918, 1-16.

Attneave, F. (1954). Some information aspects of visual perception. *Psychological Review*, 61, 183–193.

Baldi, P. (2012). Autoencoders, unsupervised learning, and deep architectures. *Proceedings of ICML Workshop on Unsupervised and Transfer Learning, Proceedings of Machine Learning Research*, 27: 37-49.

Bengio, Y., Courville, A., & Vincent, G. N. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 1798-1828.

Blau, J. J., & Wagman, J. B. (2022). *Introduction to Ecological Psychology: A lawful approach to perceiving, acting, and cognizing*. New York: Routledge.

Bootsma, R. J., & Oudejans, R. R. (1993). Visual information about time-to-collision between two objects. *Journal of experimental psychology: human perception and performance*, *19*(5), 1041-1052.

Brette, R. (2019). Is coding a relevant metaphor for the brain? *Behavioral and Brain Sciences*, *42*, e215: 1-58.

Bruineberg, J. & Rietveld, E. (2019). What's inside your head once you've figured out what your head's inside of. *Ecological Psychology*, 31(3), 198-217, DOI: 10.1080/10407413.2019.1615204

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

Dennett, D. C. (1981). *Brainstorms: Philosophical essays on mind and psychology*. MIT Press.

de-Wit, L., Alexander, D., Ekroll, V., & Wagemans, J. (2016). Is neuroimaging measuring information in the brain?. *Psychonomic bulletin & review*, *23*(5), 1415-1428.

de Wit, M. M., & Withagen, R. (2019). What should a "Gibsonian neuroscience" look like? Introduction to the special issue. *Ecological Psychology*, *31*(3), 147-151.

Foffani, G., et al. (2009). Spike timing, spike count and temporal information for the discrimination of tactile stimuli in the rat ventrobasal complex. *The Journal of Neuroscience*. 29(18), 5964-5973.

Fultot, M. Frazier, A.P., Turvey, M.T. & Carello, C. (2019). What are nervous systems *for*? *Ecological Psychology*, 31(3), 218-234, DOI: 10.1080/10407413.2019.1615205

Gallistel, C. R. & King, A. P. (2010). *Memory and the computational brain: Why cognitive science will transform neuroscience*. New York: Wiley-Blackwell.

Gallistel, C. R. (2019). Our understanding of neural codes rests on Shannon's foundations. *Behavioural and Brain Sciences*, 42, 25-26.

Gallistel, C. R. (2020). Where meanings arise and how: Building on Shannon's foundations. *Mind & Language*, *35*(3), 390-401.

Gibson, J.J. (1972). "A theory of direct visual perception." In J. R. Royce and W. W. Rozeboom, eds., *The Psychology of Knowing*, 215–240. New York: Gordon & Breach. Reprinted by permission of the publisher.

Gibson, J. J. (1979). *The ecological approach to visual perception*. New York: Psychology Press.

Gibson, J. J. (1950). *The perception of the visual world*. Boston: Houghton Mifflin.

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

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.

Hinton, G. E., & van Camp, D. (1993). Keeping the neural networks simple by minimizing the description length of the weights. *COLT '93: proceedings of the sixth annual conference on computational learning theory*, 5-13.

Hinton, G. E., & Zemel, R. S. (1993). Autoencoders, minimum description length, and Helmholtz free energy. *NIPS' 93: proceedings of the 6th international conference on neural information processing systems*, 3-10.

Hohwy, J. (2013). *The predictive mind*. Oxford: Oxford University Press.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, *2*(5), 359-366.

Jaakkola, T. S., & Haussler, D. (1999). Exploiting generative models in discriminative classifiers. *Advances in neural information processing systems*, 487-493.

Kingma, D. P. & Welling, M. (2019), An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 12(4), 307-392. http://dx.doi.org/10.1561/2200000056

Kravitz, D. J., et al. (2011). Real-world scene representations in high-level visual cortex: It's the spaces more than the places. *The Journal of Neuroscience*, 31(20), 7322-7333.

Lee, D. N. (2009). General tau theory: Evolution to date. *Perception*, 38(6), 837–858.

Lobo, L., Heras-Escribano, M., & Travieso, D. (2018). The history and philosophy of ecological psychology. *Frontiers in Psychology*, 9, 2228, 1-15.

MacKay, D. M. & McCulloch, W. S. (1952). The limiting information capacity of a neuronal link. *Bulletin of Mathematical Biophysics*, 14(2), 127–135.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press.

Nemenman, I., et al. (2008). Neural coding of natural stimuli: Information at sub-millisecond resolution. *PLoS Computational Biology*, 4(3), 1-12.

Nizami, L. (2019). Information theory is abused in neuroscience. *Cybernetics & Human Knowing*, *26*(4), 47-97.

Perkel D. & Bullock T. (1968). Neural coding: A report based on an NRP work session. *Neuroscience Research Program Bulletin 6*. MIT Press.

Raja, V., & Anderson, M. L. (2019). Radical embodied cognitive neuroscience. *Ecological Psychology*, *31*(3), 166-181.

Raja, V. (2018). A theory of resonance: Towards an ecological cognitive architecture. *Minds & Machines*, 28(1), 29–51. https://doi.org/10.1007/s11023-017-9431-8

Raja, V. (2019). From metaphor to theory: The role of resonance in perceptual learning. *Adaptive Behavior*, 27(6), 405-421.

Raja, V. (2021). Resonance and radical embodiment. *Synthese*, 199(1), 113–141. https://doi.org/10.1007/s11229-020-02610-6

Rapoport, A., & Horvath, W. J. (1960). The theoretical channel capacity of a single neuron as determined by various coding systems. *Information and Control*, 3(4), 335–350.

Rieke, F., et al. (1997). *Spikes: Exploring the neural code*. Cambridge, MA: MIT Press.

Salakhutdinov, R. (2015). Learning deep generative models. *Annual Review of Statistics and Its Application*, 2(1), 361-385.

Sedgwick, H. A. (2021). JJ Gibson's "ground theory of space perception". *i-Perception*, *12*(3), 1-55.

Sedgwick, H. A. (1973). The visible horizon: A potential source of visual information for the perception of size and distance (Doctoral Dissertation, Cornell University, 1973). Dissertation Abstracts International, 34, 1301B–1302B (University Microfilms No. 73-22530)

Segundo-Ortin, M., & Hutto, D. D. (2021). Similarity-based cognition: Radical enactivism meets cognitive neuroscience. *Synthese*, 198(1), 5-23.

Series, P., et al. (2004). Tuning curve sharpening for orientation selectivity: Coding efficiency and the impact of correlations. *Nature Neuroscience*, 7(10), 1129-1135.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379-423.

Shannon, C. E. (1956). The bandwagon. *IRE Transactions in Information Theory*, 2(1), 3.

Stone, J. V. (2012). *Vision and brain: How we perceive the world*. Cambridge, MA: MIT Press.

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

Turvey, M. T., et al. (1981). Ecological laws of perceiving and acting: In reply to Fodor and Pylyshyn. *Cognition*, 9(3), 237-304.

Turvey, M. T. (2019). *Lectures on perception: An ecological perspective*. Routledge.

van Dijk, L. & Myin, E. (2019). Ecological neuroscience: From reduction to proliferation of our resources. *Ecological Psychology*, 31(3), 254-268, DOI: 10.1080/10407413.2019.1615221

Warren, W. H. (2018). Collective motion in human crowds. *Current Directions in Psychological Science*, *27*(4), 232–240. https://doi.org/10.1177/0963721417746743

Warren, W. H. (2005). Direct perception: the view from here. *Philosophical Topics*, 33(1), 335-361.

Warren, W. H. (2021). Information is where you find it: Perception as an ecologically well-posed problem. *I-Perception*, 12(2), 1-24. https://doi.org/10.1177/20416695211000366

Warren, W. H., Kay, B. A., Zosh, W. D., Duchon, A. P., Sahuc, S. (2001). Optic flow is used to control human walking. *Nature Neuroscience*, 4(2), 213–216.

Weaver, W. (1953). Recent contributions to the mathematical theory of communication. *ETC: A Review of General Semantics*, 10(4), 261-281.

Zhou, D. X. (2020). Universality of deep convolutional neural networks. *Applied and computational harmonic analysis*, 48(2), 787-794.

Zhu, J., Zou, W., and Zhu, Z. (2017). End-to-end video-level representation learning for action recognition. *arXiv*, 1711.04161. https://arxiv.org/abs/1711.04161