Representations in Biological & Artificial Cognition

Rajsuryan Singh
Music Technology Group, Universitat Pompeu Fabra
rajsuryan.singh01@estudiant.upf.edu

***Abstract*—Deep learning has been indebted to psychology and neuroscience ever since its inception. However, recently, deep neural networks are being used as models for the brain motivating a comparison between different aspects of biological and artificial cognition. Internal representations provide an avenue for such a comparison. In the following sections of this paper, the idea of representations is elaborated upon from the perspective of philosophy, neuroscience, and artificial intelligence. Further, different dimensions for comparison of representations have been discussed, namely behavioural, structural, and functional characteristics of cognitive systems as well as the mode of acquisition of representations. These aspects provide an orthogonal and complementary set of bases for comparison of biological and artificial systems. The paper is concluded with a brief analysis of state of the art audio-visual representations using the proposed framework of analysis.**

I. Introduction

Deep learning has been indebted to psychology and neuroscience ever since its inception and ground-breaking advancements have often come about as a consequence of that influence - the deep learning literature is rife with synonyms of the phrase “brain-inspired”. Convolutional neural networks, for instance, were inspired by Hubel and Wiesel’s work on receptive fields and hierarchical processing in the visual cortex [1] and have led to human-level performance on a number of computer vision tasks. The goals of deep learning, however, are not just to mimic intelligence but also to help us understand it by the virtue of having created it in-silico. Deep neural networks, if used as models for the brain, give us third person access to first person states[1]. In this business, it has become customary to quote Richard Feynman - “What I cannot create, I do not understand”. But when the subject of one’s inquiry is the mind, the question of what it means to create it is not trivial. The ability to build systems that behaviourally resemble the brain or its parts is not a reason for complacency but rather a push towards analysing, with depth and nuance, the artificial systems against the biological ones that they mimic.

Internal representations provide an avenue for such a comparison between biological and artificial cognition. In the following sections of this paper, the idea of representations is elaborated upon from the perspective of philosophy, neuroscience, and artificial intelligence. Further, different dimensions for comparison of representations have been discussed, namely behavioural, structural, and functional characteristics of cognitive systems as well as the mode of acquisition of representations. These aspects provide an orthogonal and complementary set of bases for comparison of biological and artificial systems. The paper is concluded with a brief analysis of state of the art audio-visual representations using the proposed framework of analysis.

II. Representations

A. Representations in Mind and Brain

The word “representation” is indispensable in the common parlance of psychology, neuroscience, artificial intelligence, and philosophy of mind. But even a cursory glance at the literature reveals the dissonance between the usage of the word in the different disciplines. Since a common ground is necessary for a meaningful discussion, the following is a brief description of the different connotations of the word followed by an attempt to narrow it down for the purposes of this paper. There seem to be as many definitions of the word as there are philosophers of mind, but one can circumvent the need to open that can of worms by just reading the label i.e. summarising the representational theory of mind (RTM). RTM posits the existence of semantically evaluable intentional mental states capable of being combined and manipulated according to logical (or otherwise truth-respecting) rules. That is to say that these mental states are about things and they can be evaluated based on their semantic content. Such states are defined in relation to mental representations which by definition have to be semantically interpretable.

The neuroscientific definition is not as much a subject of debate as its philosophical counterpart. In neuroscience, representations are used to describe the empirical relationships between neural activity and features of the world [3]. The representation can be at different scales and levels of abstraction. For example, an image of a face is represented at a low level of abstraction in early visual cortex through Gabor-like filters but in highly abstract and specialised manner in later stages of visual processing. An extreme end of specialisation would be Lettvin’s idea of grandmother cells[4] where single neurons correspond to specific people. But the idea remains the same - there is a functional relationship between stimuli and neural activity such that the stimuli is represented by the activity of neurons.

B. Representations in Artificial Intelligence

The computational theory of mind (CTM) conceptualises intelligence as information processing i.e. computations over

[1] This phrase is stolen from Paul but I wasn’t quite sure how to cite something I heard in a lecture.

representations. Representations are information-bearing entities that can be operated on or processed to give rise to other representations and all mental states can be described in relation to such representations. This reminiscence of RTM is not coincidental as CTM has its roots in philosophy and is built upon RTM. And as one would expect from an idea with philosophical underpinnings, it is not free of debate; there is a sharp divide between the classicist and connectionist takes on CTM about the nature of representations and computations that it entails. The classicist view is exemplified by Alan Turing's model of computing[5] where representations are symbolic and there is a strict mapping between the symbols and the things that they represent. The computations in this model are discrete and are carried out by a central processing unit according to a set of instructions. Intelligence is hence reduced to mere symbol manipulation.

Despite the reductionism, or perhaps because of it, this paradigm of artificial intelligence was extremely fruitful in the early days of AI with programs convincingly mimicking intelligent behaviour like playing chess [6] and proving mathematical theorems[7]. The success wasn't limited to problems with low-dimensional and finite state spaces; strides were made in areas like computer vision (see [8] for a computational theory of interpretable visual representations by David Marr) and natural language processing (NLP). In 1957, Chomsky proposed a formalism for rule-based descriptions of syntactic structure [9], at a time when rule-based systems like Newel's "Logic Theorist"[7] and "General Problem Solver"[10] were gaining popularity and programming languages like LISP were being developed. The synergy between these circumstances led to great advancements in NLP with systems like SHRDLU [11] and LIFER/LADDER [12] demonstrating human-like language abilities. The linguistic bias in the classicist approach is particularly interesting because it seems to be carrying a remnant of early experimental psychology in the assumption that representations, which can be thought of as the fundamental elements of cognition and thought, are introspectable and hence have semantic qualities to them. This conserves the semantic interpretability of representations and manipulations thereof originally proposed in RTM.

The connectionist view on the other hand, engendered from McCulloch & Pitts' work that demonstrated the possibility of neurons implementing logical operations [13], does away with the assumption of semantically interpretable representations. It is proposed that representations, instead of being localised in memory as in the classical framework, are distributed and are realised as characteristic activation patterns of neurons and such neural embeddings don't lend themselves very well to semantic analysis. Two ideological differences seem to emerge between connectionist and classical takes on computation - considering Marr's three levels of analysis, there is a shift in focus from the algorithmic level in the classical approach to the implementation level in the connectionist counterpart. Moreover, connectionism doesn't assume organisms to be Turing machines, so the substrate of the computations becomes relevant and since no two organisms can have identical neuronal connections, there is a shift towards functionalism where representations are defined by the function they serve for the organism and not their precise internal constituents.

While McCulloch and Pitts, among others, continued to probe the nature of connectionist representations in biological systems [14-16], Frank Rosenblatt, in the spirit of the aforementioned Feynman quote, built a physical system that put these ideas into practice - the perceptron [17]. He formalised a probabilistic framework to, in his words, "answer the questions of how information about the physical world is sensed, in what form is information remembered, and how does information retained in memory influence recognition and behavior". Representations, being the information-bearing entities of the mind, were at the heart of his inquiry and his work demonstrated a concrete mechanism for the acquisition of representations opening the proverbial black box of the mind. However, his ideas were soon put under severe, and somewhat unjustified, scrutiny by Minsky and Papert in their book Perceptrons [18] where they not only criticized the perceptron for it being limited to linear functions, but also called its extension to multilayer systems sterile. While the criticism of the perceptron in its original form was fair, their claims on multilayer neural networks clearly didn't stand the test of time as the multilayer perceptron has been the bedrock for the deep learning revolution.

The connectionist form of computing has seen a remarkable resurgence in deep learning and has taken AI by storm rivalling, and in some cases even surpassing, human level performance in computer vision[19] and audition[20], natural language processing[21], sensory prediction[22], game playing [23] and reasoning [24]. These advancements, while being enabled by an exponential growth in high quality data and computational resources, are driven by the same principles as the perceptron - error based adjustments of weights of connections between units (the analogue of a neuron) with the goal of minimizing a cost function. The success of deep neural networks is often attributed to their ability to learn sophisticated and effective representations of high-dimensional input spaces [25]. Standard representations of real world inputs like images (matrices of pixel-wise RGB values) and audio (time series corresponding to pressure levels) do not capture any perceptually meaningful structures of the input. For instance, translating an image by a few pixels, which causes a very small change in behaviourally relevant constructs to be interpreted from the image, corresponds to a huge change in the representation space. The goal of learning representations is to find ways to transform the inputs such that the underlying structure of the input space is reflected in the representation space.

Representations in deep learning, much like in neuroscience, refer to a systematic relationship between the features of the input and the activity of the units in the neural network i.e. similar inputs should elicit similar activation patterns. It should be noted that the notion of similarity in the input space is subject to the goals of the system; representations where all cats are similar fare well if the objective is to tell cats from dogs, but it renders the system useless for classifying the breed of a cat. That being said, deep neural networks have made great progress towards learning representations that can generalize to a wide range of tasks [26]. This progress, however, comes at the expense of interpretability. These networks have millions, sometimes even billions, of parameters spread over tens of layers which has the effect of turning it into a black box, not due to inaccessibility of the internal states, but because of the intractable high-dimensionality of them. Owing to these properties, DNNs have invited comparison to another intractably high-dimensional system with billions of parameters that performs remarkably well at a plethora of diverse tasks - the brain.

In the following sections of this paper, that comparison is fleshed out considering representations as the basis for comparison. It is important to draw the distinction between sensory and conceptual representations at this point as the paper is primarily concerned with the former. While, there is an enduring debate around the nature of conceptual representations in philosophy as well as neuroscience [27], there is agreement on the fact that concepts are made up of a wide variety of sensory, behavioral, and ontological components. For instance, the concept of an apple would consist of sensory and behavioural information like its color, shape, texture, smell, taste and the fact that it can be used to make a pie. It would also include more abstract information ranging from different categorical hierarchies that it's a part of to an apple falling from a tree and setting into motion one of the most influential discoveries in all of physics. Because of this multifaceted nature, conceptual representations don't lend themselves very well to comparison between biological and artificial neural networks. It can be argued that representations in deeper layers of neural networks are somewhat conceptual as they correspond to categorical labels but that has the effect of blurring the line between conceptual and sensory representations. If the early layers are sensory and the later layers are conceptual, it implies that these categories lie on a continuous spectrum and any threshold to tell one from the other would only be arbitrary. For this discussion, it is assumed that concepts only emerge in higher levels of processing and the representations in individual sensory cortices (visual and auditory) as well as in convolutional layers of deep neural networks are non-conceptual.

III. Framework for Comparison

In this section a framework for comparing biological and artificial neural networks will be introduced and existing work on the comparison for computer vision and audition will be reviewed. The same framework will be used to extend the comparison to multimodal representations in state-of-the-art audio-visual systems.

A. Behaviour

The simplest way to compare biological cognitive systems to their artificial counterparts is to take a page from the book of Alan Turing and perform some version of the Turing test. If a machine can generate behaviour that is indistinguishable from, or convincingly close to, that of a human, we can grant it intelligence. Turing's original proposition was essentially an NLP problem, but the same idea can be extended to visual, auditory, and other modalities as is commonly done when claims of human-level performance are made by AI practitioners. Such claims are almost always backed by statistical metrics of performance like accuracy, false positive and false negative rates, confusion matrices, or any combination thereof. Human-level performance in these cases boils down to the following question - "Can an AI system make correct predictions as often as a human". This is very reminiscent of Turing's proxy question for whether machines can think, which was "Can a machine fool an interrogator just as often as a human". In both cases, it is assumed that if the answer to the question is yes, intelligence can be granted to the machine. As widely used as this dimension of comparison is, the behaviouristic underpinnings force a black-box approach to intelligence with no regard to internal representations. Consequently, it serves us no purpose if representations are of interest. Also, in light of behaviourism, the question of whether we "understand" things that we can build takes on new life. The same ends can be met with wildly different means, so AI systems that rival, or even surpass, human-level performance merely seem to be great feats of technology and not avenues for understanding the mind. Even if we build systems with desired behavioural properties, a deeper analysis of their inner workings is necessary.

B. Structure

To study the structure of neural networks, biological and artificial, is to characterise the elemental building blocks and their organisation in the network. The building blocks, for both neuroscience and deep learning, are fairly well understood but the properties that emerge from their complex spatial arrangement are an active subject of study. The motivation for structural analysis, apart from a general scientific merit, is that strong inductive biases are imposed by the structural properties of a neural network. These biases make neural networks unreasonably effective at learning to solve complex problems with a relatively small number of parameters when compared to the input space. This is perhaps because the inductive biases reflect the structure of the external world. Convolutional networks, for instance, have filters arranged in a hierarchy which results in the expansion of the receptive field of units with depth. This hierarchical architecture, inspired by the visual cortex, imposes an inductive bias that reflects structural properties of objects in natural images like invariance to shifting, rotating, zooming, or cropping the image. Poldrack argues that such inductive biases are a requirement for any intelligent system to function in a real-world scenario [28]. It is important to draw the distinction between the structural elements of the computations and that of the physical substrate that instantiates the computations because it is the former that is of interest for this discussion.

C. *Function*

A complementary but orthogonal dimension to structure is to study what function is served by the structural elements and their arrangement. Functions, much like the structure, can be studied at multiple levels of granularity. In biology, functional analysis can be at levels ranging from a single neuron to that of the organism even to that of collective populations [29]. Many aspects of biological neural networks motivate a functionalist approach. The structural characteristics have evolved in a way that maximises an organism's evolutionary fitness, which is defined by the functions the organism can perform. However, there is an information bottleneck in evolution; when an organism reproduces, it can not transfer information about all neural connections and their synaptic weights. Rather, the inheritance consists of the ability to regenerate the structure during the early stages of life by interacting with the environment via an action-perception loop. It can be reasonably assumed that the functions an organism needs to perform act as the compass for the evolution of inherited traits as well as the learning of acquired ones. For artificial neural networks, the case for a functionalist approach is even stronger; the objective function, which is the ceiling of our functional analysis, is a parameter that is under our control and is explicitly optimized during training. All changes in network parameters are solely driven by how modifications to a given parameter affect the objective function and since the objectives drive the learning, it is imperative to use biologically grounded objectives if DNNs are to be used as a tool for understanding the mind.

The fields of deep learning and neuroscience have seen tremendous cross-pollination in the last decade with DNNs borrowing from principles in neuroscience but also contributing to data analysis and model validation in neuroscientific research. One line of studies that is specially relevant for this discussion is the prediction of neural activity using DNNs trained not on neural data but on real world tasks [30-35]. It has been observed in [31] that DNNs trained on behavioural tasks outperform ones trained on neural data in predicting neural response to stimuli for the visual cortex. DNNs have also been shown to be very effective for the inverse task of generating images to maximally stimulate a neuron or a group of neurons [30]. Similar results have been observed in the auditory cortex and authors of [32] have even suggested that this approach sheds new light on the structural and computational properties of the auditory cortex. While goal-driven neural networks are proving to be powerful models of sensory processing in the brain, it is difficult to make a comparison between the two based on high-level behavioural goals because the goals of biological systems are not explicit and well-defined. However, comparing representations offers us a way to investigate the objective functions of nature. To that end, the method of representational similarity analysis [36] has been adapted to compare representation of DNNs to that of animals and humans [37], where representational similarity is measured in terms of the correlation between responses to pairs of stimuli. The idea behind this method is that for two representations to be similar, stimuli that are close together in one embedding space should also be close in the other. It has been observed that DNNs trained on object recognition have highly correlated representational geometries to that of the visual cortex.

In addition to the quantitative comparison of representations, qualitative measures have also been studied [32, 34]. In both the studies, the errors made by humans and DNNs have been used as a way to check for representational similarity. The rationale being that in DNNs, correct predictions are explicitly optimized and there is an enormous space of possible representations, so it is not unlikely that two very different representations may make similar correct predictions. But same failure cases are a much stronger indication of similar representations. It has also been shown in [34] that qualitative properties of human perception like the Thatcher effect, mirror confusion, Weber's law etc. can emerge in task-driven neural networks.

However, their study highlights a very important caveat of this approach - the representations learned by DNNs are highly task-specific. The qualities of human visual processing that were recovered in the DNNs were the ones that helped the model perform the task of image classification. Qualities that didn't offer a significant advantage for the task given the specific dataset like 3D-processing and tolerance to occlusions were not observed in the learned representations. This suggests that the conception of the task as object recognition, for instance, is incomplete and the features of the training data are a relevant characteristic in the definition of the task. The data for a DNN is analogous to the environment for an organism and one would not question the importance of the environment in shaping representations with theories like Gibson's ecological psychology heralding it as the sole driver of behaviour. While issues of data quality are considered of extreme importance for industrial applications with movements like Andrew Ng's data-centric AI gaining popularity by the day, there seems to be a lack of discourse in the cognitive science community using data-driven approaches to understand the mind. Another potential consideration for the goal-driven approach is the assumption that DNNs are tabula rasa. There are not only strong inductive biases resulting from the architecture, but also different initialisations have been shown to result in learning of convergent as well as divergent features with the same architecture, dataset, and performance [38]. The above two caveats warrant great caution when making general statements about similarities and differences between biological and artificial neural networks.

D. *Mode of Acquisition of Representations*

Traversing down Marr's levels of analysis, the algorithmic question of "how" is just as important as the functional question of "what" when comparing biological and artificial cognition. In classical Marrian analysis, the question would be how a set of computations is performed, but in neural networks, the computations are innumerably large and the behavioural properties of the network emerge from their interactions, so it is futile to attempt to disentangle the contribution of each. Moreover, the computations are not designed but learned, so a rephrasing of the question to how a set computations is learned makes for a more adequate inquiry.

Learning can occur at two distinct levels; a) experience dependent learning at the level of an organism b) evolution dependent learning at the level of the species. The word 'learning' has been used loosely for the latter in the sense that in both the cases, there are changes in the brain that result in increased competence for a given function. The innateness of representations have been a subject of raging debate since the early days of philosophy, but for the purposes of this discussion, the specifics of what is innate vs acquired is not pertinent. Infact, in light of evolution, they seem to be two sides of the same coin and directly correspond to our two ways of learning. The innate characteristics, learned through evolution, provide an organism with inductive biases and the structural apparatus that enables effective experience based learning.

The fundamental principles behind evolutionary 'learning' are fairly well understood, but the consensus on the mechanisms for experience dependent learning isn't as clear. One idea that connects several candidate theories into a coherent whole is the free energy principle [39-40]. It states that a self-organising system that is in equilibrium with its environment must minimize its free energy. For the brain, that corresponds to minimizing surprise, essentially framing the brain as a Bayesian inference engine. The notion that the brain can be surprised implies that it is constantly making predictions about future states (external and internal) and has a generative model of the world it inhabits. Indeed, there is evidence to support the idea that anticipation-based free energy minimisation is a fundamental property of information processing throughout the brain [41-43].

Any task that deep neural networks are trained on can also be formulated in this framework. In a classification problem, for instance, the neural network makes a prediction, it is contrasted against the observation i.e. the true class, and the error is propagated through the network to change the parameters in a way that minimises the error. The error can come from external supervision as in a supervised classification problem, but it can also be self-generated as is suggested for the brain [44]. Autoencoders are able to generate a supervision signal by using information compression followed by reconstruction of the original input and contrasting the reconstruction against the original. And anticipation (next word prediction) has been used as a self-supervised learning task to great effect in state of the art language models [45]. Moreover, biological and artificial neural networks aren't just on the same page at the computational level, but perhaps also the algorithmic level. Deep neural networks are trained using error-backpropagation, and while there have been doubts regarding the biological plausibility of such mechanisms, there has been recent work in favour of error gradient based learning in the hippocampus [44].

IV. Comparing Audio-Visual Representations

Vision and audition are complementary sources of information and effective integration of these senses offers significant survival advantages for an organism. The ability to localise sounds and connect them to visual objects enables a rich understanding of a dynamic environment. While visual and auditory inputs are collected independently, the senses are deeply connected in forming our perception of the world [46-48]. In this section, recent work in self-supervised learning of audio-visual representations[49-51] has been contrasted against the human audio-visual system based on the aspects of comparison discussed above.

A. Structure

There are many shared structural characteristics among the three publications considered for the analysis - they all use separate audio and vision channels followed by a downstream integration network with both the audio and vision networks using a standard CNN architecture. The integration is performed in different ways in each. L3 [49] simply concatenates auditory and visual features and uses the combination for subsequent layers. On the other hand, the representations are projected onto a common space in [50]. [51] uses a separate attention network which takes in the audio and visual representations as inputs and performs sound source localization for which cite work in psychology and cognitive science as the motivation. The use of separate vision and audio feature extractors and downstream processing of those features is consistent with separate sensory cortices and multisensory integration in the brain [52-54].

B. Function

While the three methods used similar architectural elements, they all had unique objectives. [49] was trained on an audio-visual correspondence task where the objective was to classify whether a pair of audio and video segments comes from the same recording. [50] also used an audio-visual correspondence task but the network was designed for cross modal retrieval i.e. it should be able to retrieve related segments of audio given a video and vice versa. This was achieved by minimizing the distance between corresponding audio and video samples, essentially projecting the representations onto a common space. [51] was trained for sound source localization where the objective was to generate a segmentation map to segment the sound generating objects in a video. While only sound-source localisation appears to be a biologically relevant task, all of the goals resulted in effective general purpose auditory and visual representations that performed well on benchmark tasks. Moreover, [49] and [50] evaluated their representations for sound source localisation and found that they could generate semantically meaningful segmentation maps without explicitly training their models to do so. This raises questions about the relevance of biologically motivated goals and representational similarity analysis of the learned representations between these methods can be the first step towards finding the answers. If very different goals can result in similar representations, that would mean that specifics of the goals are not relevant and the data, or the environment, plays the bigger role in shaping the representations.

C. Mode of Acquisition of Representations

All the three methods use self-supervised learning where the supervision signal was generated by the downstream task and the error is optimized using gradient descent. Multisensory integration in humans and animals has been framed as a Bayesian inference problem [52-54], consistent with the free energy principle. Uncertainty about the stimuli, caused by occlusions, noise, lack of visibility etc., is minimized by integrating information from complementary modalities. For example, while crossing a road, a vehicle out of the range of vision can be sensed by sound and that immediately forces our eyes towards the predicted position based on the sound, such that the uncertainty in the location of the vehicle can be minimized. In this regard, biological and artificial neural networks seem to be at odds with each other.

References

[1] Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. J. Physiol. 160, 106–154 (1962).

[2] Mnih, Volodymyr, et al. "Human-level control through deep reinforcement learning." nature 518.7540 (2015): 529-533.

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

[4] Gross, Charles G. "Genealogy of the "grandmother cell"." The Neuroscientist 8.5 (2002): 512-518.

[5] Turing, Alan Mathison. "On computable numbers, with an application to the Entscheidungsproblem." J. of Math 58.345-363 (1936): 5.

[6] Poe, Edgar Allan. "A history of computer chess from the Mechanical Turk to Deep Blue"

[7] Newell, A., Shaw, J. C., Simon, H. A. (1958) Elements a theory of human problem solving. Psychological Review, 65(3), 151–166.

[8] Marr, D., Poggio, T. & Ullman, S. Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information (MIT Press, 2010).

[9] Chomsky, Noam. "Syntactic Structures." (1957).

[10] Ernst, G. & Newell, A. (1969). GPS: A Case Study in Generality and Problem Solving. New York: Academic Press.

[11] Winograd, Terry. "Shrdlu: A system for dialog." (1972).

[12] Hendrix, G.G., Sacerdoti, E.D., Sagalowicz, D., Slocum, J. —Developing a natural language interface to complex data‖, in ACM Transactions on database systems,1978,pp.105- 147.

[13] McCulloch, Warren S., and Walter Pitts. "A logical calculus of the ideas immanent in nervous activity." The bulletin of mathematical biophysics 5.4 (1943): 115-133.

[14] Pitts, Walter, and Warren S. McCulloch. "How we know universals the perception of auditory and visual forms." The Bulletin of mathematical biophysics 9.3 (1947): 127-147.

[15] McCulloch, Warren S. "Information in the head." Synthese (1953): 233-247.

[16] Lettvin, J. Y., Maturana, H. R., McCulloch, W. S., & Pitts, W. H. (1959). What the frog's eye tells the frog's brain. Proceedings of the IRE, 47(11), 1940-1951.

[17] Rosenblatt, Frank. "The perceptron: a probabilistic model for information storage and organization in the brain." Psychological review 65.6 (1958): 386.

[18] Minsky, Marvin, and Seymour Papert. "Perceptrons." (1969).

[19] Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. Adv. Neural Inf. Proc. Sys. 25, 1097–1105 (2012).

[20] Hannun, A. et al. Deep speech: scaling up end-to-end speech recognition. Preprint at arXiv https://arxiv.org/abs/1412.5567 (2014).

[21] Radford, A. et al. Better language models and their implications. OpenAI Blog https://openai.com/blog/better-language-models/ (2019).

[22] Finn, C., Goodfellow, I. & Levine, S. Unsupervised learning for physical interaction through video prediction. Adv. Neural Inf. Proc. Sys. 29, 64–72 (2016).

[23] Silver, D. et al. Mastering the game of Go without human knowledge. Nature 550, 354–359 (2017).

[24] Santoro, A. et al. A simple neural network module for relational reasoning. Adv. Neural Inf. Proc. Sys. 30, 4967–4976 (2017).

[25] Bengio, Yoshua, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives." IEEE transactions on pattern analysis and machine intelligence 35.8 (2013): 1798-1828.

[26] Kolesnikov, Alexander, et al. "Big transfer (bit): General visual representation learning." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16. Springer International Publishing, 2020.

[27] Kiefer, Markus, and Friedemann Pulvermüller. "Conceptual representations in mind and brain: theoretical developments, current evidence and future directions." cortex 48.7 (2012): 805-825.

[28] Poldrack, Russell A. "The physics of representation." Synthese 199.1 (2021): 1307-1325.

[29] Dawkins, R. (2006). The Selfish Gene. Oxford University Press.

[30] Bashivan, P., Kar, K. & DiCarlo, J. J. Neural population control via deepimage synthesis. Science 364, eaav9436 (2019).

[31] Cadena, Santiago A., et al. "Deep convolutional models improve predictions of macaque V1 responses to natural images." PLoS computational biology 15.4 (2019): e1006897.

[32] Kell, Alexander JE, et al. "A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy." Neuron 98.3 (2018): 630-644.

[33] Yamins, D.L.K. et al. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. USA 111, 8619–8624 (2014).

[34] Jacob, Georgin, et al. "Qualitative similarities and differences in visual object representations between brains and deep networks." Nature communications 12.1 (2021): 1-14.

[35] Yamins, Daniel LK, and James J. DiCarlo. "Using goal-driven deep learning models to understand sensory cortex." Nature neuroscience 19.3 (2016): 356-365.

[36] Kriegeskorte, Nikolaus, Marieke Mur, and Peter A. Bandettini. "Representational similarity analysis-connecting the branches of systems neuroscience." Frontiers in systems neuroscience 2 (2008): 4.

[37] Khaligh-Razavi, S.M. & Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput. Biol. 10, e1003915 (2014).

[38] Li, Yixuan, et al. "Convergent learning: Do different neural networks learn the same representations?." FE@ NIPS. 2015.

[39] Friston K., Kilner, J. & Harrison, L. A free energy principle for the brain. J. Physiol. Paris 100, 70–87 (2006).

[40] Friston, Karl. "The free-energy principle: a unified brain theory?." Nature reviews neuroscience 11.2 (2010): 127-138.

[41] Herreros, I. and Verschure, P.F. (2015) About the goal of a goals' goal theory. Cogn. Neurosci. 6, 218–219

[42] Maffei, G. et al. (2017) The perceptual shaping of anticipatory actions. Proc. R. Soc. B Biol. Sci. 284, 20171780

[43] Herreros-Alonso, I. and Arsiwalla, X.D. (2016) A forward model at Purkinje cell synapses facilitates cerebellar anticipatory control. In Advances in Neural Information Processing Systems 29 (NIPS 2016) (Lee, D. et al., eds), pp. 3828–3836, NIPS

[44] Santos-Pata, Diogo, et al. "Epistemic autonomy: self-supervised learning in the mammalian hippocampus." Trends in Cognitive Sciences (2021).

[45] Brown, Tom B., et al. "Language models are few-shot learners." arXiv preprint arXiv:2005.14165 (2020).

[46] King, Andrew J., et al. "Developmental plasticity in the visual and auditory representations in the mammalian superior colliculus." Nature 332.6159 (1988): 73-76.

[47] Choe, Chong S., et al. "The "ventriloquist effect": Visual dominance or response bias?." Perception & Psychophysics 18.1 (1975): 55-60.

[48] McGurk, H., and J. MacDonald. "Hearing Voices and Seeing Eyes." Nature 264 (1976): 746-748.

[49] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 609–617 "

[50] R. Arandjelovic and A. Zisserman, “Objects that sound,” in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 435–451. ”

[51] Senocak, A., Oh, T. H., Kim, J., Yang, M. H., & Kweon, I. S. (2018). Learning to localize sound source in visual scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4358-4366).

[52] Ernst, Marc O., and Heinrich H. Bülthoff. ”Merging the senses into a robust percept.” Trends in cognitive sciences 8.4 (2004): 162-169.

[53] Körding, Konrad P., et al. ”Causal inference in multisensory perception.” PLoS one 2.9 (2007): e943.

[54] Ghazanfar, Asif A., and Charles E. Schroeder. ”Is neocortex essentially multisensory?.” Trends in cognitive sciences 10.6 (2006): 278-285.