Deep Predictive Learning in Neocortex and Pulvinar

Randall C. O'Reilly, Jacob L. Russin, and John Rohrlich
Department of Psychology, Computer Science, and Center for Neuroscience
University of California Davis
1544 Newton Ct
Davis, CA 95618
oreilly@ucdavis.edu

March 13, 2020

We thank Dean Wyatte, Tom Hazy, Seth Herd, Kai Krueger, Tim Curran, David Sheinberg, Lew Harvey, Jessica Mollick, Will Chapman, Helene Devillez, and the rest of the CCN Lab for many helpful comments and suggestions. Supported by: ONR grants ONR N00014-19-1-2684 / N00014-18-1-2116, N00014-14-1-0670 / N00014-16-1-2128, N00014-18-C-2067, N00014-13-1-0067, D00014-12-C-0638. This work utilized the Janus supercomputer, which is supported by the National Science Foundation (award number CNS-0821794) and the University of Colorado Boulder. The Janus supercomputer is a joint effort of the University of Colorado Boulder, the University of Colorado Denver and the National Center for Atmospheric Research. All data and materials will be available at `https://github.com/ccnlab/deep-obj-cat` upon publication.

Abstract

How does the human brain learn new concepts from raw sensory experience, without explicit instruction? We still do not have a widely-accepted answer to this central question. Here, we propose a detailed biological mechanism for the widely-embraced idea that learning is based on the differences between predictions and actual outcomes (i.e., *predictive error-driven learning*). Specifically, numerous weak projections into the pulvinar nucleus of the thalamus drive top-down predictions, and sparse, strong *driver* inputs from lower areas encode the actual outcome. Because these driver inputs originate in layer 5 intrinsic bursting (5IB) neurons, the outcome is only briefly activated, roughly every 100 msec (i.e., 10 Hz, *alpha*). Thus, the prediction error is a *temporal difference* in activation states over the pulvinar, from an earlier prediction to a subsequent burst of outcome. This temporal difference can drive local synaptic changes throughout the neocortex, supporting a biologically-plausible form of error backpropagation learning. We implemented these mechanisms in a large-scale model of the visual system, and found that the simulated inferiotemporal (IT) pathway learns to systematically categorize 3D objects according to invariant shape properties, based solely on predictive learning from raw visual inputs. We found that these categories match human judgments on the same stimuli, and are consistent with neural representations in IT cortex in primates. Thus, this biologically-based form of predictive error-driven learning can drive cognitively-useful levels of learning, directly from raw visual stimuli.

The fundamental epistemological conundrum of how knowledge emerges from raw experience has challenged philosophers and scientists for centuries. There have been significant advances in understanding the detailed biochemical basis of learning in terms of synaptic plasticity between neurons (Lüscher & Malenka, 2012), and many cognitive and computational models of learning. However, there is still no widely-accepted answer to this puzzle that is clearly supported by known biological mechanisms and also produces effective learning at computational and cognitive levels. At these levels, the idea that we learn via an active *predictive* process goes back to Helmholtz's *recognition by synthesis* proposal (von Helmholtz, 2013), and has been widely embraced in a wide range of different frameworks (Elman, 1990; Elman, Bates, Karmiloff-Smith, Johnson, Parisi, & Plunkett, 1996; Mumford, 1992; Dayan, Hinton, Neal, & Zemel, 1995; Rao & Ballard, 1999; Kawato, Hayakawa, & Inui, 1993; Friston, 2005). Here, we propose a detailed biological mechanism for a specific form of *predictive error-driven learning* based on distinctive patterns of connectivity between the neocortex and the pulvinar nucleus of the thalamus (Sherman & Guillery, 2006; Usrey & Sherman, 2018). Specifically, numerous weak projections into the thalamic relay cells (TRC's) in the pulvinar drive top-down predictions, and sparse, strong *driver* inputs from lower areas encode the actual outcome, and learning is based on the difference. Because these driver inputs originate in layer 5 intrinsic bursting (5IB) neurons, the outcome is only briefly activated, roughly every 100 msec (i.e., 10 Hz, *alpha*). Thus, the prediction error is a *temporal difference* in activation states over the pulvinar, from an earlier prediction to a subsequent burst of outcome. This temporal difference can drive local synaptic changes throughout the neocortex, supporting a biologically-plausible form of error backpropagation learning (O'Reilly, 1996; Bengio, Mesnard, Fischer, Zhang, & Wu, 2017; Ackley, Hinton, & Sejnowski, 1985; Hinton & McClelland, 1988).

One primary objective of this paper is to describe this biologically-based mechanism for predictive error-driven learning in sufficient detail that it can be clearly evaluated relative to a wide range of existing anatomical and electrophysiological data. We also describe a number of specific empirical predictions that follow from this functional view of the thalamocortical circuit, which could potentially be tested by current neuroscientific methods. Thus, a major contribution of this work is to provide a clear functional role for this distinctive, seemingly functionally relevant thalamocortical circuitry, that constrasts with existing ideas about what it might be doing, in testable ways.

A second major objective of this paper is to implement this predictive error-driven learning mechanism in a computational model that faithfully captures its essential biological features, while still being sufficiently simplified computationally that it can be used to simulate large-scale brain networks, to test whether the learning mechanism can drive the formation of cognitively-useful representations. In particular, there is a critical question for any purely predictive-learning model: can it develop high-level, abstract ways of representing the raw sensory inputs, while learning from nothing but predicting these low-level visual inputs. For instance, can predictive learning really eliminate the need for human-labeled image datasets where abstract category information is explicitly used to train object recognition models via error-backpropagation? Existing predictive-learning models based on error backpropagation (Lotter, Kreiman, & Cox, 2016) have not strongly demonstrated the development of abstract, categorical representations without additional human-labeled training. Instead, previous work has shown that predictive learning can be a useful method for pretraining networks that are subsequently trained using human-generated labels.

Through large-scale simulations based on the known structure of the visual system, we found that this biologically based predictive learning mechanism developed high-level abstract representations that systematically categorize 3D objects according to invariant shape properties, based on raw visual inputs alone. We found that these categories match human judgments on the same stimuli, and are consistent with neural representations in inferotemporal (IT) cortex in primates (Cadieu, Hong, Yamins, Pinto, Ardila, Solomon, Majaj, & DiCarlo, 2014). Furthermore, we show that comparison predictive DCNN models lacking these biological features (Lotter et al., 2016) did not learn object categories that go beyond the visual input struc-

ture. Thus, it is possible that incorporating certain biological properties of the brain can potentially provide a better understanding of human learning at multiple levels relative to existing DCCN models. However, it is important to emphasize that our objectives in this work are *not* to produce a better machine-learning (ML) algorithm per se, but rather to test the computational properties of our biologically-based, scientific theory for how the mammalian brain might learn. Thus, we explicitly dissuade readers from the inevitable desire to evaluate the importance of our model based on differences in standard ML metrics: it should instead be evaluated on its ability to explain a wide range of data across multiple levels of analysis, just as every other scientific theory is evaluated.

The remainder of the paper is organized as follows. First, we provide a concise overview of the biologically-based predictive error-driven learning framework. Next, we discuss the relevant biological data in detail, along with testable predictions that can differentiate this account of what this system does relative to existing ideas. Then, we present the large-scale model of the visual system, which learns by predicting over brief visual movies of 3D objects rotating and translating over time and space. We find that the model develops strongly categorical, shape-based representations in its upper IT layers, and these match those of human participants evaluating the same 3D objects. Furthermore, we show that these categorical representations diverge significantly from the similarity structure present in the lower layers of the network. Thus, we conclude that this form of predictive error-driven learning is capable of going beyond the surface structure of the raw sensory input, to develop higher-level abstract representations that otherwise have only been produced in neural models through explicit training via human-labeled image datasets. To further explore this space, we evaluated two other prediction-error learning models using pure error-backpropagation, based on current deep-convolutional neural network (DCNN) principles, and found that they did not develop the same kind of high-level categories, and instead remained largely tied to the similarity structure of the raw visual inputs. Thus, there may be some important features of the biologically-based model that enable this ability to learn higher-level structure beyond that of the raw inputs.

Predictive Error-driven Learning in the Neocortex and Pulvinar

Figure 1 shows the thalamocortical circuits characterized by Sherman and Guillery (2006) (see also Sherman & Guillery, 2013; Usrey & Sherman, 2018), which have two distinct projections converging on the principal thalamic relay cells (TRCs) of the *pulvinar*, which is the primary thalamic nucleus that is interconnected with higher-level posterior cortical visual areas; (Shipp, 2003). One projection consists of numerous, weaker connections originating in deep layer VI of the neocortex (the 6CT corticothalamic projecting cells). The other is a very sparse (typically one-to-one; (Rockland, 1998, 1996)) and very strong *driver* pathway that originates from lower-level layer 5 intrinsic bursting cells (5IB). These 5IB neurons fire discrete bursts roughly every 100 msec (Larkum, Zhu, & Sakmann, 1999; Franceschetti, Guatteo, Panzica, Sancini, Wanke, & Avanzini, 1995; Lorincz, Kekesi, Juhasz, Crunelli, & Hughes, 2009; Saalmann, Pinsk, Wang, Li, & Kastner, 2012), which corresponds to the widely-studied *alpha* frequency of 10 Hz that originates in cortical deep layers and has important effects on a wide range of perceptual and attentional tasks (Buffalo, Fries, Landman, Buschman, & Desimone, 2011; VanRullen & Koch, 2003; Jensen, Bonnefond, & VanRullen, 2012; Fiebelkorn & Kastner, 2019).

The existing literature generally characterizes the 6CT projection as *modulatory* (Sherman & Guillery, 2013; Usrey & Sherman, 2018), but a number of electrophysiological recordings from awake, behaving animals clearly show sustained, continuous patterns of neural firing in pulvinar TRC neurons, which is not consistent with the idea that they are only being driven by their 5IB inputs (Bender, 1982; Petersen, Robinson, & Keys, 1985; Bender & Youakim, 2001; Robinson, 1993; Saalmann et al., 2012; Komura, Nikkuni, Hirashima, Uetake, & Miyamoto, 2013). Indeed, these recordings show that pulvinar neural firing generally resembles that of the visual areas they interconnect with.

In contrast to the standard view, the core idea behind our theory is that the top-down 6CT projections drive a prediction *across the extent of the pulvinar, which is then subject to an (implicit) comparison with the subsequent activation state resulting from the strong 5IB driver inputs.* Indeed, the properties of these two pathways appear ideal for this predictive learning role. First, predictive learning requires some way of distinguishing between the *prediction* and the *outcome*, and having two separate pathways thus fills this need. Furthermore, one of these projections systematically originates in higher-order areas (according to standard ways of identifying the cortical hierarchy; Markov, Ercsey-Ravasz, Gomes, R, Lamy, Magrou, Vezoli, Misery, Falchier, Quilodran, Gariel, Sallet, Gamanut, Huissoud, Clavagnier, Giroud, Sappey-Marinier, Barone, Dehay, Toroczkai, Knoblauch, Van Essen, & Kennedy, 2014; Van Essen & Maunsell, 1983), and is thus suitable for driving top-down predictions, while the other originates in lower-order areas, and is thus suitable for driving the bottom-up *ground truth* signal. In addition, this ground truth signal is conveyed in a very direct, strong, focal manner, which would preserve the nature of the lower-level representations where it originates, whereas the top-down predictive pathway has many synapses which can then organize and coordinate inputs from multiple higher areas to form a coherent overall prediction. Our framework strongly predicts that these top-down projections are plastic, so that they can learn over time to shape better predictive representations, whereas the bottom-up drivers need not be. Finally, it is essential for a genuine prediction that the ground truth be *hidden* while the prediction is being formulated: otherwise, it just becomes a *post hoc* explanation. The phasic burst firing of the 5IB neurons thus provides this needed time interval between bursts when the top-down projections can be establishing the predicted activation state over the pulvinar.

Thus, remarkably, this thalamocortical circuit provides *precisely* the necessary ingredients to support predictive error-driven learning. To summarize (Figure 1b): we hypothesize that the top-down 6CT projections drive a pattern of activity over the pulvinar TRC neurons during the first roughly 75 msec of a 100 msec alpha cycle, that represents the prediction of the subsequent activity pattern that then emerges during the final 25 msec, which largely reflects the strong 5IB bottom-up ground-truth driver inputs. Thus, the difference or *prediction error* signal is reflected in the *temporal difference* of these activation states over time, which contrasts with most other predictive learning frameworks that hypothesize the existence of explicit error-coding neurons whose firing directly reflects the prediction error itself. In other words, our hypothesis is that the pulvinar is only ever directly representing either the top-down prediction or the bottom-up actual outcome, and the prediction-error difference between these remains as an implicit difference in these activation states over time. This is consistent with the way that the original *Boltzmann Machine* learning algorithm worked (Ackley et al., 1985), and how our subsequent biologically-plausible versions of error backpropagation also work (O'Reilly, 1996; O'Reilly & Munakata, 2000).

TODO: combine our summary DeepLeabra diagram with SG06 figure in one a, b fig 1.

This temporal difference prediction error signal in the pulvinar is broadcast back up to the neocortex through extensive reciprocal connections that target the same areas where the 6CT top-down projections originate (Shipp, 2003), thus providing the means by which synaptic plasticity in the neocortex can learn to improve the overall accuracy of the predictions being sent down to the pulvinar. For this to work correctly, the neocortical neurons must be sensitive to this temporal-difference in their activation states over time, which again has been a longstanding feature of our framework for how error-backpropagation can work in the cortex (O'Reilly, 1996). Although the available evidence for this mechanism remains indirect, there are detailed biological mechanisms with significant empirical support that can achieve this form of learning, as detailed below.

OLD: Based on this and other biological evidence, we hypothesize that this distinctive thalamocortical circuit supports predictive error-driven learning in a way that shapes learning throughout the posterior neocortex (O'Reilly, Wyatte, & Rohrlich, 2014b) (Figure 2a). Specifically, sensory predictions in posterior neocortex are generated roughly every 100 msec at the alpha rhythm, and the pulvinar represents this top-down prediction for roughly 75 msec of the alpha cycle as it develops, after which point the layer

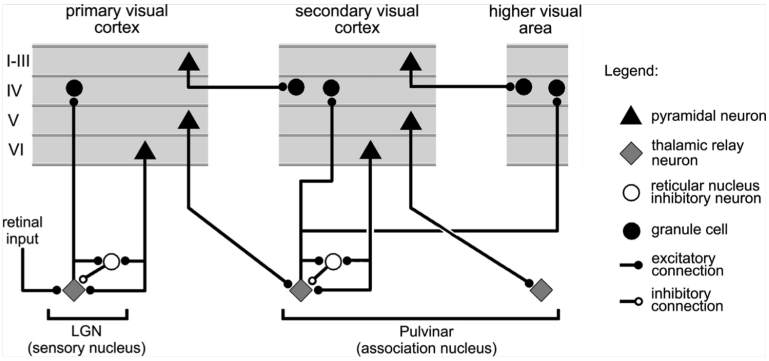

Figure 1: Summary figure from Sherman & Guillery (2006) showing the strong feedforward driver projection emanating from layer 5IB cells in lower layers (e.g., V1), and the much more numerous feedback "modulatory" projection from layer 6CT cells. We interpret these same connections as providing a prediction (6CT) vs. outcome (5IB) activity pattern over the pulvinar.

5IB intrinsic-bursting neurons send strong, bottom-up driving input to the pulvinar, representing the actual sensory stimulus. Critically, the prediction error is implicit in the temporal difference between these two periods of activity within the alpha cycle over the pulvinar, which is consistent with the biologically plausible form of error-driven cortical learning used in our models (O'Reilly, 1996). The pulvinar sends broad projections back up to all of the areas that drive top-down predictions into it (Shipp, 2003; Mumford, 1991), thus broadcasting this error signal to drive local synaptic plasticity in the neocortex. This mathematically approximates gradient descent to minimize overall prediction errors (O'Reilly, 1996). This computational framework makes sense of otherwise puzzling anatomical and physiological properties of the cortical and thalamic networks (Sherman & Guillery, 2006), and is consistent with a wide range of detailed neural and behavioral data (O'Reilly et al., 2014b).

The known biological mechanisms are widely thought to produce a form of Hebbian learning (Lüscher & Malenka, 2012; Hebb, 1949), but recent advances in deep neural network learning strongly demonstrate that error backpropagation (Rumelhart, Hinton, & Williams, 1986) is essential for computationally-powerful, cognitively-realistic learning (Krizhevsky, Sutskever, & Hinton, 2012; LeCun, Bengio, & Hinton, 2015; Schmidhuber, 2015).

Computational models with powerful learning mechanisms driven by raw images or other sensory inputs provide an attractive way to approach this problem, yet many of the current models based on deep convolutional neural networks (DCNN's) notoriously require explicit training from massive human-labeled datasets . Such models are cognitively implausible, as non-human primates and human infants learn to recognize and categorize objects without the benefit of such labeled data (Lake, Ullman, Tenenbaum, & Gershman, 2017ed). Furthermore, the biological plausibility of the core learning mechanism, *error backpropagation* , has also long been questioned on biological grounds (Crick, 1989), although various related biologically plausible mechanisms have been proposed (O'Reilly, 1996; Xie & Seung, 2003; Bengio et al., 2017).

Here we propose a form of *predictive* error-driven learning (Elman, 1990; Elman et al., 1996) that learns directly on raw sensory inputs without the need for explicit human-generated labels. This learning mechanism leverages distinctive patterns of connectivity between the neocortex and thalamus (Sherman & Guillery, 2006) (Figure 1) to achieve a biologically based form of predictive learning. In contrast to existing predictive learning frameworks (Mumford, 1992; Rao & Ballard, 1999; Kawato et al., 1993; Friston, 2005), we suggest that error signals, as differences between a prediction and what actually occurs, remain as a *temporal difference* in activation states in the network, and are not explicitly represented through error-coding neurons. Specifically, the pulvinar nucleus of the thalamus receives both top-down predictions and

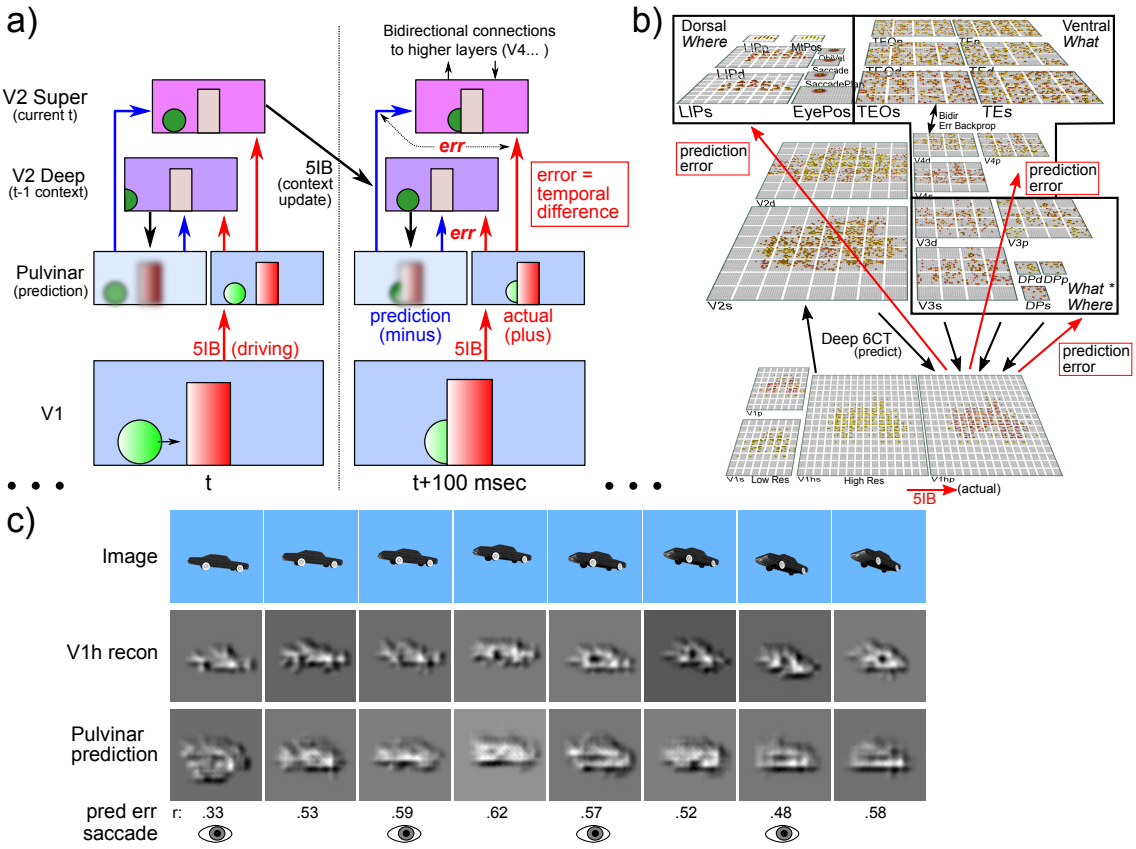

Figure 2: **a)** Temporal evolution of information flow in the DeepLeabra algorithm predicting visual sequences, over two alpha cycles of 100 msec each. In each alpha cycle, the V2 Deep layer (lamina 5, 6) uses the prior 100 msec of context to generate a prediction (*minus* phase) on the pulvinar thalamic relay cells (TRC). The bottom-up outcome is driven by V1 5IB strong driver inputs (*plus* phase); error-driven learning occurs as a function of the *temporal difference* between these phases, in both superficial (lamina 2, 3) and deep layers, sent via broad pulvinar projections. 5IB bursting in V2 drives update of temporal context in V2 Deep layers, and also the plus phase in higher area TRC, to drive higher-level predictive learning. See supporting information (SI) for more details. **b)** The *What-Where-Integration, WWI* model. The dorsal *Where* pathway learns first, using easily-abstracted *spatial blobs*, to predict object location based on prior motion, visual motion, and saccade efferent copy signals. This drives strong top-down inputs to lower areas with accurate spatial predictions, leaving the *residual* error concentrated on *What* and *What* * *Where* integration. The V3 and DP (dorsal prelunate) constitute the *What* * *Where* integration pathway, binding features and locations. V4, TEO, and TE are the *What* pathway, learning abstracted object category representations, which also drive strong top-down inputs to lower areas. *s* suffix = superficial, *d* = deep, *p* = pulvinar. **c)** Example sequence of 8 alpha cycles that the model learned to predict, with the reconstruction of each image based on the V1 gabor filters (*V1 recon*), and model-generated prediction (correlation r prediction error shown). The low resolution and reconstruction distortion impair visual assessment, but r values are well above the r's for each V1 state compared to the previous time step (mean = .38, min of .16 on frame 4 – see SI for more analysis). Eye icons indicate when a saccade occurred.

bottom-up sensory outcome signals, alternating within an *alpha* frequency cycle (10 Hz, 100 msec), via two distinctive pathways. Thus, our framework has many testable differences from these existing theories, and we argue that existing data is more consistent with our framework.

A critical question for predictive learning is whether it can develop high-level, abstract ways of representing the raw sensory inputs, while learning from nothing but predicting these low-level visual inputs. For instance, can predictive learning really eliminate the need for human-labeled image datasets where abstract category information is explicitly used to train object recognition models via error-backpropagation? Existing predictive-learning models based on error backpropagation (Lotter et al., 2016) have not demonstrated the development of abstract, categorical representations. Previous work has shown that predictive learning can be a useful method for pretraining networks that are subsequently trained using human-generated labels, but here we focus on the formation of systematic categories *de-novo*.

To determine if our biologically based predictive learning model (Figure 2b) can naturally form such categorical encodings in the complete absence of external category labels, we showed the model brief movies of 156 3D object exemplars drawn from 20 different basic-level categories (e.g., car, stapler, table lamp, traffic cone, etc.) selected from the CU3D-100 dataset (O'Reilly, Wyatte, Herd, Mingus, & Jilk, 2013). The objects moved and rotated in 3D space over 8 movie frames, where each frame was sampled at the alpha frequency (Figure 2c). There were also saccadic eye movements every other frame, introducing an additional predictive-learning challenge. An efferent copy signal enabled full prediction of the effects of the eye movement, and allows the model to capture *predictive remapping* (a widely-studied signature of predictive learning in the brain) (Duhamel, Colby, & Goldberg, 1992; Cavanagh, Hunt, Afraz, & Rolfs, 2010), and introduces additional predictive-learning challenge. The only learning signal available to the model was a prediction error generated by the temporal difference between what it predicted to see in the next frame and what was actually seen.

We performed a representational similarity analysis (RSA) on the learned activity patterns at each layer in the model, and found that the highest IT layer (TE) produced a systematic organization of the 156 3D objects into 5 categories (Figure 3a), which visually correspond to the overall shape of the objects (pyramid-shaped, vertically-elongated, round, boxy / square, and horizontally-elongated). This organization of the objects matches that produced by humans making shape similarity judgments on the same set of objects, using the V1 reconstruction as shown in Figure 2c to capture the model's coarse-grained perception (Figure 3b; see supporting information for methods and further analysis). Critically, Figure 3c shows that the overall similarity structure present in IT layers (TEO, TE) of the biological model is significantly different from the similarity structure at the level of the V1 primary visual input. Thus the model, despite being trained only to generate accurate visual input-level predictions, has learned to represent these objects in an abstract way that goes beyond the raw input-level information. Furthermore, this abstract category organization reflects the overall visual shapes of the objects as judged by human participants, suggesting that the model is extracting geometrical shape information that is invariant to the differences in motion, rotation, and scaling that are present in the V1 visual inputs. We further verified that at the highest IT levels in the model, a consistent, spatially-invariant representation is present across different views of the same object (e.g., the average correlation across frames within an object was .901). This is also evident in Figure 3a by virtue of the close similarity across multiple objects within the same category.

Further evidence for the progressive nature of representation development in our model is shown in Figure 4, which compares the similarity structures in layers V4 and IT in macaque monkeys (Cadieu et al., 2014) with those in corresponding layers in our model. In both the monkeys and our model, the higher IT layer builds upon and clarifies the noisier structure that is emerging in the earlier V4 layer. Considerable other work has also compared DCNN representations with these same data from monkeys (Cadieu et al., 2014), but it is essential to appreciate that those DCNN models were explicitly trained on the category labels, making it somewhat less than surprising that such categorical representations developed. By contrast, we

Figure 3: **a)** Category similarity structure that developed in the highest layer, TE, of the biologically based predictive learning model, showing *1-correlation* similarity of the TE representation for each 3D object against every other 3D object (156 total objects). Blue cells have high similarity, and model has learned block-diagonal clusters or categories of high-similarity groupings, contrasted against dissimilar off-diagonal other categories. Clustering maximized average *within - between* correlation distance (see SI). All items from the same basic-level object categories (N=20) are reliably subsumed within learned categories. **b)** Human similarity ratings for the same 3D objects, presented with the V1 reconstruction (see Fig 1c) to capture coarse perception in model, aggregated by 20 basic-level categories. Each cell is 1 - proportion of time given object pair was rated more similar than another pair (see SI). The human matrix shares the same centroid categorical structure as the model (confirmed by permutation testing and agglomorative cluster analysis, see SI). **c)** Emergence of abstract category structure over the hierarchy of layers. Red line = correlation similarity between the TE similarity matrix (shown in panel a) and all layers; black line shows correlation similarity between V1 against all layers (1 = identical; 0 = orthogonal). Both show that IT layers (TEO, TE) progressively differentiate from raw input similarity structure present in V1, and, critically, that the model has learned structure beyond that present in the input.

reiterate that our model has discovered its categorical representations entirely on its own, with no explicit categorical inputs or training of any kind.

Figure 5 shows the results from a purely backpropagation-based (Bp) version of the same model architecture, and a standard PredNet model (Lotter et al., 2016) with extensive hyperparameter optimization (see SI). In the Bp model, the highest layers in the network form a simple binary category structure overall, and the detailed item-level similarity structure does not diverge significantly from that present at the lowest V1 inputs, indicating that it has not formed novel systematic structured representations, in contrast to those formed in the biologically based model. Similar results were found in the PredNet model, where the highest layer representations remained very close to the V1 input structure. Thus, it is clear that the additional biologically derived properties are playing a critical role in the development of abstract categorical representations that go beyond the raw visual inputs. These properties include: excitatory bidirectional connections, inhibitory competition, and an additional Hebbian form of learning that serves as a regularizer (similar to weight decay) on top of predictive error-driven learning (O'Reilly, 1998; O'Reilly & Munakata, 2000).

Each of these properties could promote the formation of categorical representations. Bidirectional connections enable top-down signals to consistently shape lower-level representations, creating significant attractor dynamics that cause the entire network to settle into discrete categorical attractor states. By contrast, backpropagation networks typically lack these kinds of attractor dynamics, and this could contribute significantly to their relative lack of categorical learning. Hebbian learning drives the formation of representations that encode the principal components of activity correlations over time, which can help more categorical

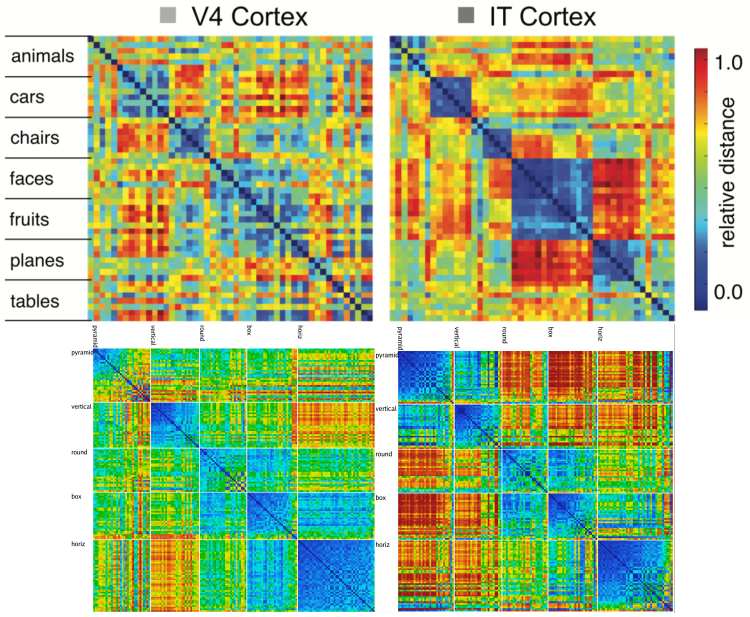

Figure 4: Comparison of progression from V4 to IT in macaque monkey visual cortex (top row, from Cadieu et al., 2014) versus same progression in model (replotted using comparable color scale). Although the underlying categories are different, and the monkeys have a much richer multi-modal experience of the world to reinforce categories such as foods and faces, the model nevertheless shows a similar qualitative progression of stronger categorical structure in IT, where the block-diagonal highly similar representations are more consistent across categories, and the off-diagonal differences are stronger and more consistent as well (i.e., categories are also more clearly differentiated). Note that the critical difference in our model versus those compared in Cadieu et al. 2014 and related papers is that they explicitly trained their models on category labels, whereas our model is *entirely self-organizing* and has no external categorical training signal.

representations coalesce (and results below already indicate its importance). Inhibition, especially in combination with Hebbian learning, drives representations to specialize on more specific subsets of the space. Ongoing work is attempting to determine which of these is essential in this case (perhaps all of them) by systematically introducing some of these properties into the backpropagation model, though this is difficult because full bidirectional recurrent activity propagation, which is essential for conveying error signals top-down in the biological network, is incompatible with the standard efficient form of error backpropagation, and requires much more computationally intensive and unstable forms of fully recurrent backpropagation (Williams & Zipser, 1992; Pineda, 1987). Furthermore, Hebbian learning requires inhibitory competition which is difficult to incorporate within the backpropagation framework.

Figure 6 shows just a few of the large number of parameter manipulations that have been conducted to develop and test the final architecture. For example, we hypothesized that separating the overall prediction problem between a spatial *Where* vs. non-spatial *What* pathway (Ungerleider & Mishkin, 1982; Goodale & Milner, 1992), would strongly benefit the formation of more abstract, categorical object representations in the *What* pathway. Specifically, the *Where* pathway can learn relatively quickly to predict the overall spatial trajectory of the object (and anticipate the effects of saccades), and thus effectively regress out that component of the overall prediction error, leaving the residual error concentrated in object feature information, which can train the ventral *What* pathway to develop abstract visual categories. Figure 6a shows that, indeed, when the *Where* pathway is lesioned, the formation of abstract categorical representations in the intact

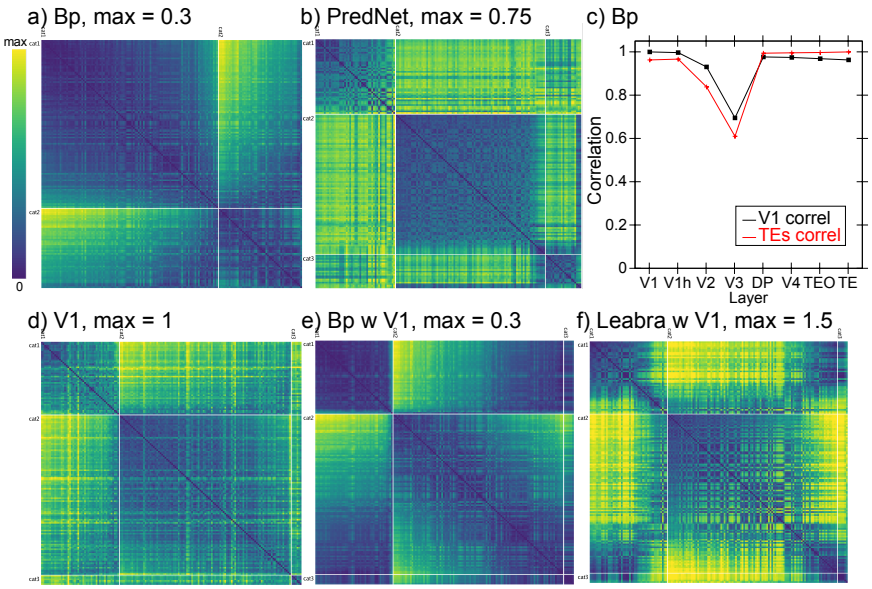

Figure 5: **a)** Best-fitting category similarity for TE layer of the backpropagation (Bp) model with the same What / Where structure as the biological model. Only two broad categories are evident, and the lower *max* distance (0.3 vs. 1.5 in biological model) means that the patterns are highly similar overall. **b)** Best-fitting similarity structure for the PredNet model, in the highest of its layers (layer 6), which is more differentiated than Bp (max = 0.75) but also less cleanly similar within categories (i.e., less solidly blue along the block diagonal), and overall follows a broad category structure similar to V1. **c)** Comparison of similarity structures across layers in the Bp model (compare to Figure 2c): unlike in the biological model, the V1 structure is largely preserved across layers, and is little different from the structure that best fits the TE layer shown in panel **a**, indicating that the model has not developed abstractions beyond the structure present in the visual input. Layer V3 is most directly influenced by spatial prediction errors, so it differs from both in strongly encoding position information. **d)** The best fitting V1 structure, which has 2 broad categories and banana is in a third category by itself. The lack of dark blue on the block diagonal indicates that these categories are relatively weak, and every item is fairly dissimilar from every other. **e)** The same similarities shown in panel **a** for Bp TE also fit reasonably well sorted according to the V1 structure (and they have a similar average within - between contrast differences, of 0.0838 and 0.0513 – see SI for details). **f)** The similarity structure from the biological model resorted in the V1 structure does *not* fit well: the blue is not aligned along the block diagonal, and the yellow is not strictly off-diagonal. This is consistent with the large difference in average contrast distance: 0.5071 for the best categories vs. 0.3070 for the V1 categories.

What pathway is significantly impaired. Figure 6b shows that full predictive learning, as compared to just encoding and decoding the current state (which is much easier computationally, and leads to much better overall accuracy), is also critical for the formation of abstract categorical representations — prediction is a "desirable difficulty" (Bjork, 1994). Finally, Figure 6c shows the impact of reducing Hebbian learning, which impairs category learning as expected.

In conclusion, we have demonstrated that learning based strictly on predicting what will be seen next is, in conjunction with a number of critical biologically motivated network properties and mechanisms, capable of generating abstract, invariant categorical representations of the overall shapes of objects. The nature of these shape representations closely matches human shape similarity judgments on the same objects. Thus, predictive learning has the potential to go beyond the surface structure of its inputs, and develop systematic, abstract encodings of the "deeper" structure of the environment. Relative to existing machine-learning-based

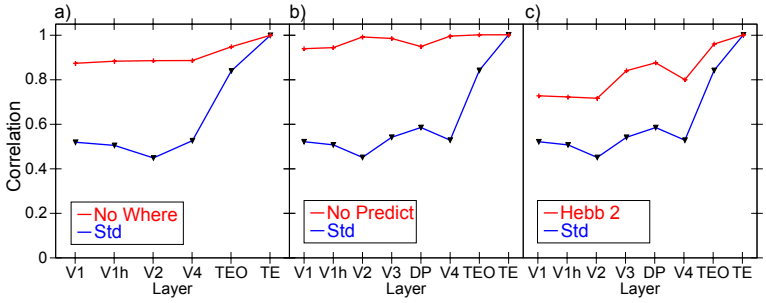

Figure 6: Effects of various manipulations on the extent to which TE representations differentiate from V1. *Std* is the same result shown in Figure 2c from the intact model for ease of comparison. All of the following manipulations significantly impair the development of abstract TE categorical representations (i.e., TE is more similar V1 and the other layers). **a)** Dorsal *Where* pathway lesions, including lateral inferior parietal sulcus (LIP), V3, and dorsal prelunate (DP). This pathway is essential for regressing out location-based prediction errors, so that the residual errors concentrate feature-encoding errors that train the *What* pathway. **b)** Allowing the deep layers full access to current-time information, thus effectively eliminating the prediction demand and turning the network into an auto-encoder, which significantly impairs representation development, and supports the importance of the challenge of predictive learning for developing deeper, more abstract representations. **c)** Reducing the strength of Hebbian learning by 20% (from 2.5 to 2), demonstrating the essential role played by this form of learning on shaping categorical representations. Eliminating Hebbian learning entirely (not shown) prevented the model from learning anything at all, as it also plays a critical regularization and shaping role on learning.

approaches in "deep learning", which have generally focused on raw categorization accuracy measures using explicit category labels or other human-labeled inputs, the results here suggest that focusing more on the nature of what is learned in the model might provide a valuable alternative approach. Considerable evidence in cognitive neuroscience suggests that the primary function of the many nested ("deep") layers of neural processing in the neocortex is to *simplify* and aggressively *discard* information (Simons & Rensink, 2005), to produce precisely the kinds of extremely valuable abstractions such as object categories, and, ultimately, symbol-like representations that support high-level cogntive processeses such as reasoning and problem-solving (Rougier, Noelle, Braver, Cohen, & O'Reilly, 2005; O'Reilly, Petrov, Cohen, Lebiere, Herd, & Kriete, 2014a). Thus, particularly in the domain of predictive or generative learning, the metric of interest should not be the accuracy of prediction itself (which is indeed notably worse in our biologically based model compared to the DCNN-based PredNet and backpropagation models), but rather whether this learning process results in the formation of simpler, abstract representations of the world that can in turn support higher levels of cognitive function.

Considerable further work remains to be done to more precisely characterize the essential properties of our biologically motivated model necessary to produce this abstract form of learning, and to further explore the full scope of predictive learning across different domains. We strongly suspect that extensive cross-modal predictive learning in real-world environments, including between sensory and motor systems, is a significant factor in infant development and could greatly multiply the opportunities for the formation of higher-order abstract representations that more compactly and systematically capture the structure of the world (Yu & Smith, 2012). Future versions of these models could thus potentially provide novel insights into the fundamental question of how deep an understanding a pre-verbal human, or a non-verbal primate, can develop (Spelke, Breinlinger, Macomber, & Jacobson, 1992; Elman et al., 1996), based on predictive learning mechanisms. This would then represent the foundation upon which language and cultural learning builds, to shape the full extent of human intelligence.

Biological Details

From (Usrey & Sherman, 2018): Second, one feature held in common by the cortico- thalamic projections from Layers 5 and 6 is that the cells of origin do not have axon branches that innervate other cortical areas (Petrof, Viaene, & Sherman, 2012), although they do have branches that pro- vide local cortical innervation. Thus the cells in Layers 5 and 6 that innervate other cortical areas do not extend subcortical branches, and those that do innervate thalamus and other subcortical sites are not involved in providing direct innervation of other cortical areas.

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, *9*(1), 147–169.

Bender, D. B. (1982). Receptive-field properties of neurons in the macaque inferior pulvinar. *Journal of neurophysiology*, *48*.

Bender, D. B., & Youakim, M. (2001). Effect of attentive fixation in macaque thalamus and cortex. *Journal of neurophysiology*, *85*, 219–234.

Bengio, Y., Mesnard, T., Fischer, A., Zhang, S., & Wu, Y. (2017). STDP-compatible approximation of backpropagation in an energy-based model. *Neural Computation*, *29*(3), 555–577.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA, US: The MIT Press.

Buffalo, E. A., Fries, P., Landman, R., Buschman, T. J., & Desimone, R. (2011). Laminar differences in gamma and alpha coherence in the ventral stream. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(27), 11262–11267.

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., & DiCarlo, J. J. (2014). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLoS Computational Biology*, *10*(12), e1003963.

Cavanagh, P., Hunt, A. R., Afraz, A., & Rolfs, M. (2010). Visual stability based on remapping of attention pointers. *Trends in Cognitive Sciences*, *14*(4), 147–153.

Crick, F. (1989). The recent excitement about neural networks. *Nature*, *337*, 129–132.

Dayan, P., Hinton, G. E., Neal, R. N., & Zemel, R. S. (1995). The Helmholtz machine. *Neural Computation*, *7*(5), 889–904.

Duhamel, J. R., Colby, C. L., & Goldberg, M. E. (1992). The updating of the representation of visual space in parietal cortex by intended eye movements. *Science*, *255*(5040), 90–92.

Elman, J., Bates, E., Karmiloff-Smith, A., Johnson, M., Parisi, D., & Plunkett, K. (1996). *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.

Elman, J. L. (1990). Finding Structure In Time. *Cognitive Science*, *14*(2), 179–211.

Fiebelkorn, I. C., & Kastner, S. (2019). A rhythmic theory of attention. *Trends in Cognitive Sciences*, *23*(2), 87–101.

Franceschetti, S., Guatteo, E., Panzica, F., Sancini, G., Wanke, E., & Avanzini, G. (1995). Ionic mechanisms underlying burst firing in pyramidal neurons: Intracellular study in rat sensorimotor cortex. *Brain Research*, *696*(1–2), 127–139.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B*, *360*(1456), 815–836.

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*(1), 20–25.

Hebb, D. O. (1949). *The Organization of Behavior*. New York: Wiley.

Hinton, G. E., & McClelland, J. L. (1988, January). Learning representations by recirculation. In D. Z. Anderson (Ed.), *Neural Information Processing Systems (NIPS 1987)*, Vol. 0 (pp. 358–366). New York: American Institute of Physics.

Jensen, O., Bonnefond, M., & VanRullen, R. (2012). An oscillatory mechanism for prioritizing salient unattended stimuli. *Trends in Cognitive Sciences*, *16*(4), 200–206.

Kawato, M., Hayakawa, H., & Inui, T. (1993). A forward-inverse optics model of reciprocal connections between visual cortical areas. *Network: Computation in Neural Systems*, *4*(4), 415–422.

Komura, Y., Nikkuni, A., Hirashima, N., Uetake, T., & Miyamoto, A. (2013). Responses of pulvinar neurons reflect a subject's confidence in visual categorization. *Nature Neuroscience*, *16*(6), 749–755.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 25* (pp. 1097–1105). Curran Associates, Inc.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017/ed). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*.

Larkum, M. E., Zhu, J. J., & Sakmann, B. (1999). A new cellular mechanism for coupling inputs arriving at different cortical layers. *Nature*, *398*(6725), 338–341.

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

Lorincz, M. L., Kekesi, K. A., Juhasz, G., Crunelli, V., & Hughes, S. W. (2009). Temporal framing of thalamic relay-mode firing by phasic inhibition during the alpha rhythm. *Neuron*, *63*(5), 683–696.

Lotter, W., Kreiman, G., & Cox, D. (2016). Deep predictive coding networks for video prediction and unsupervised learning. *arXiv:1605.08104 [cs, q-bio]*.

Lüscher, C., & Malenka, R. C. (2012). NMDA receptor-dependent long-term potentiation and long-term depression (LTP/LTD). *Cold Spring Harbor Perspectives in Biology*, *4*(6), a005710.

Markov, N. T., Ercsey-Ravasz, M. M., Gomes, R., R, A., Lamy, C., Magrou, L., Vezoli, J., Misery, P., Falchier, A., Quilodran, R., Gariel, M. A., Sallet, J., Gamanut, R., Huissoud, C., Clavagnier, S., Giroud, P., Sappey-Marinier, D., Barone, P., Dehay, C., Toroczkai, Z., Knoblauch, K., Van Essen, D. C., & Kennedy, H. (2014). A Weighted and Directed Interareal Connectivity Matrix for Macaque Cerebral Cortex. *Cerebral Cortex*, *24*(1), 17–36.

Mumford, D. (1991). On the computational architecture of the neocortex. *Biological Cybernetics*, *65*(2), 135–145.

Mumford, D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biological Cybernetics*, *66*(3), 241–251.

O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation*, *8*(5), 895–938.

O'Reilly, R. C. (1998). Six Principles for Biologically-Based Computational Models of Cortical Cognition. *Trends in Cognitive Sciences*, *2*(11), 455–462.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*. Cambridge, MA: MIT Press.

O'Reilly, R. C., Petrov, A. A., Cohen, J. D., Lebiere, C. J., Herd, S. A., & Kriete, T. (2014a). How Limited Systematicity Emerges: A Computational Cognitive Neuroscience Approach. In I. P. Calvo, & J. Symons (Eds.), *The architecture of cognition: Rethinking Fodor and Pylyshyn[1]s Systematicity Challenge*. Cambridge, MA: MIT Press.

O'Reilly, R. C., Wyatte, D., Herd, S., Mingus, B., & Jilk, D. J. (2013). Recurrent Processing during Object Recognition. *Frontiers in Psychology*, *4*(124).

O'Reilly, R. C., Wyatte, D., & Rohrlich, J. (2014b). Learning Through Time in the Thalamocortical Loops. *arXiv:1407.3432 [q-bio]*.

Petersen, S. E., Robinson, D. L., & Keys, W. (1985). Pulvinar nuclei of the behaving rhesus monkey: Visual responses and their modulation. *Journal of neurophysiology*, *54*.

Pineda, F. J. (1987). Generalization of Backpropagation to Recurrent Neural Networks. *Physical Review Letters*, *18*, 2229–2232.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.

Robinson, D. L. (1993). Functional contributions of the primate pulvinar. *Progress in brain research*, *95*.

Rockland, K. S. (1996). Two types of corticopulvinar terminations: Round (type 2) and elongate (type 1). *The Journal of comparative neurology*, *368*, 57–87.

Rockland, K. S. (1998). Convergence and branching patterns of round, type 2 corticopulvinar axons. *The Journal of Comparative Neurology*, *390*(4), 515–536.

Rougier, N. P., Noelle, D., Braver, T. S., Cohen, J. D., & O'Reilly, R. C. (2005). Prefrontal Cortex and the Flexibility of Cognitive Control: Rules Without Symbols. *Proceedings of the National Academy of Sciences*, *102*(20), 7338–7343.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(9), 533–536.

Saalmann, Y. B., Pinsk, M. A., Wang, L., Li, X., & Kastner, S. (2012). The pulvinar regulates information transmission between cortical areas based on attention demands. *Science*, *337*(6095), 753–756.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117.

Sherman, S., & Guillery, R. (2006). *Exploring the Thalamus and Its Role in Cortical Function*. Cambridge, MA: MIT Press.

Sherman, S., & Guillery, R. (2013). *Functional Connections of Cortical Areas: A New View From the Thalamus*. Cambridge, MA: MIT Press.

Shipp, S. (2003). The functional logic of cortico-pulvinar connections. *Philosophical Transactions of the Royal Society of London B*, *358*(1438), 1605–1624.

Simons, D. J., & Rensink, R. A. (2005). Change blindness: Past, present, and future. *Trends in cognitive sciences*, *9*(1), 16–20.

Spelke, E., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of Knowledge. *Psychological Review*, *99*(4), 605–632.

Ungerleider, L. G., & Mishkin, M. (1982). Two Cortical Visual Systems. In D. J. Ingle, M. A. Goodale, & R. J. W. Mansfield (Eds.), *The Analysis of Visual Behavior* (pp. 549–586). Cambridge, MA: MIT Press.

Usrey, W. M., & Sherman, S. M. (2018). Corticofugal circuits: Communication lines from the cortex to the rest of the brain. *Journal of Comparative Neurology*, *0*(0).

Van Essen, D. C., & Maunsell, J. H. R. (1983). Hierarchical organization and functional streams in the visual cortex. *Trends in Neurosciences*, *6*, 370–375.

VanRullen, R., & Koch, C. (2003). Is perception discrete or continuous? *Trends in Cognitive Sciences*, *7*(5), 207–213.

von Helmholtz, H. (2013). *Treatise on Physiological Optics, Vol III*. Courier Corporation.

Williams, R. J., & Zipser, D. (1992). Gradient-based learning algorithms for recurrent networks and their computational complexity. In Y. Chauvin, & D. E. Rumelhart (Eds.), *Backpropagation: Theory, Architectures and Applications*. Hillsdale, NJ: Erlbaum.

Xie, X., & Seung, H. S. (2003). Equivalence of backpropagation and Contrastive Hebbian Learning in a layered network. *Neural Computation*, *15*(2), 441–454.

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition*, *125*(2), 244–262.