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

June 6, 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 generate top-down predictions, and sparse, strong *driver* inputs from lower areas supply the actual outcome, originating in layer 5 intrinsic bursting (5IB) neurons. Thus, the outcome is only briefly activated, roughly every 100 msec (i.e., 10 Hz, *alpha*), resulting in a *temporal difference error signal*, which drives local synaptic changes throughout the neocortex, resulting in 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 inferotemporal (IT) pathway learns to systematically categorize 3D objects according to invariant shape properties, based solely on predictive learning from raw visual inputs. These categories match human judgments on the same stimuli, and are consistent with neural representations in IT cortex in primates.

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 functional 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 (TRCs) in the pulvinar generate top-down predictions, and sparse, strong *driver* inputs from lower areas supply 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; Ackley, Hinton, & Sejnowski, 1985; Hinton & McClelland, 1988; Bengio, Mesnard, Fischer, Zhang, & Wu, 2017; Whittington & Bogacz, 2019; Lillicrap, Santoro, Marris, Akerman, & Hinton, 2020).

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, and in contrast with other existing proposals. We provide 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, this work provides a clear functional role for the distinctive thalamocortical circuitry that contrasts with existing ideas about what it might be doing, in testable ways.

A second major objective 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 example, most current models of visual object recognition that have been compared against neurophysiological data rely on large human-labeled image datasets to explicitly train abstract category information via error-backpropagation (Cadieu, Hong, Yamins, Pinto, Ardila, Solomon, Majaj, & DiCarlo, 2014; Rajalingham, Issa, Bashivan, Kar, Schmidt, & DiCarlo, 2018). Existing predictive-learning models based on error backpropagation (Lotter, Kreiman, & Cox, 2016) have not clearly 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 our 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 et al., 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 structure. 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 DCNN 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 narrow, performance-based 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 1a 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*, the primary thalamic nucleus that is interconnected with higher-level posterior cortical visual areas; (Shipp, 2003; Arcaro, Pinsk, & Kastner, 2015). 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,

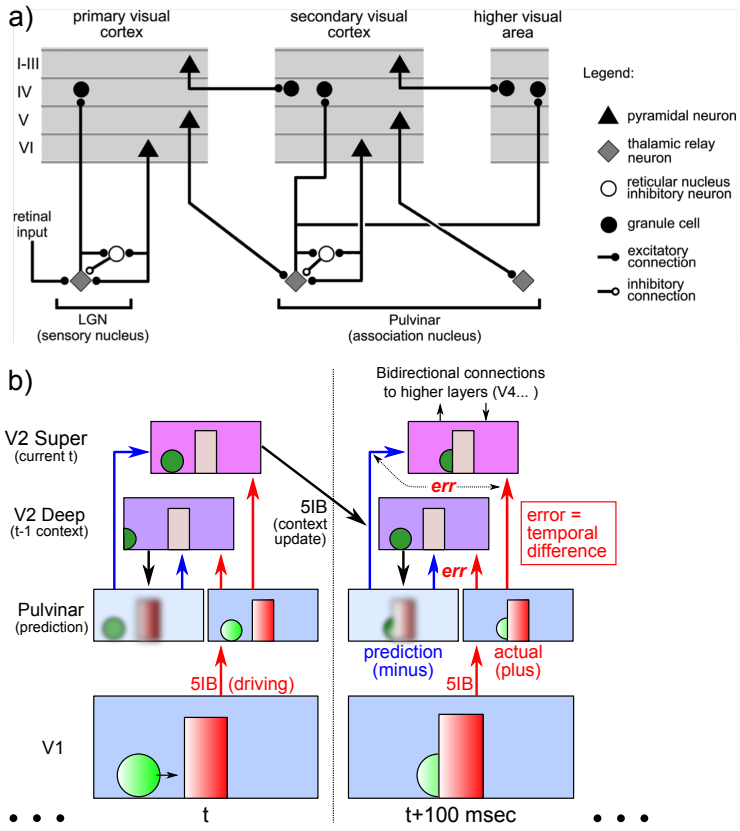

Figure 1: **a)** Summary figure from Sherman & Guillery (2006) showing the strong feedforward driver projection emanating from layer 5IB cells in lower layers (e.g., V1), and the much more numerous feedback "modulatory" projection from layer 6CT cells. We interpret these same connections as providing a prediction (6CT) vs. outcome (5IB) activity pattern over the pulvinar. **b)** Temporal evolution of information flow under our prediction error hypothesis, operating on 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 lower-level (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 updating of temporal context in V2 Deep layers (this phasic updating prevents current outcome activation in superficial layers from informing the prediction), while also driving the plus phase in higher areas of pulvinar that learn to predict V2 activation states, and so on.

Hirashima, Uetake, & Miyamoto, 2013; Zhou, Schafer, & Desimone, 2016). Indeed, these recordings show that pulvinar neural firing generally resembles that of the visual areas they interconnect with. This is important, because our predictive learning framework requires that these 6CT top-down projections be capable of driving TRC activity directly.

Specifically, 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 precedes the subsequent *outcome* state resulting from the strong 5IB driver inputs, as illustrated in Figure 1b. This prediction activity state in pulvinar TRC neurons should develop during the first roughly 75 msec of a 100 msec alpha cycle, while the final 25 msec 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. In

other words, our hypothesis is that the pulvinar is 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.

We further hypothesize that, due to pervasive neural adaptation mechanisms, outcomes that are consistent with prior predictions will continue to activate the same population of adapted neurons, whereas unpredicted outcomes will generally activate new subsets of neurons, and thus drive a phasic increase in activity at the onset of such stimuli. The 5IB neurons may be particularly responsive to these phasic increases, causing their bursting to coincide preferentially with unexpected outcomes, and driving the entrainment of the alpha cycle. Thus, the pulvinar may experience a continuous sequence of predicted states based on the top-down 6CT projections, with relatively weaker rhythmic 5IB driving inputs, until an unpredicted stimulus arises. At this point, error-driven learning would be more strongly engaged as a function of the phasic release from adaptation and 5IB burst activation. This provides a natural mechanism for alpha entrainment and learning rate modulation as a function of prediction error, and is consistent with the *expectation suppression* phenomena as discussed below.

If, as we suggest, 5IB bursting preferentially drives learning, then the prediction is essentially *defined* as the state prior to 5IB bursting, and the learning rule automatically causes that prior state to better anticipate the subsequent state. This means that even if no prediction was initially generated, the learning will work to create one to the extent possible over iterations of learning. It also means that although the alpha rhythm defines a baseline minimum prediction window, aribtrarily longer time windows would also be effective in driving this entrained form of predictive learning — learning always just happens whenever something unexpected occurs, at any point. In the typical lab experiment with phasic stimuli presented outside of an ongoing predictable temporal sequence (which is likely uncharacteristic of the natural world), there may often be no significant prediction prior to stimulus onset, and we would expect such stimuli to reliably drive 5IB bursting, which is consistent with available data.

The properties of these two pathways are notably well suited for this predictive learning role, in the following ways:

- A true prediction (i.e., about the future, as in the famous quote about what makes prediction hard: "prediction is very difficult, especially about the future", attributable to Danish author Robert Storm Petersen) must be prevented from cheating and relying on direct information about that which is being predicted: thus there must be a mechanism preventing the incoming outcome information from "contaminating" the prediction. The phasic, bursting nature of the 5IB driver inputs provides this essential feature, creating a window where no outcome signals are present, when the prediction can be represented. Further, the top-down drivers of this prediction in the 6CT deep layers are also hypothesized to be updated by 5IB bursting in their local columnar circuits, such that they are also kept isolated from current superficial-layer activation reflecting the sensory outcome being predicted (Figure 1b).

- Generating a prediction requires converging inputs from a range of higher-level cortical areas, to integrate the contributions of multiple different specialized pathways in the challenging problem of predicting what will happen next. This is consistent with the broad, integrative nature of the top-down 6CT inputs (Shipp, 2003; Mumford, 1991).

- Furthermore, it can take some time to integrate all these signals, which is consistent with the outcome bursts occupying a briefer 25 msec of the 100 msec alpha cycle, with the remainder available for this integration. The overall duration of the alpha cycle itself may represent a reasonable compromise between this integration time and the need to keep up with tracking changes in the world.

- The outcome signal should be as *veridical* as possible, and should arise from lower areas in the hierarchy relative to the corresponding 6CT inputs: the bottom-up, one-to-one nature of the 5IB driver

projections can directly convey such veridical outcome signals.

- The prediction error signal should be widely broadcast back out to the same areas that provide the top-down predictions, to provide the training signal that improves these predictions. This is also a known, distinctive property of this circuitry (Shipp, 2003; Mumford, 1991).

- For cortical neurons receiving these projections from the pulvinar, there must be some way in which the difference between prediction and outcome (i.e., the error itself) can drive learning. Here we hypothesize that this difference remains as a *temporal difference* error signal, i.e., the difference over time in pulvinar activation states, arising naturally as a prediction state followed by the outcome state. This contrasts with prevalent alternative hypotheses that require a separate population of neurons to compute a prediction error "explicitly" and transmit it directly through neural firing (Rao & Ballard, 1999; Kawato et al., 1993; Friston, 2005, 2010; Ouden, Kok, & Lange, 2012; Lotter et al., 2016). Despite many attempts to identify such explicit error-coding neurons in the cortex, no substantial body of unambiguous evidence has been discovered (Kok & de Lange, 2015; Kok, Jehee, & de Lange, 2012; Summerfield & Egner, 2009; Lee & Mumford, 2003; Walsh, McGovern, Clark, & O'Connell, 2020). Furthermore, due to the positive-only firing rate nature of neural coding, two separate populations would be required to convey both signs of prediction error signals. Thus, we think that the temporal-difference nature of the prediction error signal is more efficient and should naturally emerge from the basic operation of the circuit. Indeed, this form of error signal has the opposite problem, that it is consistent with a massive body of existing data, and thus does not represent a distinctive prediction of this framework.

- There is a long history of computational models of error-driven learning based on temporal-difference signals (Ackley et al., 1985; O'Reilly, 1996), and we have recently provided a direct biological mechanism for this form of learning based on a biologically-detailed model of spike timing dependent plasticity (STDP) (Urakubo, Honda, Froemke, & Kuroda, 2008). We showed that when activated by realistic Poisson spike trains, this STDP model produces a non-monotonic learning curve similar to that of the BCM model (Bienenstock, Cooper, & Munro, 1982), which has been widely established as resulting from competing calcium-driven postsynaptic plasticity pathways (Lüscher & Malenka, 2012). As in the BCM framework, we hypothesized that the threshold crossover point in this nonmonotonic curve moves dynamically — if this happens on the alpha timescale (Lim, McKee, Woloszyn, Amit, Freedman, Sheinberg, & Brunel, 2015), then it can reflect the prediction phase of activity, producing a net error-driven learning rule based on a subsequent calcium signal reflecting the outcome state, which mathematically approximates gradient descent to minimize overall prediction errors (O'Reilly, 1996).

Thus, remarkably, this thalamocortical circuit appears to provide *precisely* the necessary ingredients to support predictive error-driven learning. Interestingly, although Sherman and Guillery (2006) did not propose a predictive learning mechanism as just described, they did speculate about a potential role for this circuit in motor forward-model learning and the predictive remapping phenomenon (Sherman & Guillery, 2011; Usrey & Sherman, 2018). In addition, Pennartz, Dora, Muckli, and Lorteije (2019) also suggested that the pulvinar may be involved in predictive learning, but within the explict error-coding framework and not involving the detailed aspects of the above-described circuitry.

As we discuss later, this proposed predictive role for the pulvinar is not incompatible with the more widely-discussed role it may play in attention (LaBerge & Buchsbaum, 1990; Bender & Youakim, 2001; Snow, Allen, Rafal, & Humphreys, 2009; Saalmann & Kastner, 2011; Zhou et al., 2016; Fiebelkorn & Kastner, 2019). Indeed, we think these two functions are synergistic (i.e., you predict what you attend, and vice-versa), and have initial computational results consistent with this idea.

In the following section, we discuss some of the most important neural data of relevance to our hypotheses (beyond that summarized above) followed by a list of some predictions that would clearly test the validity of this framework.

Existing Neuroscience Data

Extensive biological evidence supports the alpha-frequency dynamics of the deep layer network, in contrast to a dominant gamma frequency for the superficial layers, corresponding to the 25 msec subdivision of the overall alpha cycle. This includes direct electrophysiological recording (Luczak, Bartho, & Harris, 2013), local-field-potential recordings from superficial vs. deep layers (Buffalo et al., 2011; Maier, Adams, Aura, & Leopold, 2010; Maier, Aura, & Leopold, 2011; Spaak, Bonnefond, Maier, Leopold, & Jensen, 2012; Xing, Yeh, Burns, & Shapley, 2012; Bastos, Vezoli, Bosman, Schoffelen, Oostenveld, Dowdall, De Weerd, Kennedy, & Fries, 2015; Michalareas, Vezoli, van Pelt, Schoffelen, Kennedy, & Fries, 2016), and top-down-specific synchronization (von Stein, Chiang, & König, 2000; van Kerkoerle, Self, Dagnino, Gariel-Mathis, Poort, van der Togt, & Roelfsema, 2014). There are a variety of potential mechanisms behind the generation and synchronization of these 5IB bursts (Connors, Gutnick, & Prince, 1982; Lopes da Silva, 1991; Lorincz et al., 2009; Franceschetti et al., 1995; Saalmann et al., 2012). Furthermore, the pulvinar has been shown to drive alpha-frequency synchronization of cortical activity across areas in the alpha band (Saalmann et al., 2012). Behaviorally, there is extensive evidence of alpha-frequency effects on perception consistent with our framework (Nunn & Osselton, 1974; Varela, Toro, John, & Schwartz, 1981; VanRullen & Koch, 2003; Jensen et al., 2012).

The 6CT neurons exhibit regular spiking behavior, in contrast to the 5IB bursting (Thomson, 2010; Thomson & Lamy, 2007). Also, they do not have axonal branches that project to other cortical areas — the subpopulation that projects to the pulvinar only project there and not to other cortical areas (Petrof, Viaene, & Sherman, 2012), whereas there are other layer 6 neurons that do project to other cortical areas. This distinct connectivity is consistent with a specific role of this neuron type in generating predictions in the pulvinar. The 6CT synaptic inputs on pulvinar TRCs have metabatropic glutamate receptors (mGluR) that have longer time-scale temporal dynamics consistent with the alpha period (100 msec) and even longer (Sherman, 2014), and significantly more plasticity-inducing NMDA receptors compared to the 5IB projections (Usrey & Sherman, 2018). These properties are both consistent with the 6CT inputs driving a longer-integrated prediction signal that is subject to learning, whereas the 5IB are likely non-plastic and their effects are highly localized in time.

The 5IB inputs often have a distinctive *glomeruli* structures at their synapses onto pulvinar neurons, which contain a complete feedforward inhibition circuit involving a local inhibitory interneuron, in addition to the direct strong excitatory driver input (Wilson, Bose, Sherman, & Guillery, 1984). Computationally, this can provide a balanced level of excitatory and inhibitory drive so as to not overly excite the receiving neuron, while still dominating its firing behavior.

Although there are well-documented and widely-discussed burst vs. tonic firing modes in pulvinar neurons (Sherman & Guillery, 2006), there is not much evidence of these playing a clear role in the awake, behaving state, and as noted above the growing electrophysiological evidence shows a remarkable correspondence between cortical and pulvinar response properties across multiple different pulvinar areas in this awake state. Nevertheless, there may be important dynamics arising from these firing modes that are more subtle or emerge in particular types of state transitions that may have yet to be identified.

Predictions for Predictive Learning

The following discussion provides a basis for developing direct, testable predictions from this framework, organized according to a set of major themes, within which existing relevant data and experimental

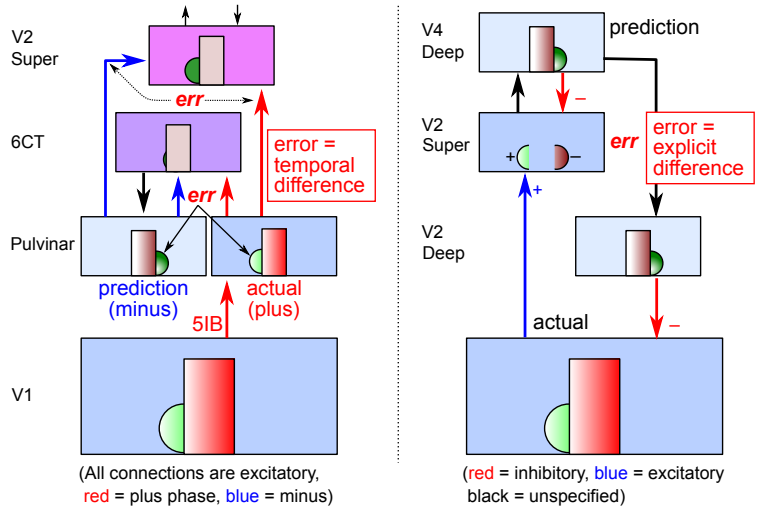

Figure 2: Comparison between the proposed thalamocortical temporal-difference predictive learning model (left side) versus the Bayesian-style explicit error (EE) coding model on the right (Rao & Ballard, 1999; Friston, 2010, Bastos et al., 2012), in a situation where the prediction is clearly erroneous (ball predicted to emerge on right, actually emerges on left). The EE model holds that superficial (2/3) error-coding neurons receive the prediction via a net inhibitory top-down projection from higher-level deep layer neurons, and an excitatory bottom-up projection representing the actual outcome, such that their activation represents the difference. To encode both signs of the error with positive-only spike rates (omissions, false alarms) two separate populations of EE neurons would be required. Unambiguous evidence of such EE coding neurons has not been found (Walsh et al, 2020). In contrast, error signals in our proposed framework remain as a temporal difference between the two states of prediction vs. outcome, *which enables all connectivity between cortical areas to be excitatory and always represent a positive encoding of either the prediction or outcome*. In contrast, under EE, after one error subtraction at the lowest level, only error signals are hypothesized to flow forward to higher layers, meaning that the prediction at the next layer is of this *error*, not the actual outcome. This is also inconsistent with extensive available data.

paradigms are discussed. In addition to potentially informing future experiments, this discussion also helps to clarify the exact nature of the theory in contrast to other frameworks for predictive learning, in particular those based on explicit error-coding (*EE*) neurons in the cortex (Rao & Ballard, 1999; Friston, 2010; Bastos, Usrey, Adams, Mangun, Fries, & Friston, 2012; Lotter et al., 2016) (Figure 2). As noted above, unambiguous evidence in support of these EE neurons is lacking (Walsh et al., 2020). Furthermore, a key feature of our approach is that by retaining error signals in the form of a temporal difference of activation states, *all connections between cortical layers are excitatory and represent the positive encoding of either the prediction or outcome state, at different levels of abstraction.* This is overwhelmingly supported by extensive electrophysiological data about the hierarchical organization of representations, e.g., in the visual object recognition pathway (Kobatake & Tanaka, 1994; Cadieu et al., 2014), and is consistent with the widely-supported biased competition model for excitatory top-down attentional effects (Desimone & Duncan, 1995; Reynolds, Chelazzi, & Desimone, 1999; Miller & Cohen, 2001).

By contrast, the EE approach requires net inhibitory top-down predictions, and it sends error signals forward, not positive representations of the actual state at a given level of abstraction. Thus a literal interpretation would have only error signals encoded beyond the lowest level, which is consistent with working implementations (Lotter et al., 2016). However, there are various different ways of reformulating the neural implementation of EE that can avoid some of these issues (Spratling, 2008; Bastos et al., 2012), but perhaps

this flexibility renders the framework difficult to falsify (Kogo & Trengove, 2015). In any case, an extensive treatment of the issues with EE is beyond the scope of this paper — we just focus on some of the core differences as a way to clarify the positive predictions of our framework by way of contrast.

Learning effects: Measuring effects on learning would clearly be the most direct test of a predictive learning theory. Critically, the type of learning must be matched to the pulvinar area in question. For example, in lower-level visual pulvinar driven by V1 inputs and predicted by V2, V4 top-down deep projections, it might be best to introduce novel sequential "physics" in movies at the alpha time scale, that are inconsistent with standard physics (e.g., changing properties such as gravity, inertia, or elasticity). At higher visual levels (e.g., IT cortex), it might be possible to use simple sequences of different objects, although it is not clear to what extent the hippocampus or prefrontal cortex might also contribute (Gavornik & Bear, 2014; Fiser, Mahringer, Oyibo, Petersen, Leinweber, & Keller, 2016). To distinguish pulvinar learning effects from pervasive motor learning supported by other brain areas, it would be most effective to directly measure activity in the pulvinar and / or associated perceptual neocortical areas, instead of involving overt behavioral performance.

- *Lesioning / inactivating the pulvinar should impair learning.* This is perhaps the most obvious prediction, but also challenging to test, for multiple reasons. First, much of the learning in posterior sensory cortex should take place early in development, requiring very early developmental interventions. Indeed, if a primary function of this system is for predictive learning *to train the neocortex*, then once this learning has been achieved, the contributions of this circuit may be much more strongly weighted toward its role in attention, as we discuss below. In any case, it might be difficult to uniquely attribute deficits to learning *per se*, given these additional attentional contributions. Furthermore, existing evidence suggests that inactivation of pulvinar has dramatic effects on cortical activity, raising further interpretational difficulties if learning deficits are seen (Zhou et al., 2016; Purushothaman, Marion, Li, & Casagrande, 2012).

- *Blocking synaptic plasticity in pulvinar should impair learning*. This can provide a way around the broader effects of inactivating the pulvinar. However, according to our computational model, most of the learning occurs in the neocortex as a result of pulvinar broadcasting the temporal difference error signal into neocortex. As noted above, only the 6CT projection should exhibit plasticity effects. And 6CT learning effects are most important very early in the learning process, as the pulvinar learns to map the neocortical representations into the space defined by the 5IB projections. Thus, these effects would require very early interventions.

- *Pulvinar, and the corresponding top-down 6CT neurons that project to it, should have predictive, anticipatory representations within 100 msec of a predictable stimulus.* This would constitute an important piece of consistent evidence, although not as causally definitive as direct learning effects. Interestingly, Barczak, O'Connell, McGinnis, Ross, Mowery, Falchier, and Lakatos (2018) recently showed that the auditory pulvinar in monkeys indeed exhibited earlier predictive activity than A1, using a carefully controlled auditory sequence that had no first-order acoustic differences from a background noise signal. The deep layers of higher auditory areas that contribute to the formation of the pulvinar prediction were not recorded in this study, so their role in generating the prediction could not be determined. Thus, this paradigm would appear ideal for further tests of this predictive learning framework.

  Another study demonstrated modulation of pulvinar activity by *confidence* driven by relative ambiguity in a random dot motion categorization task (Komura et al., 2013). Critically for the present framework, this confidence modulation only emerged in the period after the first 100 msec of processing, and manifested as a positive correlation with confidence (i.e., more unambiguous stimuli

resulted in higher firing rates). We can interpret this as reflecting an ongoing generative prediction of the stimulus signal, with stronger firing associated with more unambiguous predictions based on the current internal representation. Note that this directionality is the opposite of EE neurons, which would presumably increase with increasing error / ambiguity in the prediction. Interestingly, inactivation of these pulvinar neurons resulted in a substantial (200%) increase in opt-out choices on the most ambiguous stimuli, suggesting a level of metacognitive awareness of the pulvinar signal (or at least a direct effect of pulvinar on relevant metacognitive processes). Predictive accuracy would be an ideal source of metacognitive confidence signals across a wide range of domains, suggesting another important contribution of pulvinar even after initial learning.

- *Higher-level cortical activation of predicted stimuli should excite associated lower-level representations, in anticipation of stimulus onset.* This prediction is less specific to the pulvinar-based framework, but it does potentially differentiate from the EE models, and we include it here for that purpose, even though it is actually one of the most consistent positive findings in the literature reviewed by Walsh et al. (2020), and is thus is more of a confirmed prediction at this point. For example, the widely replicated predictive remapping effect is of this nature (Duhamel, Colby, & Goldberg, 1992; Cavanagh, Hunt, Afraz, & Rolfs, 2010). Whereas our framework predicts fully excitatory effects of top-down predictions (consistent with the data), EE instead requires that these predictions should have a net inhibitory effect on error-coding neurons (see the expectation suppression discussion below for more details). The top-down effect of predictions onto lower-level prediction-coding neurons in the EE paradigm is less clear, but could be excitatory. Thus, more detailed tests that distinguish these different populations would be needed to fully differentiate these models.

Nature of the Prediction Error Signal: A central feature of our thalamocortical framework is that prediction error signals are represented as temporal differences between prediction and outcome states, which contrasts directly with the EE model of neurons directly coding for error.

- *Errors should be encoded as a new state of activity representing the unexpected outcome, immediately following any prior predictions.* This prediction has the "unfortunate" status of being overwhelmingly supported by a wide range of existing data — new stimuli almost always drive a new state of activity through feedforward pathways. Furthermore, as noted earlier, the prediction state is essentially defined as the state prior to an unexpected stimulus, by the hypothesized nature of the temporal-difference error-driven learning rule, triggered by the 5IB burst activation. In this way, temporal-difference learning represents a very "natural" form of error-driven learning that follows directly from the basic dynamics of the network. Nevertheless, in contrast with the EE framework, our model provides a biological mechanism that explains how widely observed predictive learning phenomena can arise without requiring empirically unsupported EE coding neurons.

- *Temporal differences on an alpha cycle timescale actually drive synaptic plasticity in an error-driven learning manner, in neocortical pyramidal neurons and in 6CT inputs to pulvinar*. That is, if a pre / post pair of neurons across a synapse is more active in the prediction than the subsequent outcome, the synapse should experience LTD (long term depression), and vice-versa if the activity pattern is reversed (long term potentiation, LTP, for more activity in outcome than prediction). Furthermore, if activity is essentially stable across both prediction and outcome phases, then weights should not change (modulo a small level of Hebbian learning). This should be directly testable using current synaptic plasticity experimental methods, and is perhaps the single most important empirical test of this entire framework, and it also underlies many other current approaches to error-driven learning in the brain (Bengio et al., 2017; Whittington & Bogacz, 2019; Lillicrap et al., 2020). One general consideration is the extent to which a awake *in vivo* preparation would be required to capture all the

neuromodulatory and other factors present when this learning normally takes place. Some suggestive evidence in such a preparation is generally consistent with a sensitivity to relatively short-term temporal dynamics (Lim et al., 2015), although these results lacked the direct measurement of individual neural activity across a synapse.

- *Expectation suppression is due to neural adaptation and other dynamics, not explicit error coding.* The strongest evidence typically cited in favor of EE is *expectation suppression*, where expected inputs elicit suppressed neural responses (Summerfield, Trittschuh, Monti, Mesulam, & Egner, 2008; Todorovic, van Ede, Maris, & de Lange, 2011; Meyer & Olson, 2011; Bastos et al., 2012). However, multiple comprehensive reviews conclude that it is difficult to distinguish these effects from the neural adaptation effects that underlie the well-documented *repetition suppression* effect (Walsh et al., 2020; Vinken & Vogels, 2017; Kok et al., 2012; Lee & Mumford, 2003). Furthermore, detailed single-neuron level recordings are the least likely to show these effects — instead, they are most evident in aggregate signals such as the BOLD response in fMRI, suggesting that they are likely due to population-level differences in activity, rather than individual explicit error coding neurons. As noted earlier, accurately predicted outcomes would result in a continued adaptation of the neural response carrying over from the prediction to the outcome state, whereas unexpected outcomes would be associated with two distinct patterns of activity over a given area: first the prediction and then the outcome. The outcome state would not be subject to the prior neural adaptation effects, and furthermore the time-integrated aggregate activity over these two patterns would be greater compared to the single activity state associated with an accurately predicted outcome. Thus, our model explains expectation suppression without invoking EE neurons, meaning that considerably more detailed and replicable experimental paradigms using single-neuron resolution techniques are needed to distinguish EE from our framework.

Alpha Frequency Effects: The 5IB bursting is known to have an intrinsic 100 msec period, which, along with other thalamocortical network effects, defines the predictive learning cycle in our framework. There is a large and growing literature about the behavioral and biological properties of alpha frequency oscillations in thalamocortical networks. Much of this literature is generally consistent with the idea that alpha is important for predictive learning and sensory processing, but it is also clear that there are important attentional modulation effects associated with alpha as well. We will consider these issues later as well, but briefly it is clear that reducing cortical activity and excitability reduces the Poisson random jittering associated with direct excitatory synaptic inputs, and thus enables intrinsic oscillatory drivers to exert a much stronger synchronizing effect on cortical firing, increasing associated EEG power (Zhou et al., 2016; Klimesch, Sauseng, & Hanslmayr, 2007; Jensen & Mazaheri, 2010). Thus, a reduction of attentional focus brings an associated reduction in activity and excitability, and hence an increase in EEG alpha power. This, however, does not directly disconfirm the idea that predictive learning and attention operate within circuits driven by 5IB alpha-bursting inputs. Instead, it is important to look at other measures such as intertrial phase coherence (ITPC), which more cleanly reflect just the direct effects of these 5IB bursts. Here, the evidence is much more consistent in showing increased ITPC for conditions associated with increased prediction and attention (MaryPapersMayer, Schwiedrzik, Wibral, Singer, & Melloni, 2016).

- The extensively studied event related potential (ERP) components unfold over waves defined in 100msec increments, which suggests that these may be related to the underlying alpha frequency dynamics (Makeig, Westerfield, Jung, Enghoff, Townsend, Courchesne, & Sejnowski, 2002; Gruber, Klimesch, Sauseng, & Doppelmayr, 2005; Klimesch, 2011).

- A number of studies have shown alpha-synchronized predictive dynamics (Mayer et al., 2016). TODO more!

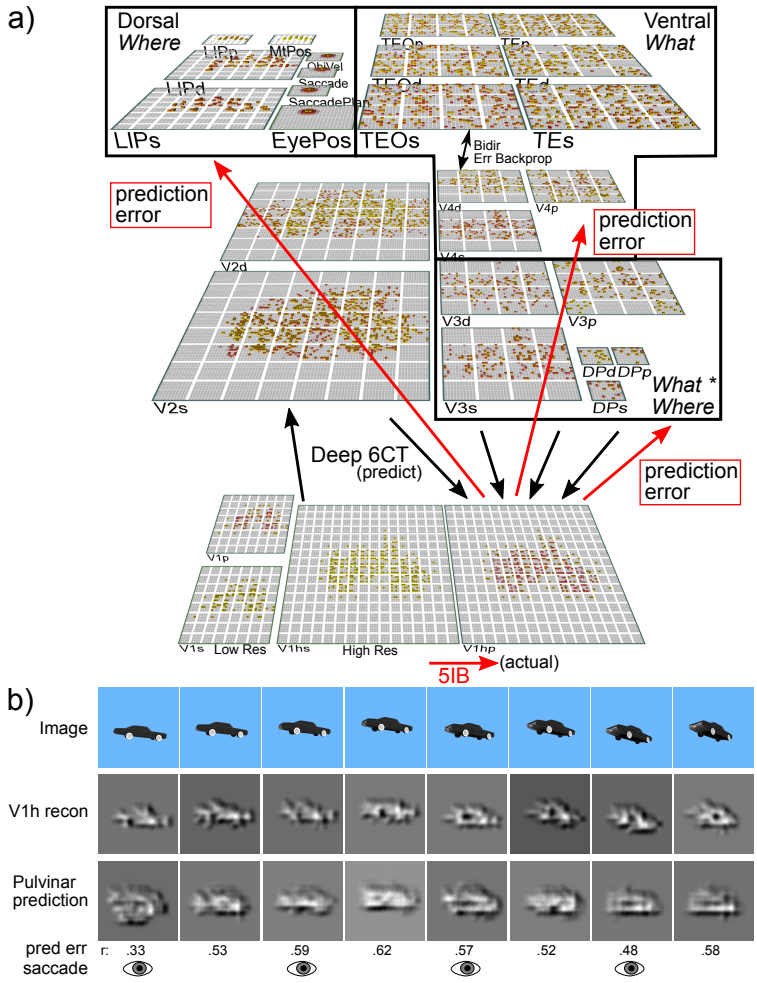

Figure 3: **a)** The *What-Where-Integration, WWI* deep predictive learning model. The dorsal *Where* pathway learns first, using easily-abstracted *spatial blobs*, to predict object location based on prior motion, visual motion, and saccade efferent copy signals. This drives strong top-down inputs to 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. Suffixes: s = 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 (*V1h 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.

- Eye fixation dynamics are also strongly entrained at alpha intervals. 200 msec dwell, exactly 2 alpha cycles, minimum required for initial processing of unexpected input followed by another cycle where input is then predicted based on resulting internal state.

Predictive Learning of Object Categories in IT Cortex

Now we turn to our implementation of the proposed thalamocortical predictive error-driven learning framework, in a large-scale model of visual predictive learning (Figure 3). Our second major objective, and a critical question for predictive learning, is whether the model can develop high-level, abstract ways of representing the raw sensory inputs, while learning from nothing but predicting these low-level visual inputs. Existing predictive-learning models based on error backpropagation (Lotter et al., 2016JakeAddOthers) have not demonstrated the development of abstract, categorical representations. Instead these models have generally shown that predictive learning can be a useful method for pretraining networks that are subsequently trained using human-generated labels. But for a biologically and ecologically plausible model of learning, it is important to determine what can be learned in the absence of any such external supervised training inputs.

To determine if our biologically based predictive learning model 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 for their overall shape diversity 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 3b). 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 the signature predictive remapping phenomenon (Duhamel et al., 1992; Cavanagh et al., 2010), and introduces an additional predictive-learning challenge. The *only* learning signal available to the model was the prediction error generated by the temporal difference between what it predicted to see in the V1 input in the next frame and what was actually seen.

As described in detail in the supporting information, our model was constructed to capture critical features of the visual system, including the major division between a dorsal *Where* and ventral *What* pathway (Ungerleider & Mishkin, 1982), and the overall hierarchical organization of these pathways derived from detailed connectivity analyses (Rockland & Pandya, 1979; Felleman & Van Essen, 1991; Markov, Vezoli, Chameau, Falchier, Quilodran, Huissoud, Lamy, Misery, Giroud, Ullman, Barone, Dehay, Knoblauch, & Kennedy, 2014b; 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, 2014a). In addition to these biological constraints, we conducted extensive exploration of the connectivity and architecture space, and found a remarkable convergence between what worked functionally and the known properties of these pathways (O'Reilly, Wyatte, & Rohrlich, 2017). For example, the feedforward pathway has projections from lower-level superficial layers to superficial layers of higher levels, while feedback originated in both the superficial and deep and projected back to both (Rockland & Pandya, 1979; Felleman & Van Essen, 1991). Also, consistent with the core features of the pulvinar pathways discussed above, deep layer predictive (6CT) inputs originated in higher levels, while driver (5IB) inputs originated in lower levels. For simplicity we organized the model layers in terms of these driver inputs, whereas the topographic organization of pulvinar in the brain is organized more according to the 6CT projection loops (Shipp, 2003). Another important set of parameters are the strength of deep-layer self recurrent projections, which influence temporal integration time window, producing a simple biologically-based version of *slow feature analysis* (Wiskott & Sejnowski, 2002; Foldiak, 1991). We followed the biological data suggesting that recurrence increases progressively up the visual hierarchy (Chaudhuri, Knoblauch, Gariel, Kennedy, & Wang, 2015). Results from various informative model architecture and parameter manipulations are discussed below after the primary results from the standard intact model. Learning curves and other model details are shown in the supporting information.

To analyze the learned representations, we performed a representational similarity analysis (RSA) on the activity patterns at each layer in the model, and found that the highest IT layer (TE) produced a systematic

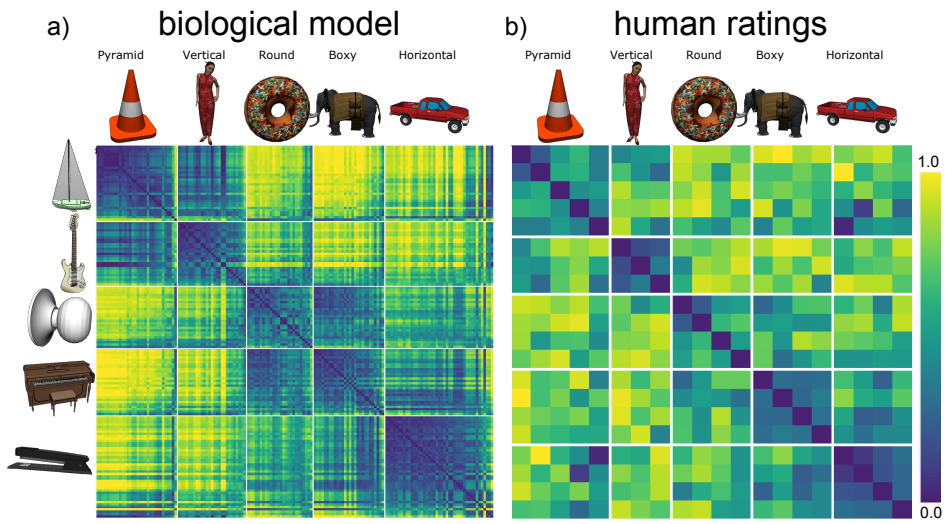

Figure 4: **a)** Category similarity structure that developed in the highest layer, TE, of the biologically based predictive learning model, showing *1-correlation* similarity of the TE representation for each 3D object against every other 3D object (156 total objects). Blue cells have high similarity, 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), and clearly corresponded to the shown shape-based categories, with exemplars from each category shown. Also, 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 1b) to capture coarse perception in model, aggregated by 20 basic-level categories (156 x 156 matrix was too large to sample densely experimentally). 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), indicating that human raters used the same shape-based category structure.

organization of the 156 3D objects into 5 categories (Figure 4a). These categories clearly correspond to the overall shape of the objects, as shown by the object exemplars in the figure (pyramid-shaped, vertically-elongated, round, boxy / square, and horizontally-elongated). Given that the model only learns from a passive visual experience of the objects, it has no access to any of the richer interactive multi-modal information that people and animals would have. Furthermore, as evident in Figure 3b, the relatively low resolution of the V1 layers (required to make the model tractable computationally) means that complex visual details are not reliably encoded (and even so, are not generally reliable across object exemplars), such that the overall object shape is the most salient and sensible basis for categorization.

Although these object categories appeared sensible to us, we ran a simple experiment to test whether a sample of 30 human participants would use the same category structure in evaluating the pairwise similarity of these objects. Figure 4b shows the results, confirming that indeed this same organization of the objects emerged in their similarity judgments. These judgments were based on the V1 reconstruction as shown in Figure 3b to capture the model's coarse-grained perception; see supporting information for methods and further analysis.

Another important question concerns the way in which these categorical representations at the highest level in the *What* pathway of the model emerge through the deep hierarchy of layers progressing upward from V1. This has been investigated in recent comparisons between monkey electrophysiological recordings

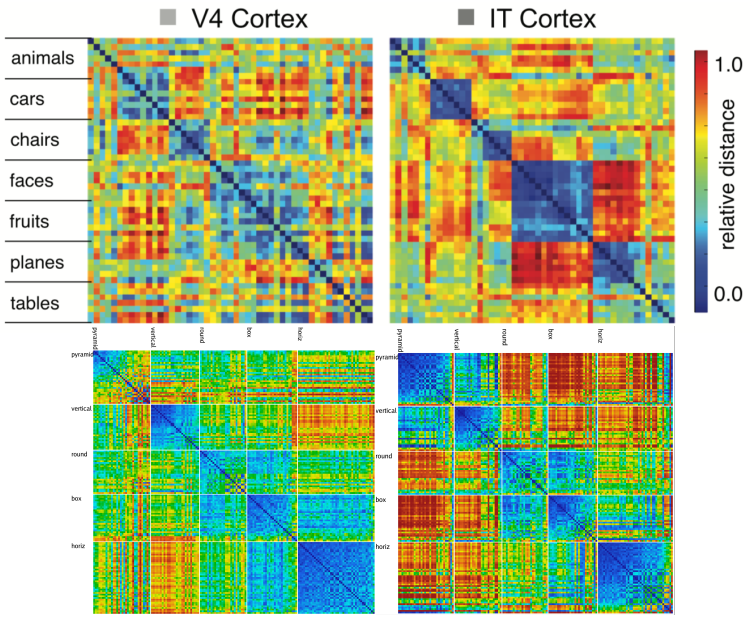

Figure 5: Comparison of progression from V4 to IT in macaque monkey visual cortex (top row, from Cadieu et al., 2014) versus same progression in model (replotted using comparable color scale). Although the underlying categories are different, and the monkeys have a much richer multi-modal experience 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.

and deep convolutional neural networks (DCNNs), which provide a reasonably good fit the the overall progressive pattern of increasingly categorical organization (Cadieu et al., 2014). However, these DCNNs were trained using error backpropagation with large datasets of human-labeled object categories, and it is perhaps not too surprising that the higher layers closer to these category output labels exhibited a greater degree of categorical organization — this is an intrinsic property of the error backpropagation gradients. In contrast, because the only source of learning in our model comes from prediction errors over the V1 input layers, the graded emergence of an object hierarchy here would reflect a truly self-organizing learning process. Figure 5 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, showing that our model replicates the essential qualitative hierarchical progression in the brain. After presenting a few more analyses, we explore the critical factors that lead to this result.

We can more precisely quantify the emergence of categorical representations in our model by computing the meta-similarity across the similarity matricies computed at each layer in the network (Figure 6). This shows the extent to which the similarity matrix across objects in one layer is itself similar to the object similarity matrix in another layer, in terms of a correlation measure across these similarity matricies. Starting from either V1 compared to all higher layers, or TE compared to all lower layers, we found a consistent pattern of progressive emergence of the object categorization structure in the upper IT pathway (TEO, TE).

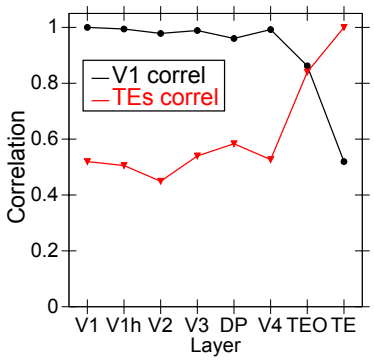

Figure 6: Emergence of abstract category structure over the hierarchy of layers. Red line = correlation similarity between the TE similarity matrix (shown in Figure 2a) and all layers; black line shows correlation similarity between V1 against all layers (1 = identical; 0 = orthogonal). Both show that 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.

Critically, this analysis shows that the IT category structure is significantly different from that present 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. 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 4a by virtue of the close similarity across multiple objects within the same category.

In summary, the model learned an abstract category organization that reflects the overall visual shapes of the objects as judged by human participants, in a way that is invariant to the differences in motion, rotation, and scaling that are present in the V1 visual inputs. We are not aware of any other model that has accomplished this signature computation of the ventral *What* pathway in a purely self-organizing manner operating on realistic 3D visual objects, without any explicit supervised category labels, much less using a learning algorithm directly based on detailed properties of the underlying biological circuits in this pathway.

Backpropagation Comparison Models

To help discern some of the factors that contribute to the categorical learning in our model, and provide a comparison with more widely-used error backpropagation models, we tested a backpropagation-based (Bp) version of the same *What* vs. *Where* architecture as our biologically-based predictive error model, and we also tested a standard *PredNet* model (Lotter et al., 2016) with extensive hyperparameter optimization (see SI). Due to the constraints of backpropagation, we had to eliminate any bidirectional connectivity loops in the Bp version, but we were able to retain a form of predictive learning by configuring the V1p pulvinar layer as the final target output layer, with the target being the next visual input relative to the V1 inputs. As shown in Figure 7, the highest layers of the Bp model 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

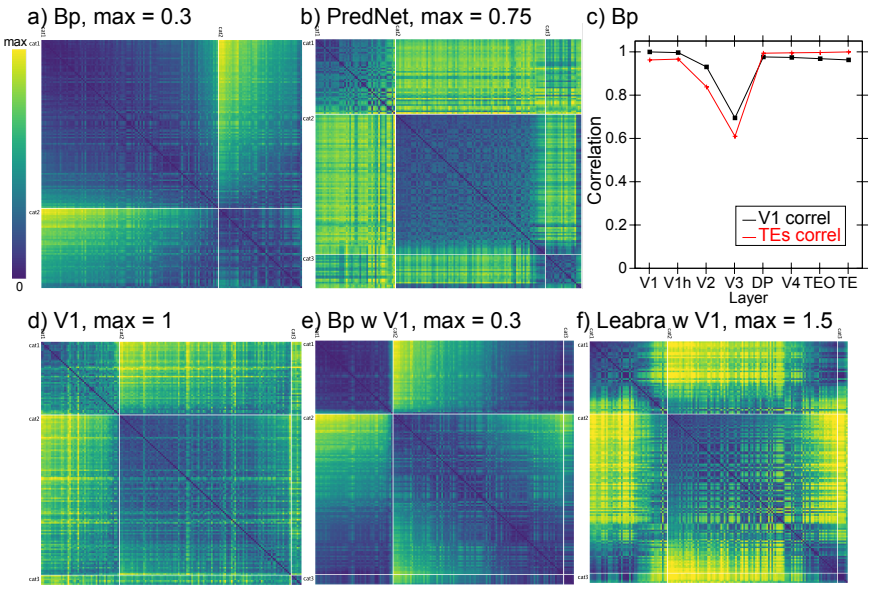

Figure 7: **a)** Best-fitting category similarity for TE layer of the backpropagation (Bp) model with the same What / Where structure as the biological model. Only two broad categories are evident, and the lower *max* distance (0.3 vs. 1.5 in biological model) means that the patterns are highly similar 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.

(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. Furthermore, the recurrent connections within the TEO and TE layers likely play an important role by biasing the temporal dynamics toward longer persistence (Chaudhuri et al., 2015). 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 representations coalesce (and results below already indicate its importance). Inhibition, especially in combination with Hebbian learning,

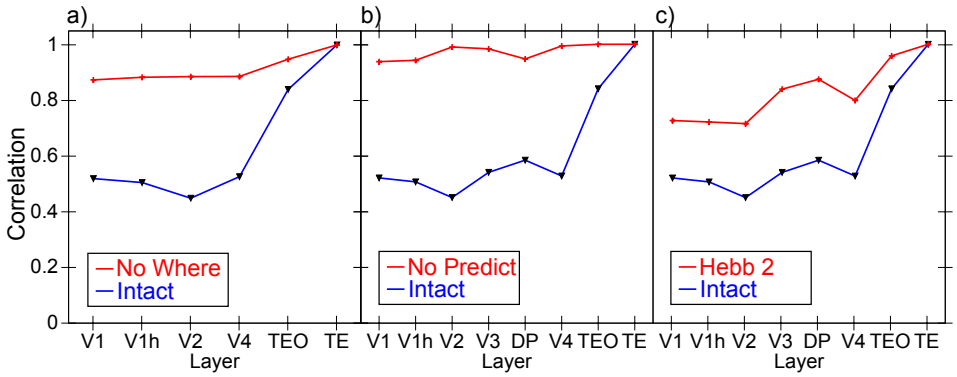

Figure 8: Effects of various manipulations on the extent to which TE representations differentiate from V1. For all plots, *Intact* is the same result shown in Figure 5 from the intact model for ease of comparison. All of the following manipulations significantly impair the development of abstract TE 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.

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 significantly more computationally intensive and unstable forms of fully recurrent backpropagation (Williams & Zipser, 1992; Pineda, 1987). Furthermore, Hebbian learning requires dynamic inhibitory competition which is difficult to incorporate within the backpropagation framework.

Architecture and Parameter Manipulations

Figure 8 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 8a shows that, indeed, when the *Where* pathway is lesioned, the formation of abstract categorical representations in the intact *What* pathway is significantly impaired. Figure 8b 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 8c shows the impact of reducing Hebbian learning, which impairs category learning as expected.

Predictive Behavior

A signature example of predictive behavior at the neural level in the brain is the *predictive remapping* of visual space in anticipation of a saccadic eye movements (Duhamel et al., 1992; Colby, Duhamel, & Goldberg, 1997; Gottlieb, Kusunoki, & Goldberg, 1998; Nakamura & Colby, 2002; Marino & Mazer, 2016) (Figure 9a). Here, parietal neurons start to fire at the *future* receptive field location where a currently-visible stimulus will appear after a planned saccade is actually executed. Remapping has also been shown for border ownership neurons in V2 (O'Herron & von der Heydt, 2013) and in area V4 (Neupane, Guitton, & Pack, 2016). These are examples, we believe, of a predictive process operating throughout the neocortex to predict what will be experienced next. A major consequence of this predictive process is the perception of a stable, coherent visual world despite constant saccades and other sources of visual change.

Figure 9b shows that our model exhibits this predictive remapping phenomenon. Specifically, LIP, which is most directly interconnected with the saccade efferent copy signals, is the first to predict the new location, and it then drives top-down activation of lower layers. This top-down dynamic is consistent with the account of predictive remapping given by Wurtz (2008) and Cavanagh et al. (2010), who argue that the key remapping takes place at the high levels of the dorsal stream, which then drive top-down activation of the predicted location in lower areas, instead of the alternative where lower-levels remap themselves based on saccade-related signals. The lower-level visual layers are simply too large and distributed to be able to remap across the relevant degrees of visual angle — the extensive lateral connectivity needed to communicate across these areas would be prohibitive.

Discussion

We have hypothesized a novel computational function for the distinctive features of thalamocortical circuits (Sherman & Guillery, 2006), as supporting a specific form prediction-error driven learning, where predictions arise from the numerous top-down layer 6CT projections into the pulvinar, and the strong sparse driving 5IB inputs supply the bottom-up sensory-driven outcome. The phasic bursting nature of the 5IB inputs results in a natural temporal-difference error signal of prediction followed by outcome, consistent with extensive neural recording data. This temporal dynamic is also essential for enabling predictions to be generated without contamination from current sensory inputs, and predicts a characteristic alpha frequency prediction cycle based on the 10hz bursting cycle of the 5IB inputs, consistent with the pervasive influence of alpha on perception and neural dynamics. In short, the hypothesized predictive learning function fits remarkably well with a number of well-established properties of these thalamocortical circuits, and we also provided a number of additional predictions that could be tested to further evaluate this theory, especially in contrast to the widely-discussed alternative of explicit error coding neurons, which have not been unambiguously supported across a range of empirical studies (Walsh et al., 2020).

Furthermore, we implemented this theory in a large scale model of the visual system, and 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 environment. We found that comparison models based on standard error backpropagation learning did not learn a

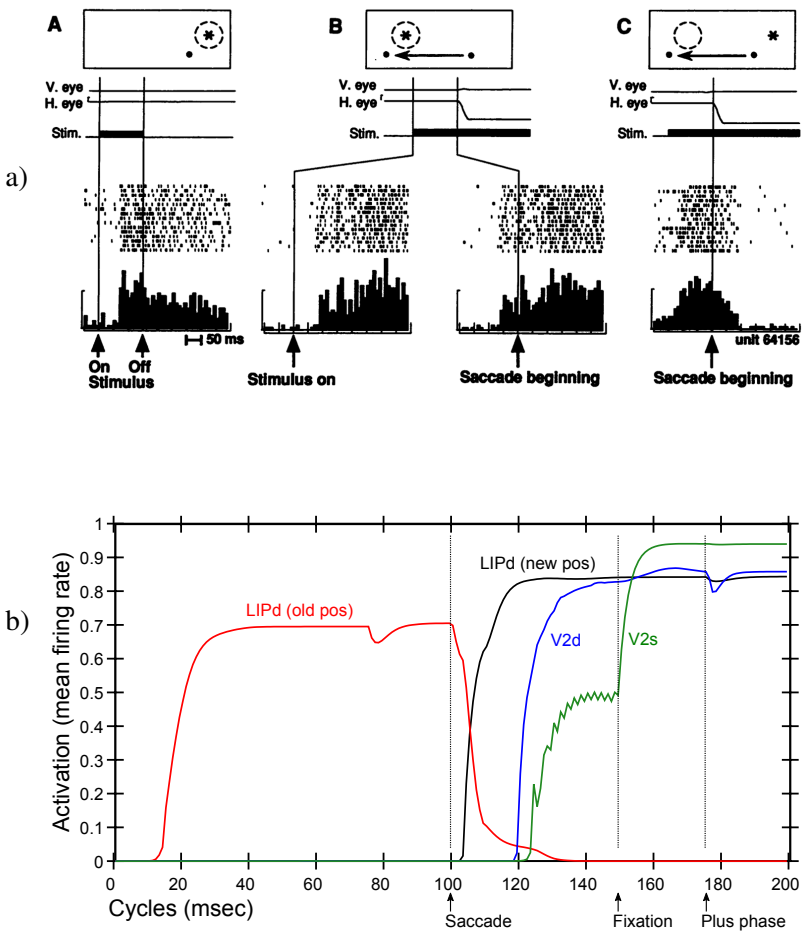

Figure 9: Predictive Remapping. **a)** Original remapping data in LIP from Duhamel et al (1992). A) shows stimulus (star) response within receptive field (dashed circle) relative to fixation dot (upper right of fixation). B) Just prior to monkey making a saccade to new fixation (moving left), stimulus is turned on in receptive field location that *will be* upper right of the new fixation point, and the LIP neuron responds to that stimulus in advance of the saccade completing. The neuron does not respond to the stimulus in that location if it is not about to make a saccade that puts it within its receptive field (not shown). This is predictive remapping. C) response to the old stimulus location goes away as saccade is initiated. **b)** Data from our model, from individual units in LIPd, V2d, and V2s, showing that the LIP deep neurons respond to the saccade first, activating in the new location and deactivating in the old, and this LIP activation goes top-down to V3 and V2 to drive updating there, generally at a longer latency and with less activation especially in the superficial layers. When the new stimulus appears at the point of fixation (after a 50 msec saccade here), the *primed* V2s units get fully activated by the incoming stimulus. But the deep neurons are insulated from this superficial input until the plus phase, when the cascade of 5IB firing drives activation of the actual stimulus location into the pulvinar, which then reflects up into all the other layers.

categorical structure that went beyond the surface similarity present in the visual input layers, and future work is focused on narrowing down the specific mechanisms required to drive this learning.

In addition to the predictive learning functions of the deep / thalamic layers, these same circuits are also likely critical for supporting powerful top-down attentional mechanisms that have a net multiplicative effect on superficial-layer activations (Bortone, Olsen, & Scanziani, 2014; Olsen, Bortone, Adesnik, & Scanziani, 2012; Bortone et al., 2014; Olsen et al., 2012). The importance of the pulvinar for attentional processing has been widely documented (e.g., LaBerge & Buchsbaum, 1990; Bender & Youakim, 2001; Saalmann et al., 2012), and there is likely an additional important role of the thalamic reticular nucleus (TRN), which can contribute a surround-inhibition contrast-enhancing effect on top of the incoming attentional signal from the cortex (Crick, 1984; Pinault, 2004; Wimmer, Schmitt, Davidson, Nakajima, Deisseroth, & Halassa, 2015). In our computational framework, these attentional modulation signals cause the iterative constraint satisfaction process in the superficial network to focus on task-relevant information while down-regulating responses to irrelevant information — in the real world, there are typically too many objects to track at any given time, so predictive learning must be directed toward the most important objects. A subsequent paper will explore the attentional aspects of the DeepLeabra model and its synergy with the predictive learning aspect.

Synergy between attention and prediction: (Richter & de Lange, 2019) – but in context of expectation suppression as discussed above.

(Keller & Mrsic-Flogel, 2018) – alternative predictive model??

TODO: more specific effects.

TODO: Kastner papers. (Halassa & Kastner, 2017) (Fiebelkorn, Pinsk, & Kastner, 2018) (Fiebelkorn & Kastner, 2019)

(Jaramillo, Mejias, & Wang, 2019) – integrative large-scale theory – read fk19a preview first.

Relative to existing machine-learning-based 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 cognitive processes such as reasoning and problem-solving (Rougier, Noelle, Braver, Cohen, & O'Reilly, 2005; O'Reilly, Petrov, Cohen, Lebiere, Herd, & Kriete, 2014). 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.

Supplemental Information

All of the materials described here, including the experimental study, the computational models, and the code to perform the representational similarity analysis, are all available on our github account at: `https://github.com/ccnlab/deep-obj-cat` For the computational models in particular, the most complete understanding can only be had by directly examining the code for the models, as there are a number of details that are not efficiently captured in this suppplementary materials text.

Representational Similarity Analysis Methods

The different representations being compared here are:

Leabra: The DeepLeabra (biological model) TE layer representations (specifically TEs = superficial – results are very similar for deep as well).

Bp: The TEs layer representations from the backpropagation version of biological model, including *What, Where* and *What * Where* integration layers, trained with the V1p and V1hp (low and high resolution pulvinar) layers as final output layers, using the time t target pattern from the $t-1$ input (i.e., as a predictive network).

V1: The gabor-filtered representation of the visual input to both of the above models, which was identical across them.

PredNet: Highest layer (6th Layer) of the PredNet architecture.

Expt: Similarity matrix constructed from human pairwise similarity judgments (see *Behavioral Experiment Methods*).

An optimal category cluster can be defined as one that has high within-cluster similarity and low between-cluster similarity. This can be operationalized by the *contrast* distance metric, based on a 1-correlation (*correlation distance*) measure, as the difference between the average within-cluster similarity and the average between-cluster similarity:

$$cd = \langle 1 - r_{in} \rangle - \langle 1 - r_{out} \rangle \tag{1}$$

With distance-like 1-correlation values, this contrast distance should be minimized (it is typically negative), or equivalently the contrast on raw correlation values can be maximized (it is typically a positive number – just the sign flip of distance value). We refer to the positive numbers and maximization here as that is more natural.

Starting with an initial set of clusters, a permutation-based hill-climbing strategy was used to determine a local minimum in this measure: each item was tested in each of the other possible categories, and if that configuration reduced the overall average contrast distance metric across all items, then it was adopted and the process iterated until no such permutation improved the metric. This algorithm can only decrease the number of clusters (by moving all items out of a given cluster), so different numbers of initial clusters can be used to search the overall space.

Figure 10 shows the resulting categories. The Bp model converged on the same cluster state from all starting configurations tested, varying from 5 to 2 initial categories. This is the cluster set shown in Figure 5a of the main paper, and has an average contrast distance (*acd*) of 0.0838 (this is relatively low because the patterns were overall quite similar). Likewise, the V1 patterns (which were the same across Leabra and Bp models) reliably converged on the same pattern (shown in Figure 5d), with *acd* = 0.2448.

Centroid

1. pyramid
   - banana
   - layercake
   - trafficcone
   - sailboat
   - trex
2. vertical
   - person
   - guitar
   - tablelamp
3. round
   - doorknob
   - donut
3. round cont'd
   - handgun
   - chair
4. box
   - slrcamera
   - elephant
   - piano
   - fish
5. horiz
   - car
   - heavycannon
   - stapler
   - motorcycle

Bp

1. cat1
   - banana
   - layercake
   - trafficcone
   - sailboat
   - trex
   - person
   - guitar
   - tablelamp
   - doorknob
   - donut
1. cat1 cont'd
   - handgun
   - chair
   - slrcamera
   - elephant
   - piano
   - fish
   - car
2. cat2
   - heavycannon
   - stapler
   - motorcycle

V1

1. cat1
   - trafficcone
   - sailboat
   - person
   - guitar
   - tablelamp
   - chair
2. cat2
   - layercake
   - trex
   - doorknob
   - donut
2. cat2 cont'd
   - handgun
   - slrcamera
   - elephant
   - piano
   - fish
   - car
   - heavycannon
   - stapler
   - motorcycle
3. cat3
   - banana

PredNet

1. cat1
   - trafficcone
   - sailboat
   - person
   - guitar
   - tablelamp
   - layercake
2. cat2
   - trex
   - donut
   - banana
   - handgun
2. cat2 cont'd
   - slrcamera
   - elephant
   - fish
   - car
   - heavycannon
   - stapler
   - motorcycle
3. cat3
   - chair
   - doorknob
   - piano

Figure 10: Shape categories used for similarity matrix plots in main paper. *Centroid* shape categories are near-best for both the Leabra model and the Expt results, and fit our visual intuitions about overall shape. *Bp* are reliably optimal for Bp model from all starting points. *V1* are reliably optimal for V1 inputs, and also were close to the best for the Bp and PredNet layer 6 representations. *PredNet* are best stable solution for PredNet layer 6.

For the PredNet layer 6 representations, starting from the V1 categories gave the best results of any other set (*acd* = 0.1967), and a few permutations resulted in a reliable solution that was arrived at from all other 3 category starting points tested, shown in Figure 10 (*acd* = 0.2820). This indicates that PredNet did not go much beyond the structure present in the input, even though it did not use the V1 gabor filtering used in the Leabra and Bp models (i.e., this V1-level encoding well-captures the structure of the visual inputs in general). The PredNet pixel and layer 1 representations both converged on essentially a single monolitic category with very low acd (0.0018, 0.0013).

For the Leabra TE representations, we found a set of *centroid* shape categories that are near-best when considering both the Leabra model and the results from the human behavioral experiment (Expt). Starting from these categories, the permutation analysis converged on reducing the size of the vertical and round

categories to one item each, over a sequence of 5 steps. This is consistent with the observation from Figure 3a that there are three broader categories within which the 5 finer-grained categories are embedded (i.e., vertical and pyramid are overall similar to each other, as are round and box). Nevertheless, our initial visual intuition about the broad shape categories, along with a bias against having single-item categories, reinforced the use of the finer-grained centroid selection. The average contrast difference of our centroid selection is 0.5071, while the maximal result from the permutation was 0.5526, which is a relatively small proportional difference.

Furthermore, once we had collected the human experimental data (*Expt*), it was clear that it strongly coincided with our original shape intuitions, and with the finer-grained 5 category centroid structure. Starting from the centroid categories, the maximal permutation made only 3 changes, moving trex (T-rex) and handgun into the horizontal category, and chair into the pyramid, going from a distance score of 0.3083 to 0.3225, which is a relatively small improvement. However, using the maximal *Expt* clusters directly on the Leabra model gives a lower *acd* measure of 0.3745 (compared to 0.5071 for centroid), so the centroid categories represent a good middle-ground between experiment and the model, and this strong shared similarity structure with near-optimal cluster structures confirms that the model and people are encoding largely the same information.

In contrast, if we organize the experiment similarity matrix using the Bp categories, it produces a very poor average contrast distance measure of 0.0643 (compared to 0.3083 for the centroid categories), strongly suggesting that people's shape representations are not compatible with that simple structure.

Another approach to determining clusters from similarity matrices, *agglomerative clustering*, starts with all items as singletons, and iteratively combines the closest two into a new cluster. The results for the Leabra and Expt similarity matrices are shown in Figure 11, which has also color-coded the items in terms of their category status according to the centroid structure. Due to a strong history dependency in the clustering process, and the indeterminacy of reducing a high-dimensional similarity structure down to two dimensions, structure beyond the leaf level is not very reliable (ties are also broken by a random number generator), but nevertheless you can clearly see that in both cases items from the same cluster are almost always together as leaves in the plots. This then provides additional converging support for the idea that the model is learning the same kind of shape categories as people have.

Behavioral Experiment Methods

The behavioral experiment was conducted on Amazon.com's MTurk web platform under University of Colorado IRB approval (19-0176), using 30 participants each categorizing up to 800 image pairs as shown in Figure 12, using the standard *simple image categorization* framework with a lightly customized script. Objects were drawn from the 156 3D object set, but data was aggregated in terms of the 20 basic-level categories (car, stapler, etc) because we could not sample all 156 x 156 object pairs. Thus, the resulting data was aggregated for each category pair in terms of the proportion of times when that pair was selected when presented.

The individual images were produced by reconstructing from the V1 transform that the computational model used in its high resolution V1 input layer, to give human participants as similar of an experience as possible to how the model "saw" the objects, and to reduce the influence of existing semantic knowledge which was entirely missing in our model (Figure 12).

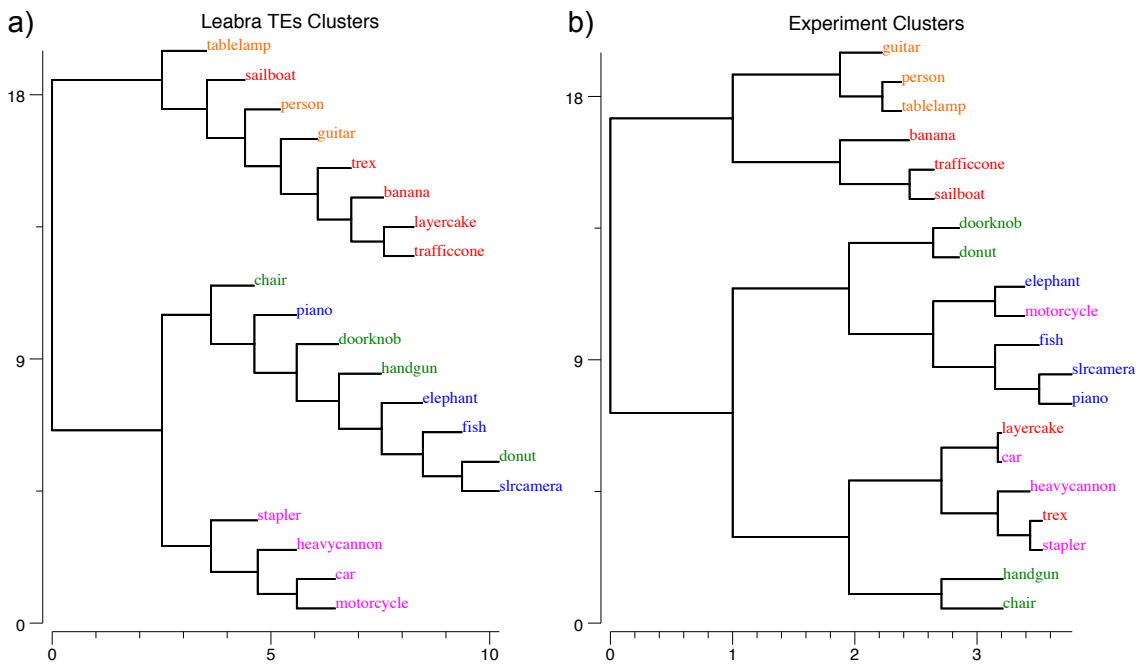

Figure 11: Agglomerative clustering on the Leabra and Expt representations, with the centroid categories color coded. The most reliable information from this is the leaf-level groupings, as the rest of the structure is indeterminante and history dependent in reducing higher-dimensional structure down to a 2D plot. Both cluster plots show a strong tendency to group leaf items together in the same centroid categories, with a few exceptions in each case. Also, the Leabra plot nicely captures the broader 3-category structure evident in the similarity matrix plots, within which the 5 finer-grained centroid categories are organized. Overall, this provides further confirmation that the model and the human subjects are organizing the shapes in largely the same way.

Figure 12: Example stimulus from the behavioral experiment, using the V1 reconstruction of the actual input images presented to the model, to better capture the coarse-grained perception of the model. Subjects were requested to choose which of the two pairs, Left or Right, was most similar in terms of *overall shape*.

Biological Model Methods

This section provides more information about the *DeepLeabra What-Where Integration (WWI)* model. The purpose of this information is to give more detailed insight into the model's function beyond the level provided in the main text, but with a model of this complexity, the only way to really understand it is to explore the model itself. It is available for download at: `https://github.com/ccnlab/deep-obj-cat/sims/C++` Furthermore, the best way to understand this model is to understand the framework in which it is implemented, which is explained in great detail, with many running simulations explaining specific elements of functionality, at `http://ccnbook.colorado.edu`

		Units		Pools		
Area	**Name**	**X**	**Y**	**X**	**Y**	**Receiving Projections**
V1	V1s	4	5	8	8	
	V1p	4	5	8	8	V1s V2d V3d V4d TEOd
V1h	V1hs	4	5	16	16	
	V1hp	4	5	16	16	V1s V2d V3d V4d TEOd
Eyes	EyePos	21	21			
	SaccadePlan	11	11			
	Saccade	11	11			
Obj	ObjVel	11	11			
V2	V2s	10	10	8	8	V1s LIPs V3s V4s TEOd V1p V1hp
	V2d	10	10	8	8	**V2s** V1p V1hp LIPd LIPp V3d V4d V3s TEOs
LIP	MtPos	1	1	8	8	V1s
	LIPs	4	4	8	8	MtPos ObjVel SaccadePlan EyePos LIPp
	LIPd	4	4	8	8	**LIPs** LIPp ObjVel Saccade EyePos
	LIPp	1	1	8	8	**MtPos** V1s LIPd
V3	V3s	10	10	4	4	V2s V4s TEOs DPs LIPs V1p V1hp DPp TEOd
	V3d	10	10	4	4	**V3s** V1p V1hp DPp LIPd DPd V4d V4s DPs TEOs
	V3p	10	10	4	4	**V3s** V2d DPd TEOd
DP	DPs	10	10			V2s V3s TEOs V1p V1hp V3p TEOp
	DPd	10	10			**DPs** V1p V1hp DPp TEOd
	DPp	10	10			**DPs** V2d V3d DPd TEOd
V4	V4s	10	10	4	4	V2s TEOs V1p V1hp
	V4d	10	10	4	4	**V4s** V1p V1hp V4p TEOd TEOs
	V4p	10	10	4	4	**V4s** V2d V3d V4d TEOd
TEO	TEOs	10	10	4	4	V4s V1p V1hp TEs
	TEOd	10	10	4	4	**TEOs TEOd** V1p V1hp V4p TEOp TEp TEd
	TEOp	10	10	4	4	**TEOs** V3d V4d TEOd TEd
TE	TEs	10	10	4	4	TEOs V1p V1hp
	TEd	10	10	4	4	**TEs TEd** V1p V1hp V4p TEOp TEp TEOd
	TEp	10	10	4	4	**TEs** V3d V4d TEOd

Table 1: Layer sizes, showing numbers of units in one pool (or entire layer if Pool is missing), and the number of Pools of such units, along X,Y axes. Each area has three associated layers: *s* = superficial layer, *d* = deep layer (context updated by 51B neurons in same area, shown in bold), *p* = pulvinar layer (driven by 5IB neurons from associated area, shown in bold).

Layer Sizes and Structure

Figure 2 in the main text shows the general configuration of the model, and Table 1 shows the specific sizes of each of the layers, and where they receive inputs from.

All the activation and general learning parameters in the model are at their standard Leabra defaults.

Projections

The general principles and patterns of connectivity are shown in Figure 13 (and Figure 1 in the main text). As noted in the main text, the connectivity and overall structure obeys the established principles identified in neocortical anatomy (Rockland & Pandya, 1979; Felleman & Van Essen, 1991; Markov et al., 2014b;

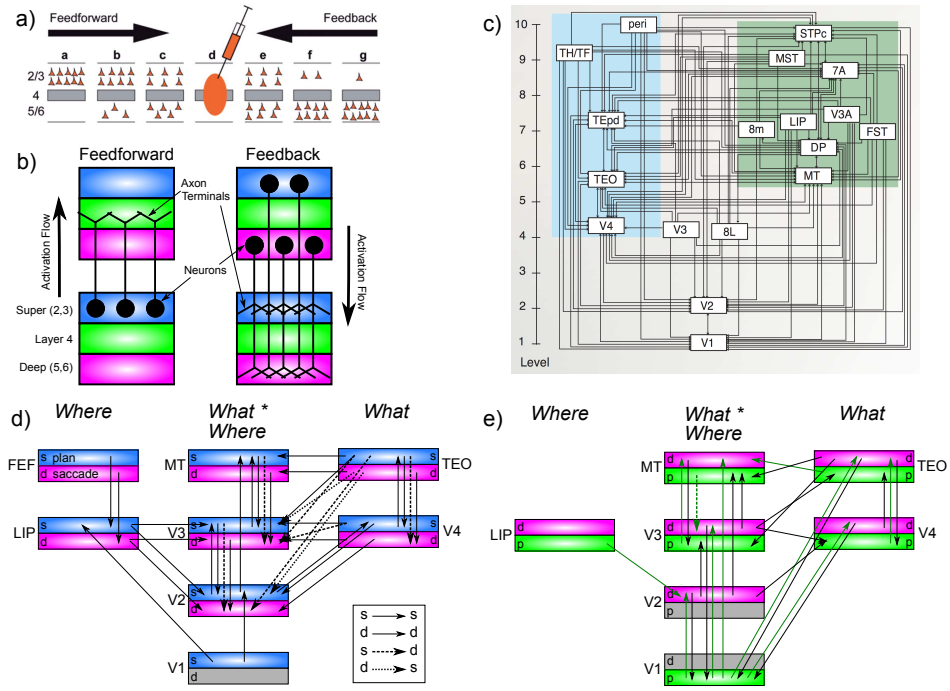

Figure 13: Principles of connectivity in DeepLeabra. **a)** Markov et al (2014) data showing density of *retrograde* labeling from a given injection in a middle-level area (d): most feedforward projections originate from superficial layers of lower areas (a,b,c) and deep layers predominantly contribute to feedback (and more strongly for longer-range feedback). **b)** Summary diagram showing most feedforward connections originating in superficial layers of lower area, and terminating in layer 4 of higher area, while feedback connections can originate in either superficial or deep layers, and in both cases terminate in both superficial and deep layers of the lower area (adapted from Felleman & Van Essen, 1991). **c)** Anatomical hierarchy as determined by percentage of superficial layer source labeling (SLN) by Markov et al (2014) — the hierarchical levels are well matched for our model, but we functionally divide the dorsal pathway (shown in green background) into the two separable components of a *Where* and a *What * Where* integration pathway. **d)** Superficial and deep-layer connectivity in the model. Note the repeating motif between hierarchically-adjacent areas, with bidirectional connectivity between superficial layers, and feedback into deep layers from both higher-level superficial and deep layers, according to canonical pattern shown in panels a and b. Special patterns of connectivity from TEO to V3 and V2, involving crossed super-to-deep and deep-to-super pathways, provide top-down support for predictions based on high-level object representations. **e)** Connectivity for deep layers and pulvinar in the model, which generally mirror the corticocortical pathways (in d). Each pulvinar layer (p) receives 5IB driving inputs from the labeled layer (e.g., V1p receives 5IB drivers from V1). In reality these neurons are more distributed throughout the pulvinar, but it is computationally convenient to organize them together as shown. Deep layers (d) provide predictive input into pulvinar, and pulvinar projections send error signals (via temporal differences between predictions and actual state) to *both* deep and superficial layers of given areas (only d shown). Most areas send deep-layer prediction inputs into the main V1p prediction layer, and receive reciprocal error signals therefrom. The strongest constraint we found was that pulvinar outputs (colored green) must generally project only to higher areas, not to lower areas, with the exceptions of DPp $\rightarrow$ V3 and LIPp $\rightarrow$ V2. V2p was omitted because it is largely redundant with V1p in this simple model.

Markov et al., 2014a).

Detailing each of the specific parameters associated with the different projections shown in Table 1 would take too much space — those interested in this level of detail should download the model from the link shown above. There are topographic projections between many of the lower-level retinotopically-mapped layers, consistent with our earlier vision models (O'Reilly et al., 2013). For example the 8x8 unit groups in V2 are reduced down to the 4x4 groups in V3 via a 4x4 unit-group topographic projection, where neighboring units have half-overlapping receptive fields (i.e., the field moves over 2 unit groups in V2 for every 1 unit group in V3), and the full space is uniformly tiled by using a wrap-around effect at the edges. Similar patterns of connectivity are used in current deep convolutional neural networks. However, we do not share weights across units as in a true convolutional network.

The projections from ObjVel (object velocity) and SaccadePlan layers to LIPs, LIPd were initialized with a topographic sigmoidal pattern that moved as a function of the position of the unit group, by a factor of .5, while the projections from EyePos were initialized with a gaussian pattern. These patterns multiplied uniformly distributed random weights in the .25 to .75 range, with the lowest values in the topographic pattern having a multiplier of .6, while the highest had a multiplier of 1 (i.e., a fairly subtle effect). This produced faster convergence of the LIP layer when doing *Where* pathway pre-training compared to purely random initial weights. In addition to exploring different patterns of overall connectivity, we also explored differences in the relative strengths of receiving projections, which can be set with a `wt_scale.rel` parameter in the simulator. All feedforward pathways have a default strength of 1. For the feedback projections, which are typically weaker (consistent with the biology), we explored a discrete range of strengths, typically .5, .2, .1, and .05. The strongest top-down projections were into V2s from LIP and V3, while most others were .2 or .1. Likewise projections from the pulvinar were weaker, typically .1. These differences in strength sometimes had large effects on performance during the initial bootstrapping of the overall model structure, but in the final model they are typically not very consequential for any individual projection.

Training Parameters

Training typically consisted of 512 alpha trials per epoch (51.2 seconds of real time equivalent), for 1,000 such epochs. Each trial was generated from a virtual reality environment in the emergent simulator, that rendered first-person views with moving eye position onto the object tumbling through space with fixed motion and rotation parameters over the sequence of 8 frames (see Figure 2c in main text for representative example). Because the start of each sequence of 8 frames is unpredictable, we turned off learning for that trial, which improves learning overall. We have recently developed an automatic such mechanism based on the running-average (and running variance) of the prediction error, where we turn off learning whenever the current prediction error z-normalized by these running average values is below 1.5 standard deviations, which works well, and will be incorporated into future models. Biologically, this could correspond to a connection between pulvinar and neuromodulatory areas that could regulate the effective learning rate in this way.

Figure 14a shows the learning trajectory of the model, indicating that it learns quite rapidly. This rapid initial learning is likely facilitated by the extensive use of shortcut connections convering from all over the simulated visual system onto the V1 pulvinar layers, and direct projections back from these pulvinar layers. Thus, error signals are directly communicated and can drive learning quickly and efficiently. However, there are also extensive indirect, bidirectional connections among the superficial layers, which can drive indirect error backpropagation learning as well.

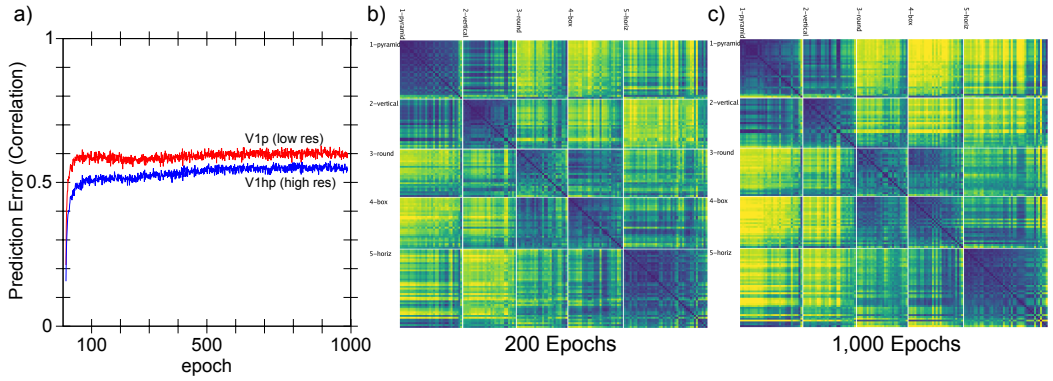

Figure 14: **a)** Predictive learning curve for DeepLeabra, showing the correlation between prediction and actual over the two different V1 layers. Initial learning is quite rapid, followed by a slower but progressive learning process that reflects development of the IT representations (e.g., manipulations that interfere with those areas selectively impair this part of the learning curve). Overall prediction accuracy remains far from perfect, as shown in Figure 2c in main text, and significantly worse than the backpropagtion-based models. This is a typical finding from Leabra models which are significantly more constrained as a result of bidirectional attractor dynamics, Hebbian learning, and inhibitory competition – i.e., the very things that are likely important for forming abstract catgorical representations. **b)** Similarity matrix over TEs layer at 200 epochs, which has less contrast and definition compared to the 1,000 epoch result (**c** also shown in Figure 3a in main text).

Testing Parameters

To be able to monitor similarity metrics as the model trained, we used a running-average integration of neural activity across trials to accumulate the patterns used for the RSA analysis described above. Specifically, for each object, and each frame, the current activation pattern across each layer was recorded and averaged unit-by-unit with a time constant of $\tau = 10$. Critically, by integrating separately for each frame, this running-average computation did not introduce any bias for temporally-adjacent frames to be more similar. Nevertheless, when we computed the frame-to-frame similarities for TE, they were quite high (.901 correlation on average across all objects).

Model Algorithms

The biologically-based model was implemented using the Leabra framework, which is described in detail in previous publications (O'Reilly, Munakata, Frank, Hazy, & Contributors, 2012; O'Reilly & Munakata, 2000; O'Reilly, 1998, 1996), and summarized here. There are two main implementations of Leabra, one in the *C++ emergent* software, and a new one using Go and Python language at: `https://github.com/emer/leabra`. There are also other simpler implementations in Python and MATLAB, see `https://grey.colorado.edu/emergent/index.php/Leabra`. Both of the preceeding links contain a full detailed description of the algorithm. These same equations and standard parameters have been used to simulate over 40 different models in (O'Reilly et al., 2012; O'Reilly & Munakata, 2000), and a number of other research models. Thus, the model can be viewed as an instantiation of a systematic modeling framework using standardized mechanisms, instead of constructing new mechanisms for each model. Here, we only detail properties of the predictive learning algorithm that go beyond the basic Leabra model.

Deep Context

At the end of every plus phase, a new deep-layer context net input is computed from the dot product of the context weights times the sending activations, just as in the standard net input:

$$\eta = \langle x_i w_{ij} \rangle = \frac{1}{n} \sum_i x_i w_{ij} \tag{2}$$

This net input is then added in with the standard net input at each cycle of processing.

The relative strength of these context layer inputs was set progressively larger for higher layers in the network, with a maximum of 4 in V4, TEO, and TE. In addition, TEO and TE received *self* context projections which provide an extended window of temporal context into the prior 200 msec interval, consistent with multiple sources of neural data (Chaudhuri et al., 2015). These self projections were connected only within the narrower Pool level of units, enabling these neurons to develop mutually-excitatory loops to sustain activations over the multiple trials when the same object was present. We hypothesize that these modifications correspond to biological adaptations in IT cortex that likewise support greater sustained activation of object-level representations.

Learning of the context weights occurs as normal, but using the sending activation states from the *prior* time step's activation.

Computational and Biological Details of SRN-like Functionality

Predictive auto-encoder learning has been explored in various frameworks, but the most relevant to our model comes from the application of the SRN to a range of predictive learning domains (Elman, 1990; Elman et al., 1996). One of the most powerful features of the SRN is that it enables error-driven learning, instead of arbitrary parameter settings, to determine how prior information is integrated with new information. Thus, SRNs can learn to hold onto some important information for a relatively long interval, while rapidly updating other information that is only relevant for a shorter duration. This same flexibility is present in our DeepLeabra model. Furthermore, because this temporal context information is hypothesized to be present in the deep layers throughout the entire neocortex (in every microcolumn of tissue), the DeepLeabra model provides a more pervasive and interconnected form of temporal integration compared to the SRN, which typically just has a single temporal context layer associated with the internal "hidden" layer of processing units.

An extensive computational analysis of what makes the SRN work as well as it does, and explorations of a range of possible alternative frameworks, has led us to an important general principle: *subsequent outcomes determine what is relevant from the past*. At some level, this may seem obvious, but it has significant implications for predictive learning mechanisms based on temporal context. It means that the information encoded in a temporal context representation cannot be learned at the time when that information is presently active. Instead, the relevant contextual information is learned on the basis of what happens next. This explains the peculiar power of the otherwise strange property of the SRN: the temporal context information is preserved as a *direct copy* of the state of the hidden layer units on the previous time step (Figure 15), and then learned synaptic weights integrate that copied context information into the next hidden state (which is then copied to the context again, and so on). This enables the error-driven learning taking place in the *current* time step to determine how context information from the *previous* time step is integrated. And the simple direct copy operation eschews any attempt to shape this temporal context itself, instead relying on the learning pressure that shapes the hidden layer representations to also shape the context representations. In other words, this copy operation is essential, because there is no other viable source of learning signals to shape the nature of the context representation itself (because these learning signals require future outcomes, which are by definition only available later).

The direct copy operation of the SRN is however seemingly problematic from a biological perspective:

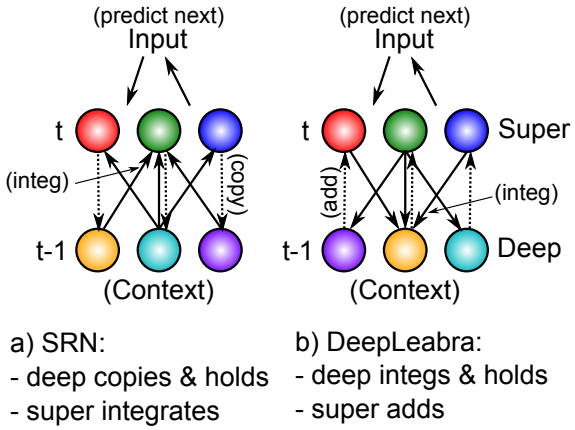

Figure 15: How the DeepLeabra temporal context computation compares to the SRN mathematically. **a)** In a standard SRN, the context (deep layer biologically) is a copy of the hidden activations from the prior time step, and these are held constant while the hidden layer (superficial) units integrate the context through learned synaptic weights. **b)** In DeepLeabra, the deep layer performs the weighted integration of the soon-to-be context information from the superficial layer, and then holds this integrated value, and feeds it back as an additive net-input like signal to the superficial layer. The context net input is pre-computed, instead of having to compute this same value over and over again. This is more efficient, and more compatible with the diffuse interconnections among the deep layer neurons. Layer 6 projections to the thalamus and back recirculate this pre-computed net input value into the superficial layers (via layer 4), and back into itself to support maintenance of the held value.

how could neurons copy activations from another set of neurons at some discrete point in time, and then hold onto those copied values for a duration of 100 msec, which is a reasonably long period of time in neural terms (e.g., a rapidly firing cortical neuron fires at around 100 Hz, meaning that it will fire 10 times within that context frame). However, there is an important transformation of the SRN context computation, which is more biologically plausible, and compatible with the structure of the deep network (Figure 15). Specifically, instead of copying an entire set of activation states, the context activations (generated by the phasic 5IB burst) are immediately sent through the adaptive synaptic weights that integrate this information, which we think occurs in the 6CC (corticortical) and other lateral integrative connections from 5IB neurons into the rest of the deep network. The result is a *pre-computed net input* from the context onto a given hidden unit (in the original SRN terminology), not the raw context information itself. Computationally, and metabolically, this is a much more efficient mechanism, because the context is, by definition, unchanging over the 100 msec alpha cycle, and thus it makes more sense to pre-compute the synaptic integration, rather than repeatedly re-computing this same synaptic integration over and over again (in the original feedforward backpropagation-based SRN model, this issue did not arise because a single step of activation updating took place for each context update — whereas in our bidirectional model many activation update steps must take place per context update).

There are a couple of remaining challenges for this transformation of the SRN. First, the pre-computed net input from the context must somehow persist over the subsequent 100 msec period of the alpha cycle. We hypothesize that this can occur via NMDA and mGluR channels that can easily produce sustained excitatory currents over this time frame. Furthermore, the reciprocal excitatory connectivity from 6CT to TRC and back to 6CT could help to sustain the initial temporal context signal. Second, these contextual integration synapses require a different form of learning algorithm that uses the sending activation from the prior 100 msec, which is well within the time constants in the relevant calcium and second messenger pathways

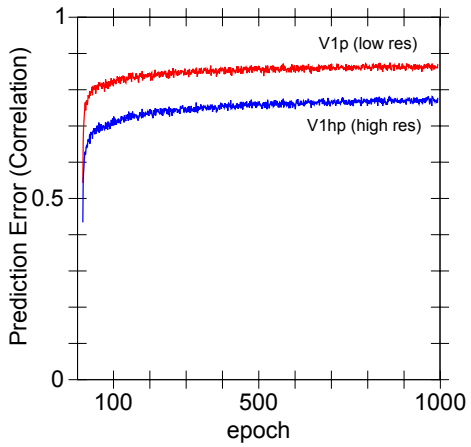

Figure 16: Learning curves for the backpropagation version of the WWI model. Although it achieves better predictive accuracy than the DeepLeabra version, it fails to acquire abstract object category structure, indicating a potential tradeoff between simplifying and categorizing inputs, versus predicting precisely where the low-level visual features will move.

involved in synaptic plasticity.

Backpropagation Model Methods

The backpropagation version of the WWI model has exactly the same layer sizes and *feedforward* patterns of connectivity as the DeepLeabra version. Topographically, the V1p and V1hp pulvinar layers serve as output layers at the highest level of the network, receiving all the various connections from deep layers as shown in Table 1. Likewise, the LIPp served as a target output layer for the Where pathway. To achieve predictive learning, the V1 pulvinar targets were from the scene at time t, while the V1s inputs were from the scene at time $t - 1$. We also ran a comparison auto-encoder model that had inputs and target outputs from the same time step, and it showed even less systematic organziation of its higher-level representations, further supporting the notion that predictive learning is important, across all frameworks. The learning curve for the predictive version is shown in Figure 16, which shows better overall prediction accuracy compared to the DeepLeabra model. However, as the RSA showed, this backpropagation model failed to learn object categories that go beyond the input similarity structure, indicating that perhaps it was paying too much "attention" in learning to this low-level structure, and lacked the necessary mechanisms to enable it to impose a simplifying higher-level structure on top of these inputs.

PredNet Model Methods

The PredNet architecture was designed to incorporate principles from predictive coding theory into a neural network model for predicting the next frame in a video sequence. Details of the model can be found in the original paper (Lotter et al., 2016), but here we provide a brief overview of the architecture.

Architecture

PredNet is a deep convolutional neural network that is composed of layers containing discrete modules. The lowest layer generates a prediction of incoming inputs (i.e. the pixels in the next frame), while each

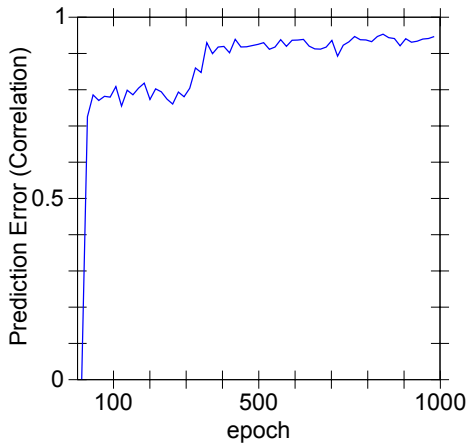

Figure 17: Learning curves for the PredNet model. This model achieves the best overall prediction performance but also has the least well differentiated, categorical representations.

of the higher layers attempts to predict the *errors* made by the previous layer. Each layer contains an input convolutional module (A_l), a recurrent representational module (R_l), a prediction module ($\hat{A}_l$), and a representation of its own errors (E_l). The input convolutional module (A_l) transforms its input with a set of standard convolutional filters, a rectified linear activation function, and a max-pooling operation. The recurrent representation module (R_l) is a convolutional LSTM, which is a recurrent convolutional network that replaces the matrix multiplications in the standard LSTM equations with convolutions, allowing it to maintain a spatially organized representation of its inputs over time. The prediction module ($\hat{A}_l$) consists of another standard convolutional layer and rectified linear activation that is used to generate predictions from the output of R_l. These predictions are then compared against the output of the input convolutional module (A_l). The errors generated in this comparison are represented explicitly in E_l, which applies a rectified linear activation to a concatenation of the positive ($A_l - \hat{A}_l$) and negative ($\hat{A}_l - A_l$) prediction errors. These errors then become the inputs to the next layer.

$$A_l^t = \begin{cases} x_t, & \text{if } l = 0 \\ MaxPool(ReLU(Conv(E_{l-1}^t))), & \text{if } l > 0 \end{cases} \tag{3}$$

$$\hat{A}_l^t = ReLU(Conv(R_l^t)) \tag{4}$$

$$E_l^t = [ReLU(A_l^t - \hat{A}_l^t); ReLU(\hat{A}_l^t - A_l^t)] \tag{5}$$

$$R_l^t = ConvLSTM(E_l^{t-1}, R_l^{t-1}, UpSample(R_{l+1}^t)) \tag{6}$$

At each time step in the video sequence, PredNet generates a prediction of the next frame. This is done as follows: first, the R_l is computed for each layer starting from the top of the hierarchy (because each R_l^t depends on input from R_{l+1}^t), and then the A_l^t, $\hat{A}_l^t$ and E_l^t are computed in a feed-forward fashion (becauase each A_l^t depends on input from the layer below, E_{l-1}^t).

All analyses in the RSA were conducted using the representations from the R_l layers.

Implementation details

All experiments with the PredNet architecture were performed using PyTorch. An informal hyperparameter search was conducted to find the settings that maximized representational similarity to the human

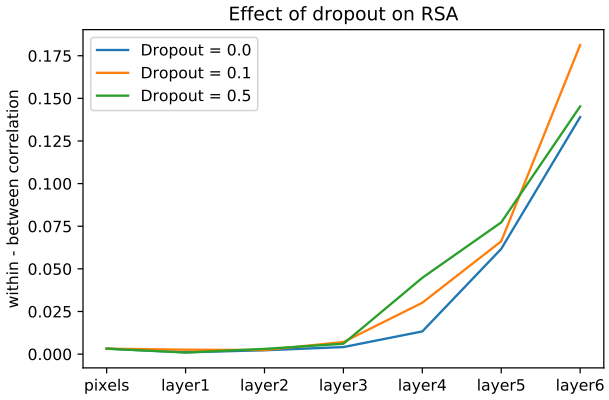

Figure 18: Effect of dropout in PredNet on RSA, as measured by the difference between the average within-category correlation and the average between category correlation (using the Centroid categories derived from human data). Dropout marginally improves the category structure learned in PredNet.

judgments. This was done by conducting RSA on each layer for each hyperparameter setting, and computing, according to the Centroid categories derived from the human data, the difference between the average within-category similarity and the average between-category similarity. Our final architecture had 6 layers with 3, 16, 32, 64, 128, and 256 filters in the A_l and R_l modules, and 3x3 kernels throughout the whole network. We also found that using sigmoid and tanh activation functions in fully-connected convolutional LSTMs slightly improved performance, so these were used for all experiments.

The weights in the PredNet model are trained using error backpropagation. Predictions are generated and errors are computed at all levels of the hierarchy, but the model performs better when only the lowest layer's errors are backpropagated (Lotter et al., 2016). We confirmed these results with experiments that backpropagated the errors in higher layers, in which performance (in terms of mean squared error) was marginally reduced but the RSA results were similar. For this reason, all reported experiments used a PredNet that was trained by only backpropagating the lowest level error.

The model was trained using a batch size of 8 and an Adam optimizer with a learning rate of 0.0001, with no scheduler, for 150,000 batches. A training curve is shown in Figure 17, showing that it achieves the best overall prediction accuracy of any model we tested, and yet does not have representations that are as differentiated or categorical as our biologically based model, as shown in the main paper.

Regularization experiments

As discussed in the main paper, our biologically based model includes a number of important biologically motivated properties that may be contributing to the development of its categorical representations. These properties, including excitatory bidirectional connections, inhibitory competition, and an additional form of Hebbian learning, may be acting as regularizers that encourage categorical learning. We therefore tested whether standard regularization methods used in deep learning would have similar effects on the representations developed in the PredNet architecture. We tested 1) batch normalization, 2) dropout (0.1, 0.3, and 0.5), and 3) weight decay (0.01,0.001,0.0001,0.00001). All experiments with batch normalization and weight decay showed reduced performance (in terms of both prediction error on the test set and within-category correlation). As shown in figure 18, dropout marginally improved the within-category correlation while also slightly improving prediction accuracy, so a dropout rate of 0.1 was used for the comparison to our biologically based model in the main paper.

References

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

Arcaro, M. J., Pinsk, M. A., & Kastner, S. (2015). The Anatomical and Functional Organization of the Human Visual Pulvinar. *Journal of Neuroscience*, *35*(27), 9848–9871.

Barczak, A., O'Connell, M. N., McGinnis, T., Ross, D., Mowery, T., Falchier, A., & Lakatos, P. (2018). Top-down, contextual entrainment of neuronal oscillations in the auditory thalamocortical circuit. *Proceedings of the National Academy of Sciences*, *115*(32), E7605–E7614.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*(4), 695–711.

Bastos, A. M., Vezoli, J., Bosman, C. A., Schoffelen, J.-M., Oostenveld, R., Dowdall, J. R., De Weerd, P., Kennedy, H., & Fries, P. (2015). Visual Areas Exert Feedforward and Feedback Influences through Distinct Frequency Channels. *Neuron*, *85*(2), 390–401.

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.

Bienenstock, E. L., Cooper, L. N., & Munro, P. W. (1982). Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *The Journal of Neuroscience*, *2*(2), 32–48.

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.

Bortone, D. S., Olsen, S. R., & Scanziani, M. (2014). Translaminar inhibitory cells recruited by layer 6 corticothalamic neurons suppress visual cortex. *Neuron*, *82*.

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.

Chaudhuri, R., Knoblauch, K., Gariel, M.-A., Kennedy, H., & Wang, X.-J. (2015). A Large-Scale Circuit Mechanism for Hierarchical Dynamical Processing in the Primate Cortex. *Neuron*, *88*(2), 419–431.

Colby, C. L., Duhamel, J. R., & Goldberg, M. E. (1997). Visual, presaccadic, and cognitive activation of single neurons in monkey lateral intraparietal area. *Journal of neurophysiology*, *76*, 2841.

Connors, B. W., Gutnick, M. J., & Prince, D. A. (1982). Electrophysiological properties of neocortical neurons in vitro. *Journal of Neurophysiology*, *48*(6), 1302–1320.

Crick, F. (1984). Function of the thalamic reticular complex: The searchlight hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, *81*, 4586–4590.

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

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*(1), 193–222.

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.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed Hierarchical Processing in the Primate Cerebral Cortex. *Cerebral Cortex*, *1*(1), 1–47.

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

Fiebelkorn, I. C., Pinsk, M. A., & Kastner, S. (2018). A dynamic interplay within the frontoparietal network underlies rhythmic spatial attention. *Neuron*, *99*(4), 842–853.e8.

Fiser, A., Mahringer, D., Oyibo, H. K., Petersen, A. V., Leinweber, M., & Keller, G. B. (2016). Experience-dependent spatial expectations in mouse visual cortex. *Nature Neuroscience*, *19*(12), 1658–1664.

Foldiak, P. (1991). Learning Invariance from Transformation Sequences. *Neural Computation*, *3*(2), 194–200.

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.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

Gavornik, J. P., & Bear, M. F. (2014). Learned spatiotemporal sequence recognition and prediction in primary visual cortex. *Nature Neuroscience*, *17*(5), 732–737.

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

Gottlieb, J. P., Kusunoki, M., & Goldberg, M. E. (1998). The representation of visual salience in monkey parietal cortex. *Nature*, *391*, 481.

Gruber, W. R., Klimesch, W., Sauseng, P., & Doppelmayr, M. (2005). Alpha Phase Synchronization Predicts P1 and N1 Latency and Amplitude Size. *Cerebral Cortex*, *15*(4), 371–377.

Halassa, M. M., & Kastner, S. (2017). Thalamic functions in distributed cognitive control. *Nature Neuroscience*, *20*(12), 1669.

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.

Jaramillo, J., Mejias, J. F., & Wang, X.-J. (2019). Engagement of Pulvino-cortical Feedforward and Feedback Pathways in Cognitive Computations. *Neuron*, *101*(2), 321–336.e9.

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

Jensen, O., & Mazaheri, A. (2010). Shaping functional architecture by oscillatory alpha activity: Gating by inhibition. *Frontiers in Human Neuroscience*, *4*(186).

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.

Keller, G. B., & Mrsic-Flogel, T. D. (2018). Predictive Processing: A Canonical Cortical Computation. *Neuron*, *100*(2), 424–435.

Klimesch, W. (2011). Evoked alpha and early access to the knowledge system: The P1 inhibition timing hypothesis. *Brain Research*, *1408*, 52–71.

Klimesch, W., Sauseng, P., & Hanslmayr, S. (2007). EEG alpha oscillations: The inhibition-timing hypothesis. *Brain Research Reviews*, *53*(1), 63–88.

Kobatake, E., & Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway. *Journal of Neurophysiology*, *71*(3), 856–867.

Kogo, N., & Trengove, C. (2015). Is predictive coding theory articulated enough to be testable? *Frontiers in Computational Neuroscience*, *9*.

Kok, P., & de Lange, F. P. (2015). Predictive Coding in Sensory Cortex. In *An Introduction to Model-Based Cognitive Neuroscience* (pp. 221–244). Springer, New York, NY.

Kok, P., Jehee, J. F. M., & de Lange, F. P. (2012). Less Is More: Expectation Sharpens Representations in the Primary Visual Cortex. *Neuron*, *75*(2), 265–270.

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.

LaBerge, D., & Buchsbaum, M. S. (1990). Positron emission tomographic measurements of pulvinar activity during an attention task. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *10*, 613–9.

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.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America*, *20*(7), 1434–1448.

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., & Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience*, *21*(6), 335–346.

Lim, S., McKee, J. L., Woloszyn, L., Amit, Y., Freedman, D. J., Sheinberg, D. L., & Brunel, N. (2015). Inferring learning rules from distributions of firing rates in cortical neurons. *Nature Neuroscience*, *18*(12), 1804–1810.

Lopes da Silva, F. (1991). Neural mechanisms underlying brain waves: From neural membranes to networks. *Electroencephalography and Clinical Neurophysiology*, *79*(2), 81–93.

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]*.

Luczak, A., Bartho, P., & Harris, K. D. (2013). Gating of sensory input by spontaneous cortical activity. *The Journal of Neuroscience*, *33*(4), 1684–1695.

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.

Maier, A., Adams, G. K., Aura, C., & Leopold, D. A. (2010). Distinct Superficial and Deep Laminar Domains of Activity in the Visual Cortex during Rest and Stimulation. *Frontiers in Systems Neuroscience*, *4*(31).

Maier, A., Aura, C. J., & Leopold, D. A. (2011). Infragranular sources of sustained local field potential responses in macaque primary visual cortex. *The Journal of Neuroscience*, *31*(6), 1971–1980.

Makeig, S., Westerfield, M., Jung, T. P., Enghoff, S., Townsend, J., Courchesne, E., & Sejnowski, T. J. (2002). Dynamic Brain Sources of Visual Evoked Responses. *Science*, *295*, 690–693.

Marino, A. C., & Mazer, J. A. (2016). Perisaccadic Updating of Visual Representations and Attentional States: Linking Behavior and Neurophysiology. *Frontiers in Systems Neuroscience*, *10*.

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. (2014a). A Weighted and Directed Interareal Connectivity Matrix for Macaque Cerebral Cortex. *Cerebral Cortex*, *24*(1), 17–36.

Markov, N. T., Vezoli, J., Chameau, P., Falchier, A., Quilodran, R., Huissoud, C., Lamy, C., Misery, P., Giroud, P., Ullman, S., Barone, P., Dehay, C., Knoblauch, K., & Kennedy, H. (2014b). Anatomy of hierarchy: Feedforward and feedback pathways in macaque visual cortex: Cortical counterstreams. *Journal of Comparative Neurology*, *522*(1), 225–259.

Mayer, A., Schwiedrzik, C. M., Wibral, M., Singer, W., & Melloni, L. (2016). Expecting to See a Letter: Alpha Oscillations as Carriers of Top-Down Sensory Predictions. *Cerebral Cortex*, *26*(7), 3146–3160.

Meyer, T., & Olson, C. R. (2011). Statistical learning of visual transitions in monkey inferotemporal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(48), 19401–19406.

Michalareas, G., Vezoli, J., van Pelt, S., Schoffelen, J.-M., Kennedy, H., & Fries, P. (2016). Alpha-Beta and Gamma Rhythms Subserve Feedback and Feedforward Influences among Human Visual Cortical Areas. *Neuron*, *89*(2), 384–397.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202.

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.

Nakamura, K., & Colby, C. L. (2002). Updating of the visual representation in monkey striate and extrastriate cortex during saccades. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(6), 4026–4031.

Neupane, S., Guitton, D., & Pack, C. C. (2016). Two distinct types of remapping in primate cortical area V4. *Nature Communications*, *7*, 10402.

Nunn, C. M. H., & Osselton, J. W. (1974). The Influence of the EEG Alpha Rhythm on the Perception of Visual Stimuli. *Psychophysiology*, *11*(3), 294–303.

O'Herron, P., & von der Heydt, R. (2013). Remapping of border ownership in the visual cortex. *Journal of Neuroscience*, *33*(5), 1964–1974.

Olsen, S., Bortone, D., Adesnik, H., & Scanziani, M. (2012). Gain control by layer six in cortical circuits of vision. *Nature*, *483*(7387), 47–52.

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., Munakata, Y., Frank, M. J., Hazy, T. E., & Contributors (2012). *Computational Cognitive Neuroscience*. Wiki Book, 1st Edition, URL: http://ccnbook.colorado.edu.

O'Reilly, R. C., Petrov, A. A., Cohen, J. D., Lebiere, C. J., Herd, S. A., & Kriete, T. (2014). 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. R., & Rohrlich, J. (2017). Deep predictive learning: A comprehensive model of three visual streams. *arXiv:1709.04654 [q-bio]*.

Ouden, H. E. M., Kok, P., & Lange, F. P. (2012). How prediction errors shape perception, attention, and motivation. *Frontiers in Psychology*, *3*(548).

Pennartz, C. M., Dora, S., Muckli, L., & Lorteije, J. A. (2019). Towards a Unified View on Pathways and Functions of Neural Recurrent Processing. *Trends in Neurosciences*.

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*.

Petrof, I., Viaene, A. N., & Sherman, S. M. (2012). Two populations of corticothalamic and interareal corticocortical cells in the subgranular layers of the mouse primary sensory cortices. *Journal of Comparative Neurology*, *520*(8), 1678–1686.

Pinault, D. (2004). The thalamic reticular nucleus: Structure, function and concept. *Brain research*, *46*.

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

Purushothaman, G., Marion, R., Li, K., & Casagrande, V. A. (2012). Gating and control of primary visual cortex by pulvinar. *Nature Neuroscience*, *15*(6), 905–912.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *bioRxiv*, 240614.

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.

Reynolds, J. H., Chelazzi, L., & Desimone, R. (1999). Competitive mechanisms subserve attention in macaque areas V2 and V4. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *19*, 1736–1753.

Richter, D., & de Lange, F. P. (2019). Statistical learning attenuates visual activity only for attended stimuli. *eLife*, *8*, e47869.

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.

Rockland, K. S., & Pandya, D. N. (1979). Laminar origins and terminations of cortical connections of the occipital lobe in the rhesus monkey. *Brain Research*, *179*(1), 3–20.

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.

Saalmann, Y. B., & Kastner, S. (2011). Cognitive and perceptual functions of the visual thalamus. *Neuron*, *71*(2), 209–223.

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.

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.

Sherman, S. M. (2014). The function of metabotropic glutamate receptors in thalamus and cortex. *The Neuroscientist*, *20*(2), 146–149.

Sherman, S. M., & Guillery, R. W. (2011). Distinct functions for direct and transthalamic corticocortical connections. *Journal of Neurophysiology*, *106*(3), 1068–1077.

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.

Snow, J. C., Allen, H. A., Rafal, R. D., & Humphreys, G. W. (2009). Impaired attentional selection following lesions to human pulvinar: Evidence for homology between human and monkey. *Proceedings of the National Academy of Sciences*, *106*(10), 4054–4059.

Spaak, E., Bonnefond, M., Maier, A., Leopold, D. A., & Jensen, O. (2012). Layer-specific entrainment of gamma-band neural activity by the alpha rhythm in monkey visual cortex. *Current Biology*, *22*(24), 2313–2318.

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

Spratling, M. W. (2008). Reconciling predictive coding and biased competition models of cortical function. *Frontiers in Computational Neuroscience*, *2*(4), 1–8 (online).

Summerfield, C., & Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in Cognitive Sciences*, *13*(9), 403–409.

Summerfield, C., Trittschuh, E. H., Monti, J. M., Mesulam, M. M., & Egner, T. (2008). Neural repetition suppression reflects fulfilled perceptual expectations. *Nature Neuroscience*, *11*(9), 1004–1006.

Thomson, A. M. (2010). Neocortical layer 6, a review. *Frontiers in Neuroanatomy*, *4*(13).

Thomson, A. M., & Lamy, C. (2007). Functional maps of neocortical local circuitry. *Frontiers in Neuroscience*, *1*(1), 19–42.

Todorovic, A., van Ede, F., Maris, E., & de Lange, F. P. (2011). Prior Expectation Mediates Neural Adaptation to Repeated Sounds in the Auditory Cortex: An MEG Study. *Journal of Neuroscience*, *31*(25), 9118–9123.

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.

Urakubo, H., Honda, M., Froemke, R. C., & Kuroda, S. (2008). Requirement of an allosteric kinetics of NMDA receptors for spike timing-dependent plasticity. *The Journal of Neuroscience*, *28*(13), 3310–3323.

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 Kerkoerle, T., Self, M. W., Dagnino, B., Gariel-Mathis, M.-A., Poort, J., van der Togt, C., & Roelfsema, P. R. (2014). Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. *Proceedings of the National Academy of Sciences U.S.A.*, *111*(40), 14332–14341.

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

Varela, F. J., Toro, A., John, E. R., & Schwartz, E. L. (1981). Perceptual framing and cortical alpha rhythm. *Neuropsychologia*, *19*(5), 675–686.

Vinken, K., & Vogels, R. (2017). Adaptation can explain evidence for encoding of probabilistic information in macaque inferior temporal cortex. *Current Biology*, *27*(22), R1210–R1212.

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

von Stein, A., Chiang, C., & König, P. (2000). Top-down processing mediated by interareal synchronization. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(26), 14748–14753.

Walsh, K. S., McGovern, D. P., Clark, A., & O'Connell, R. G. (2020). Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Annals of the New York Academy of Sciences*, *1464*(1), 242–268.

Whittington, J. C. R., & Bogacz, R. (2019). Theories of error back-propagation in the brain. *Trends in Cognitive Sciences*, *23*(3), 235–250.

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.

Wilson, J. R., Bose, N., Sherman, S. M., & Guillery, R. W. (1984). Fine structural morphology of identified X- and Y-cells in the cat's lateral geniculate nucleus. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *221*(1225), 411–436.

Wimmer, R. D., Schmitt, L. I., Davidson, T. J., Nakajima, M., Deisseroth, K., & Halassa, M. M. (2015). Thalamic control of sensory selection in divided attention. *Nature*, *526*(7575), 705–709.

Wiskott, L., & Sejnowski, T. J. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, *14*, 715–770.

Wurtz, R. H. (2008). Neuronal mechanisms of visual stability. *Vision Research*, *48*(20), 2070–2089.

Xing, D., Yeh, C.-I., Burns, S., & Shapley, R. M. (2012). Laminar analysis of visually evoked activity in the primary visual cortex. *Proceedings of the National Academy of Sciences*, *109*(34), 13871–13876.

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

Zhou, H., Schafer, R. J., & Desimone, R. (2016). Pulvinar-cortex interactions in vision and attention. *Neuron*, *89*, 209–220.