Dear Dr. Postle:

We are hereby submitting our manuscript, “Deep Predictive Learning in the Neocortex and Pulvinar” for publication in the *Journal of Cognitive Neuroscience*. We had submitted an earlier version to *eLife* and are trying out the Neuroscience Peer Review Consortium so that you can hopefully get those reviews, for submission number: 26-06-2020-RA-eLife-60436. We have extensively revised the manuscript to address the reviewer’s concerns, and thus increased its length beyond the parameters for that journal, by providing more details about our computational model, including an initial presentation of a much simpler, smaller-scale model that should facilitate in the understanding of the basic mechanisms, as requested by the reviewers.

All three of the prior reviewers appeared to be from a computationally-oriented background, and had fairly consistent comments. If you decide to seek additional reviews, it would be great if you could get someone from the systems neuroscience perspective, e.g., someone like Sabine Kastner at Princeton, or Kenneth D. Harris at UCL.

In the remainder of this letter, we summarize the changes made in this revision relative to the comments from the eLife reviews, followed by a detailed discussion of the reviewer comments.

First, the most critical problem was that the basic ideas behind our theory were not communicated effectively, and the reviewers could not understand it. We made major revisions to address this. We spent a lot of effort redoing the critical Figure 2, which provides a summary of the theory. Trying to capture a dynamic that unfolds across time and over layers is tricky, and the previous version of this figure did a relatively poor job, in retrospect, consistent with the reviewer’s comments. We think the new version is a lot clearer, and hopefully goes a long way toward communicating the essential ideas better. The associated exposition of the model was also significantly clarified, and in particular the section “Computational Properties of Predictive Learning in the Thalamocortical Circuits” was completely rewritten and significantly improved. As requested by the reviewers, we moved the more detailed discussion of the neuroscience and predictions to after the presentation of the computational models, and also improved that significantly, also including a better figure there.

The other major issues centered around the nature of the 3D object dataset, and the logic for the RSA-style analysis. We significantly clarified and reorganized this presentation, emphasizing the initial presentation of the objective, easily-interpretable differentiation of representations across the hierarchy, followed by a presentation of the intuitively-appealing RDM plots with more information about the corresponding 3D objects in each category, which we think clarify the nature of the learned categories significantly. Reviewer 3 provided the most detailed and constructive feedback on these issues, so our responses are mostly detailed there. Thank you for your consideration,



Randall C. O’Reilly, on behalf of the co-authors.

### Reviewer 1:

My biggest concern with this work is that it was very difficult to evaluate. The introduction is written like a review paper rather than a research manuscript. Also, the introduction is fully focused on reviewing different aspects of pulvinar biology (which is great), but then the result section seems to be quite disconnected with no clear focus on how the pulvinar is wired up to the cortical model and how exactly the weights or selectivity of its connections change with learning.

In general, I'm not seeing how the introduction and the three elements above play out in the results. I think I was initially quite excited about the premise of the paper, but could not find enough data within the manuscript to support the ideas laid out in the introduction. Also, as a general comment, I found the figures difficult to follow; they could be made more informative.

We think that the improved figures, added simulation, and improved discussion should address these issues.

Specific comments:

1. There has been a surge of "biologically plausible approximations of backpropagation" recently and several of them show that the learned representations match performance (ML/neuro) of standard backprop. But often these algorithms tend to not scale and break down e.g. at the scale of ImageNet. You could argue that we care about the brain and not ImageNet, but I always found that ML benchmark to be a good judge of how well something can scale up to a challenging task.

Indeed we would like to argue that we care about the brain and not ImageNet :) In particular, almost all of the principles that underlie current DCNN models were invented back in the 1980’s and 90’s, and the main difference now is largely due to advances in “engineering” that has enabled very large, deep models to run fast on massive datasets (GPU’s and the like). Thus, from a purely scientific perspective, the ImageNet results have not necessarily changed our understanding of the essential nature of learning in the brain.

Furthermore, there are significant tradeoffs between including more biologically-based properties vs. pure computational speed. For example, we show that our models learn abstract representations in ways that other DCNN-based models do not, and that this may depend on the presence of full simultaneous bidirectional activation flow, which allows top-down influences to shape the overall attractor dynamics of the network. The dynamic nature of this network, which captures both the bidirectional connectivity and widely-demonstrated top-down influences in the brain, means that it takes well over 100 times longer to train than a corresponding feedforward DCNN. Just the bidirectional connectivity alone imposes a huge factor of 2 in memory and processing for all the connections -- this is actually much larger in practice because we find that shortcut connections in both directions are useful, as also widely found in the brain. Our models take over a day to run across 32 CPUs on our compute cluster, and they are already pushing the maximum memory throughput bottlenecks, so scaling them up significantly is currently not feasible. They also do not work well on GPUs at this point due to critical optimizations for sparse activity -- this is an ongoing area of work in the lab.

2. The RDM category similarity used in figs 4 and 5 are noisy and easy to score well on. The authors really need to use neural predictivity and similarity similar to that used by labs like Jim DiCarlo's (e.g. Brainscore). I suspect that they can't do this now is because their model gets video input, not static image input, and monkey physiologists haven't curated large-scale video response datasets. However, without gaining access to such datasets, it is almost impossible to validate their model.

Yes our model does require video input. The responses to Reviewer 3 address these similar concerns.

3. Violation of expectations. If their model is good, it should be able to reproduce violation of expectations results from psychology. There are lots of open-source datasets and benchmarks for that (from both machine learning people and cognitive scientists).

Indeed, and there are lots of other such data that could be accounted for by our model. We included Figure 12 because it was the most iconic, important predictive learning result that connects directly to the neural areas simulated in the model. As might be expected, we are planning many other applications of the model to explain many other findings, including behavioral experiments that we have designed to specifically test this model -- these experiments include violation of expectations conditions, and our model certainly predicts such effects, as predicted...

4. The authors need to run the model on more than one dataset. They only ran the model on one dataset --- a dataset of rotating 3D objects from their lab (that almost nobody else in the literature uses). Thus I'm worried that their model is overfit to that dataset. They could easily run it on more standard video datasets, and I have no idea why they don't. Running on other datasets is a way to cross-validate the model architecture, which is essential to show that the model is robust, particularly for a complicated model like this one with so many hand-crafted hyperparameters.

These issues are addressed in response to Reviewer 3. In addition we have run the model on numerous other datasets, including a project applying the model to prediction in speech, which is nearly ready for writing up. Standard video datasets do not typically contain a suitably sampled set of object shapes that would enable answering the central question regarding the process of developing abstract invariant shape representations across the hierarchy in the model. The ones we are familiar with focus on driving and rotating faces.

Reviewer #2:

The authors propose a computational theory that pulvinar is central to predictive error-driven learning (learning by comparing predictions with outcomes) of sensory representation. In the authors' model, pulvinar integrates both top-down future predictions from cortical L6 neurons, and actual sensory outcome from L5 bursting neurons. The authors showed that a computational model of the visual cortex trained with predictive learning shows more object-selective representation in higher layers compared to lower layers, unlike structurally-analogous models trained with back-propagation. The authors study the important question of the neural mechanism of representation learning in sensory cortices, and propose a novel theory based on the combination of predictive learning and higher-order thalamus. However, I cannot support publication of the manuscript in its current form due to the following major concerns (especially the first two).

Major:

(1) This study lacks critical details for understanding the inner working of the model.

The authors go through great length of motivate and explain their model conceptually. However, very little quantitative details are provided to aid the readers' understanding and reproducing of results. The following are several indications for this lack of details.

The authors acknowledged that their model is very complex ("the only way to really understand the model is to explore the model itself"). They pointed the readers to a github repo (p. 28) for more insights, however, the linked repo doesn't exist (publicly). This alone would be OK if the authors plan to open-source the code after publication and sufficient details are provided to reproduce the core phenomena without the authors' code.

There appears to be no mathematical description of the core predictive learning algorithm. Only 6 equations exist throughout the paper, Eq. 1 data analysis (clustering), Eqs. 3-6 are for the PredNet model from Lotter et al. 2016. And Eq. 2 provides very little information about the predictive learning algorithm. So I couldn't find out how the core predictive learning algorithm proposed by this paper is implemented. This is in contrast with common computational papers that clearly define the model used in the Methods section.

As noted previously, much of the extensive revisions were focused on addressing these comments, common across reviewers.

(2) The model is overly complicated, obscuring the key point being made by the authors.

The authors propose a specific role for pulvinar, yet the model contains a great amount of details that may not be important for demonstrating the role of pulvinar. For example, the model contains a saccade component, is it really necessary? It certainly makes the model more powerful, but it may make it harder for the readers to understand the model. Similarly, could the authors have shown the same results without the where pathway, when there is no saccade and the object remains centered (two reasons cited for the necessity of the model where pathway, p.21)?

The addition of the much simpler sequence-learning model hopefully addresses at least some of this issue, providing a much more transparent, simple application of the core predictive learning theory.

In the larger-scale WWI model, every detail of the model is there because either it makes a measurable impact on the way the model learns, or it is consistent with well-established biological data (and reassuringly often, both). The appendix provides some extensive figures and documentation explaining how the model connectivity aligns with this biological data. We also provided a tiny sample of the many, many such tests we have run with the model to validate the importance of its properties, in Figure 11. This figure specifically shows the significant effects of lesioning the Where pathway.

If no saccades are included, then we cannot address the predictive remapping effects (Figure 12), and we also think this role of predicting the effects of motor actions is an essential part of the predictive learning story in the brain. So, yes, we could probably make a simpler model that didn’t have that component, but it does significantly contribute to the difficulty of the prediction task, in ways that we think capture how it works in the brain. So, in our judgment, this was important enough to merit the extra complexity in the model.

If there are no saccades and the object doesn’t move, then almost certainly the prediction task will become too easy (“degenerate”) and it just won’t learn much at all because it will very quickly achieve high levels of accuracy. The more difficult the prediction task, the more prediction errors, and the more the model can learn.

I believe the manuscript could be much more approachable to a broad readership if the authors to explain their key model mechanism in a simpler model setup, preferably not relying on models that are only understandable when reading through the code.

We have now done this.

(3) The manuscript is too long, with extensive discussion and literature review that are more appropriate for a review/opinion piece instead of a research article.

The paper is very long (~12,000 words, compared to ~5,000 words expected in regular eLife research articles), perhaps substantially longer than average eLife papers. This is not a major issue because the authors can trim it.

The main issue is that the authors spend too much space proposing the pulvinar as suitable for predictive coding (page 5-16), substantially delaying the introduction of the actual model (actual results start from page 16). This structural issue makes it difficult to evaluate this manuscript as a research article because a substantial proportion is literature review and synthesis in nature (there is nothing wrong with thorough literature review, simply less appropriate for a research article).

We agree that it was too difficult to provide a satisfying explanation of this model in the shorter eLife format, and are hopeful that it will find a more suitable home here at JOCN. We did move the more extensive aspect of the literature review to after the models are introduced. We think this neuroscience data is essential, because the major contribution of this work is theoretical in nature: we are advancing a specific biological theory, and it is essential therefore to discuss the relevant biological data that bears on this theory. Furthermore, we include a number of novel testable predictions that follow from this theory, which we feel is also a critical contribution of theoretical and modeling work such as this.

Reviewer #3:

(Please also see attached PDF)

Summary: This work proposes to use the temporal difference between the neuronal predictions for the next stimuli and the subsequent inputs from earlier layers in the pulvinar as error signals to drive the learning of the other visual cortical areas. The authors implement this learning mechanism in a large-scale model of the visual system and show that the simulated inferotemporal (IT) pathway develops abstract representations of inputs. It is further claimed that the learned IT pathway categorizes and groups 3D objects according to their shapes and the resulting grouping matches human judgements. Moreover, the authors also show that this biologically plausible framework surpasses alternatives in how abstract the learned IT pathway is.

Overall Reaction: The problem addressed in this work is undoubtedly important for the community: how the visual system develops its strong representations without supervision. I also appreciate the authors' efforts in building a large-scale model and in formulating a predictive-error-driven learning mechanism together to simulate the learning of the visual cortex. Furthermore, both the model and the learning mechanism are biologically plausible and integrate many neuroscience experimental findings, making it easier to imagine how the proposed model can be instantiated by real organisms. The authors also claim that the learned representations in the simulated IT pathway are similarly abstract and category-selective as human judgements. According to the comparisons shown in the paper, the proposed model also surpasses alternatives including PredNet and a backpropagation model with the same architecture and error signal.

We appreciate this summary statement, which we think captures the contributions of the paper well.

Nevertheless, I'm concerned that the paper doesn't really support its claims in a strong way. My concerns are a set of issues that are pretty serious in my view, half-way between mere technical cavils and high-level conceptual gaps, so I think are quite important to address before publication.

Specific comments:

1. In this article, the authors claim that the proposed model develops abstract representations and surpasses alternatives. However, these claims are seriously undermined by the fact that both the model and the alternatives are trained and tested on the same small dataset. This fact hurts the support for the claims in at least two ways. First, as the proposed model (WWI) is trained on the same set of objects that are used to measure the Representational Dissimilarity Matrixes (RDMs) that are later compared to human subject judgements and other models, it is highly likely that observed category-selective representations are a result of overfitting on this dataset and cannot generalize to other categories. Secondly, both the PredNet and the backpropagation models are only trained on the small dataset and in fact evaluated on the same training dataset. Although it is shown in the paper that both models reasonably fit to the training videos, they may very well overfit to the training set and yield trivial solutions that do not generalize to held-out categories. The authors may argue that the fact that these alternatives easily overfit while the WWI model can yield non-trivial solutions also supports the superiority of the proposed algorithm. However, it is unclear whether this overfitting can be avoided through properly tuning the hyperparameters or the training curriculum, as even the hyperparameter search conducted for the alternatives is done on the same dataset. Therefore, it will be good to see the evaluations of these models and the comparisons on them to human judgements done on a dataset with different categories or at least with different configurations (object orientations, size, rotation speed, etc).

Although the general concern for overfitting is understandable, we do not think that it applies in this case, at least not in the standard kinds of ways. Unlike most models, this one is not trained explicitly on a small fixed set of object category labels. The only inputs are bitmap images. Furthermore, each of these images is unique, across the 512,000 images that the model receives. There is extensive, completely randomly sampled variability (using uniformly distributed floating point numbers) in trajectory, 3D rotation, and saccades, across all of the images, which are also sampling randomly across 156 different 3D objects. Thus, it is actually quite a large dataset from this perspective (e.g., it is massive in terms of underlying systematic dimensions of variation, and dimensionality of inputs, compared to the MNIST digit recognition dataset, which is still widely used in current biologically-based learning papers). And again, unlike MNIST and ImageNet, etc, there is no low-dimensional supervised training signal that the model could seize upon to overfit to, and thus the standard notion of a training and testing dataset are not as clearly applicable. We could introduce entirely new objects, but assessing the nature of the model’s response to such objects is not well-defined, and we are planning to engage in a much more detailed such analysis once we can scale up the model to the point of being able to train on a much larger set of images for which relevant empirical data is available.

Furthermore, we now note that this model was initially developed using an entirely different dataset, with objects rendered as arbitrary combinations from a feature vocabulary, which enabled us to look specifically at held-out such combinations during testing. The details of this dataset and results can be found in an earlier preprint: <https://arxiv.org/abs/1709.04654> The reviewer’s reaction to this paper was that our dataset was too artificial: we needed to run on “real” bitmap images with realistic objects, which we have now done. In our tests, the same model features that were important on one dataset were also important on the other. Thus, there is no sense in which the model itself was overfit to this specific dataset, and as noted earlier we have applied similar models with similar parameters to auditory speech data.

2. Another concern I have is that although the WWI model has a hierarchical architecture that mimics the visual cortex, including V1, V2, V3, V4, and IT cortical areas, the learned representations seem to only have two levels of representations: V1-like layers and IT layers. This is especially contradicting the authors' claim that the WWI model develops progressive representations that are similar to the progression from V4 to IT in macaque monkey visual cortex, which is shown in Figure 5. In fact, the very figure that follows Figure 5 shows that all the layers that are lower than IT layers are almost identical to V1. It is difficult to imagine how a model with only two effective levels of representations can yield representations in its higher layers that are claimed to be similar to the IT area, whose functions are supported by the cascades of a series of cortical areas. The fact that the WWI model only has two levels of representations further strengthens the worry that the similarity between the model's representations and human judgements may only be a result on this particular set of objects that the model is trained on and cannot generalize to other stimuli. I believe it is critical for the authors to clearly explain this inconsistency between the WWI model and the visual cortex.

This is a consequence of aggregating across the different views of each object exemplar, which obscures important differences in lower levels of the network. These layers have much more dynamically varying representations across the course of the trajectories, as a function of spatial differences, but because we’re not factorizing the data in that way, this variation is all lost, and the layers appear similar to V1 in this aggregate measure. This is a limitation of trying to reduce a very high-dimensional space down to a single overall summary number as shown in the new Figure 6. Thus, actually examining the RDM’s, as we do next, is important for then revealing the more progressive nature of the transformations across layers.

As we note in the paper, this overall pattern is consistent with what is known about the primate visual system, where higher-order categorical structure emerges only at higher layers (including up into PFC), while lower layers have much more variable representations, and are building up spatial invariance more with respect to features, rather than developing object category structure (e.g., the seminal work of Tanaka and colleagues).

3. I am also concerned by the way the authors compare different models or compare the WWI model to the human judgements. In my view, the authors compare the representation in two ways: one way is through plotting the models' RDMs and eyeballing them to tell which one is closer to the target (Figure 4 and Figure 5) or which one is more "reasonable" (Figure 7); the other way is to first get clusters of categories through iteratively adjusting the cluster assignments according to the category similarities measured by the models and then compare different cluster assignments generated by models to see which one is better aligned with that from humans. Both ways are qualitative or even subjective. Having qualitative comparisons can help readers develop intuitively understanding. However, it also leaves space for readers to interpret the results themselves and can weaken the supports for the claims. For example, Figure 7 attempts to compare the backpropagation models (BP), PredNet, and WWI. According to the plotted RDMs, I am somewhat convinced that BP is weaker than WWI, as the RDM in Fig 7.a only roughly gives two big categories and is therefore less category-selective. This impression is strengthened by the correlation result shown in Fig 7.c, which shows that even the IT layers in BP are similar to V1. The comparison between PredNet and WWI is, nevertheless, less convincing to me. The authors claim that the RDM in Fig 7.b is "less cleanly similar within categories" and "overall follows a broad category structure similar to V1". But I find that the two RDMs in Fig 7.b and Fig 7.f (the RDM of WWI model in category orders of V1 clustering) share some similarities, such as messy diagonals and boundaries (see Fig. 1 of this review in PDF, boxes of corresponding colors). It is possible that PredNet also generates abstract representations, which can be better shown with a different category order. If the authors can show correlation analysis between V1 and all the layers in PredNet, it will then provide a quantitative measure about the quality of the representations. In fact, if the authors can conduct similar correlation analysis between similarity structure of human subjects and that of models for all models, it will also make the comparisons between different models more convincing.

We now include the Prednet equivalent for Figure 7c, initially in the new Figure 9, which is computed on raw exemplar-level similarity matrices and has no subjective element at all. This also shows no progressive differentiation across higher level layers. We agree that this raw similarity comparison is important for avoiding any subjective judgments, and all the critical results are substantiated by these analyses, so we feel confident that they are not due to any subjective judgments. The revision emphasizes this point by placing these figures first, and emphasizing the distinctions. We also appreciate the acknowledgement that the categorized figures are much more intuitively understandable, and when we present these results in talks, it is much more difficult to convey the meaning of plots like 7c compared to the others, so we think including both provides the best overall understanding.

4. I also think the claim in the abstract that "These categories ... are consistent with neural representations in IT cortex in primates.", one of the central claims in this paper, is not supported at all by the presented evidence. The major evidence for this claim is Figure 5 in the paper, which qualitatively compares the RDMs of the WWI model at V4 and IT layers to the RDMs of primate V4 and IT areas. However, the stimuli used to compute the RDMs of the WWI model are totally different from those used for the RDMs of primate V4 and IT areas, which makes this comparison not scientifically meaningful. It is not only that the categories of the stimuli are drastically different, but also that the complexity of the stimuli presented to primates is much bigger than that of the stimuli shown to the WWI model. More specifically, the categories of the primate stimuli are seven finegrained categories like animals, cars, and chairs. But the categories of the WWI model stimuli are five meta-categories, each of which includes several different objects that can be very different from each other. Although the authors put their categories into a specific order so that the RDMs of the WWI model are visually similar to the primate RDMs, this comparison is not meaningful at the beginning. In fact, even just the fact that primate V4 and IT neurons successfully distinguish these fine-grained categories while the WWI model similarly responds to the categories in the meta-categories is already clearly saying that the two representations are totally different. Moreover, the stimuli presented to primates are generated by combining irrelevant and naturalistic backgrounds and 3D objects with greatly varied orientations, scales, and positions, while the model stimuli are simply the 3D objects without any backgrounds and it is unclear how big the variations of the orientations, scales, and positions of the authors are. With such high-variation stimuli, the neuronal responses of primates are still category-selective. In contrast to this, the WWI model groups many categories into meta-categories even with such low-variation stimuli. The authors may argue that it is the progression from V4 to IT that they want to show in this figure. However, as I have mentioned in previous points, the WWI model only has two levels of representations: V1-like layers and IT layers. This figure is just showing that the WWI-IT representation is slightly more category selective than the WWI-V1 representation, but the category-selective property of the WWI-IT representation is significantly worse than that of primate IT or even V4 representations, due to the reasons described above in this point. In summary, I believe the authors need to either totally remove their point about the similarity between the model and the primate responses as well as the misleading Figure 5, or present much stronger evidence supporting this point than Figure 5.

We acknowledged in the paper that the data could only be compared at a very qualitative level, and there are many, many differences in the learning life of a primate compared to our model, that would shape these kinds of representations. Nevertheless, we do think that the qualitative similarity of the differences from V4 to IT is relevant, and as noted above it is not the case that there are only 2 different kinds of reps in the model. Furthermore, as mentioned above, there really is a very high level of variation in the inputs.

We nevertheless think that the qualitative nature of the differences between V4 and IT in the two cases are readily apparent visually in the figure, and provide additional insight into the model relative to the extant data on these brain areas. Furthermore, as noted below (and in the paper), the existing models applied to this data can be questioned on the basis of their reliance on human labeled data -- how much of their fit depends on the implicit knowledge conveyed in these labels, which we are not relying upon at all in our model? Is there any other model that shows the emergence of category structure like this without depending on explicit category inputs? Including this data provides an important occasion to discuss these issues.

5. As the authors are attempting to model how real organisms develop their visual system from real world stimuli, I believe it is critical to show the algorithm's ability in learning from much more general inputs in addition to the currently used toy-like datasets. The dataset used in this paper is different from real world stimuli in many ways like they do not have backgrounds, each of the video only has one object, and the transformations of the objects are very simple and consistent across different frames. The authors should test their algorithm on more realistic datasets such as SAYCam [1], a dataset collected by head-cameras mounted on infants during their development. If the model can be trained on that dataset, it will be important to analyze the learned representations. Otherwise, the authors need to provided explanations about why it cannot leverage real world stimuli, which are actually used by infants to develop their visual system.

Currently, the model does not have a functioning attentional system. In real brains, visual attention mechanisms, dependent in part on the parietal lobe spatial representations, are essential for processing cluttered visual scenes. For example, people with parietal damage suffer from significant impairments in processing cluttered visual displays (balint’s syndrome, simultagnosia, etc). Thus, until we incorporate such mechanisms, we do not think it would be reasonable to test our model on cluttered scenes.

Indeed, the fact that existing DCNN models succeed in processing cluttered visual images without such attentional mechanisms could be considered grounds for invalidating these models as accurate models of the primate visual system. Furthermore, the way in which they achieve this feat involves an extreme sensitivity to texture-level information, not shape information, and this texture-based solution is a critical reason why they are so susceptible to “adversarial images” that the human visual system has no difficulties with. Thus, this example demonstrates the limitations of a purely engineering-focused, ML-style approach for understanding at a scientific level the way in which vision works in the primate brain. Although one might argue that these models have been validated through tests like Brainscore, that does not address these points. The highest correlations of these models with brain data is only around .5, and it is likely that a significant amount of this is driven by the fact that they are explicitly trained on human-generated category labels.

The discussion section includes mention of the synergistic role of predictive learning and attention, and we have several promising results along these lines. However, a sufficient treatment of these issues requires at least one (and likely several) full papers on its own, which we are excited to embark upon. Thus, there is no way that we could include such results in this paper, where one of the primary complaints has been that the model is already too complex.

6. This work also needs more ablation studies and controls to support its claims. One control I think is important is to have an untrained model with the same architecture to show the effects of learning. For now, it is unclear how much the developed representations are a result of the proposed learning mechanism alone or a joint result of both the learning mechanism and the architecture. The architecture includes a complicated connection map across different layers. Although these connections may have some neuroscience experimental results as supports, systematic ablation studies showing the influence of different connections can still help readers better understand the architecture and the proposed learning algorithm.

As noted above, we have run extensive ablation studies, and included one with the Where pathway lesioned, in Figure 8 (new Figure 11). Also, the fact that these parameter manipulations result in very large reductions in the differentiation across layers shows that these results are not a simple consequence of the architecture alone. Furthermore, it is important to note that the BP model has exactly the same architecture, so clearly the architecture alone is not sufficient. These points were added to the paper.

7. Finally, I think the authors need to provide more explanations about how the learning works, even just explanations that give the readers some intuitions about how the algorithm works. For example, after reading the whole paper and some external links provided by the authors about the codes and the Leabra framework, I still find it hard to imagine how the temporal difference signal is used in the model. This temporal difference signal is the prediction signal between the actual outcomes and predicted outcomes and is hypothesized to be the only supervision signal used to learn the weights. It is therefore critical to understand how it is actually used in the model. In my understanding, the general weight updates in the model are done by the STDP-like learning rule described in <https://github.com/emer/leabra>, however, neither the paper nor the link explains how the temporal difference is integrated in this learning rule. Although the authors provide the source codes for the model (<https://github.com/ccnlab/deep-obj-cat/tree/master/sims/cemer>), I find it impossible to look into the codes, as they are mega-bytes files in specific formats without any explanations about how they can be read.

We have now added a much-improved Figure 2, which includes the critical learning equation, and more discussion about the nature of this temporal difference learning rule. Also, the github repository now contains a new from-the-ground-up replication of this model, implemented in our new simulation framework, which provides a much more transparent picture of exactly what goes into constructing and running the model. It is just one source code file.

We have also now included a much simpler model first, with just a single “hidden” layer, and a standard sequence prediction task, that can be downloaded and explored much more easily, to understand the essential computational properties of the model. We include a new Figure 4 that illustrates the behavior of this model over time, and we hope this should now make it much easier to understand how everything works.