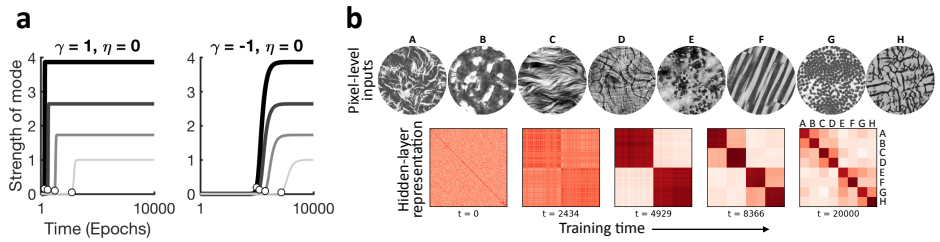

Rebuttal Figure 1: (**a**) Learning trajectories for two algorithms: $\gamma = 1$ vs. $\gamma = -1$. (**b**) Generalizing task to pixel inputs. Top row: texture stimuli (A to H). Bottom row: Covariance matrices of hidden-unit activities [for gradient descent ($\gamma \to 0, \eta = 0$)] exhibit progressive differentiation in the internal representation of hierarchical semantic structure.

- Thanks to R1 for the improved notation and wording. We agree O'Reilly's work is highly relevant and we should have (and will) cite/discuss Leabra. We will also link to the repository containing code for replicating all results.

- R2 had two technical concerns: (1) For progressive differentiation (paper Fig. 2a), R2 suggested an alternative measure normalized by overall learning speed. We had investigated several alternatives like this before settling on our metric. Consider Rebuttal Fig. 1a. These settings both show qualitatively similar progressive differentiation, which is captured by our mean lag metric. Further dividing by the time to learn the smallest mode would make the $\gamma < 0$ case appear much less stage-like, because the overall learning rate is faster for $\gamma > 0$ but slower for $\gamma < 0$ (notably it is not simply determined by the absolute value of γ). We found this normalization would make the 2D-map visualization unintuitive. We will clarify these points in the paper. (2) R2 argued that there is no learning when $\gamma = 0$: First, we note that the target is still clamped at the output for $\gamma = 0$ (see blue nodes in paper Fig. 1a) and could participate in learning. We will clarify in paper. Second, however, it is well known that CHL [Eq. (3)] is undefined when $\gamma = 0$ due to division by γ. When we wrote $\gamma = 0$, we always meant $\gamma \to 0$, i.e., the limit where top-down feedback is infinitesimal. We will correct these labels to be more precise. This limit $\gamma \to 0, \eta = 0$ reduces to standard gradient descent in which target info is back-propagated to enable learning, not that there is no target info. We agree with R2 and will tone down the bio-plausibility of the CHL network, and our work has no intention to compare the bio-plausibility of different CHL per se (Detorakis et al., 1999). R2 pointed out the similarity of our work to Saxe et al. (2019): Our work is indeed indebted to Saxe et al., but addresses a key limitation of their study. How distinctive are the phenomena that they identify and associate with gradient descent? We show that in fact a swath of learning rules give qualitatively similar behavior.

- R3 argued our 2D space is too restrictive: In fact we believe a major contribution of our work is proposing a minimalist space that nevertheless encompasses five commonly discussed learning rules. Our metrics do provide a methodology that can be used to characterize any desired learning rule. We will include the references on three-factor learning suggested by R3 since they are indeed relevant. R3 questioned the significance of paper Fig. 6. Fig. 6 shows the degree to which hidden-layer features are learned via error backprop vs. unsupervised Hebbian learning, a key distinction in many theories of neural learning. γ and η do scale the update size, but this does not directly translate to the integrated changes. E.g., if error is driven near zero, the CHL component will stop learning even with large γ. The integrated synaptic strength changes in the network model correspond to important training-induced plasticity that can be measured in electrophysiological experiments (Ahissar & Hochstein, 2004). R3 also argued that illusory correlations depend on the inversion symmetry. We explain here why this is not correct: The input-output map is invariant to this symmetry and our metric directly measures the illusory correlation in the output layer, and hence is not dependent on the SVD. Also we note that nonlinear models (whose dynamics cannot be described by the SVD) show similar phenomena (Rogers & McClelland, 2003).

- R2 and R3, deep network terminology: Saxe et al. showed that for deep linear networks, the real difference is between 0 vs. 1 hidden layer, and we are following this terminology. We will clarify in the text. R4: We do not claim that Hebbian learning can never yield progressive differentiation. We agree that in other tasks (particularly those where the unsupervised statistics are hierarchical too), progressive differentiation could occur. Our goal here was to start with a task environment well studied in prior work, that specifically does not have the target structure embedded in the inputs.

- R1 and R4 commented on the generalizability of our results to more realistic data sets and tasks beyond the ones we used in the paper. We have now simulated learning dynamics from pixels using distinct images as inputs and have seen similar results (Rebuttal Fig. 1b). The network is trained to learn novel semantic properties (e.g., "cells are made of a molecule X") assigned to perceptually distinct images representing skin samples of hypothetical alien creatures. We use this description of the stimuli because it is a cover story that could fit directly to a human cognitive-learning experiment on this topic (one example of how our theory could inspire or guide future experiments). Multivariate pattern analysis could reveal the evolution of the similarity structure of neural population vectors (recorded using fMRI/MEG) in response to different stimuli over the course of semantic learning. These neural data could be compared to the various theoretical predictions from our framework.