Models of hierarchical knowledge structures and error driven learning: proof of concept and benchmark tests using a “re-categorization” paradigm.

Tim Sparer

UIC

Abstract

Human learners inhabit dynamic environments: not only does a learner's environment change, but those changes themselves are subjects to further, systemic change. In these settings, how do learners distinguish between noisy signal, and "meaningful" (systematic, environmentally driven) change in that signal? This paper approaches these questions from the perspective of two complementary mechanisms: error driven learning, and hierarchically organized knowledge structures. It proposes that simple error driven learning algorithms, coupled hierarchically, can efficiently separate environmental signal from noise, and (to some degree) tease out underlying environmental drivers of surface level change. As illustrated in the paper, these structures give rise to "non-monotonic" learning of the type observed in higher level "conceptual change" processes. These ideas are operationalized in a computational simulation-model using hierarchical Bayesian learning. The model is tested on a set-shifting task (the re-categorization paradigm), and the resulting simulated data is compared to actual participants' data. The simulated data is seen to fit the participant data quite well.

Introduction

Human learners inhabit dynamic environments: Not only do observable environmental features change, but the deeper drivers of observable change are themselves also subject to change. To complicate matters, learning itself is a process of change, albeit "internal" change: learners must represent their environment and subsequently update their representations. How do learners trade-off updating their internal representation relative to both observable and un-observable environmental changes? If there is a nontrivial cost to learners changing internal representations, learners must balance between accurately capturing environmental phenomena and costly overfitting (changing their internal model too frequently). Consider the learner faced with some time varying signal, whose goal it is to capture the "true" value of the signal over time. If the signal varies around some mean, the learner can ignore the observed variance as noise: however, if there is some deeper change (a change in the actual mean of the signal) the learner would like to update his/her internal representation to reflect this. How, though, can a learner distinguish between meaningful signal change, and simple noise[1]?

Various techniques have been developed for these sorts of problem. Two such approaches are representation via hierarchical models, and error driven learning. In this paper, I suggest that a combination of the two approaches that makes for a powerful and

---

[1] An example of this sort of question is in the "debate" over global warming: climate deniers insist that any observed change in global temperature is normal variance, and that no underlying environmental change is occurring. Climate scientists however, suggest that normal patterns of variation alone cannot account for observed change.

flexible learning apparatus (which also exhibits similar behavior to so-called "non-monotonic learning" in human learners). The remainder of the paper is structured as follows: a brief background on the relevant learning mechanisms, a comment on non-monotonic learning in humans, and finally a more technical introduction to the specific benchmark task (re-categorization) and the learning mechanisms used to model subjects' behavior in the task.

Learning Mechanisms

An elegant solution to sorting noise from meaningful change can be found by combining error driven approaches to learning with hierarchical (Bayesian) models. The crux of this argument is that learners' maintain a model of the world "in their heads" and that this model is hierarchically organized, whereby constructs at higher levels of the model explain variance unaccounted for by lower model levels. The parameters of this model are, in turn, updated through online, error driven, learning.

Theoretical arguments for hierarchically structured cognition may come in many flavors. Two such arguments are mentioned here. The first argument, relative to the learning system, is the "cognitive economy" (efficiency of learning, compression, and search operations) of hierarchical knowledge structures. The second argument relates the operation of the learning system to ontological "facts of the matter" in the world: namely that, the world itself is structured as a hierarchy of changing phenomena, with otherwise

hidden factors (or indeed layers of hidden factors) driving observable change[2]. The two arguments are not at odds with one another; in an environment that contains hierarchically nested change mechanisms (as, arguably, humans' environment does), hierarchically structured learning can be both accurate and efficient. Evidence for these sorts of knowledge structures can be found in both the cognitive (e.g. Knoblich et al. 1999) and neuro-physiological literatures (Keibel et al. 2008). However, hierarchical structures can seem somewhat static: If hierarchical knowledge structures are accurate and efficient encodings of learner's environment, how does a learner go about changing them, if and when they are in error?

When faced with information that contradicts existing knowledge (error), learners can update their knowledge by taking the contradiction into account. So-called error driven learning is a common updating mechanism. Error driven learning has been observed in skill acquisition (Ohlsson 1996), neurobiological contexts (e.g. reward learning, (Shultz et. al, 1997)), and is generally considered a fast and efficient learning method (Mathys et al. 2011). Error driven learning algorithms operate by making a prediction about the environment, making an observation, comparing the observation and the prediction (the difference between the two is referred to as the prediction error), and then incorporating the weighted prediction error into a new prediction. The weighting

---

[2] Continuing the climate related theme from earlier, an example of these sorts of layers of change is the relationship between "weather" (a short term, observable phenomenon) and climate (long-term patterns of weather), and the underlying factors driving both (including "hidden drivers" of change at both short (day-to day), intermediate (seasonal) long-term (geologic climactic trends) time-scales. Here the differing time-scales can be thought of as differing "hidden levels/drivers of the relevant weather/climate phenomena).

factor on the prediction error is referred to as the learning rate, and adjusts how quickly errors are incorporated into new predictions.

Though common, and clearly useful, error driven learning does not appear to be "final word" in the business of learning from contradictions. Learners often appear resistant to change. For example, students may take years to learn that the earth is round, even when confronted with evidence to the contrary. Is it possible to adjust prediction error algorithms to account for this resistance?

One approach is to nuance the rate at which errors are incorporated into new predictions. Instead of a simple, static weight, the learning rate can be made to depend on (inferred) environmental variability. So-called precision weighted (error driven) learning, weights the importance of errors relative to inferred environmental variability: If errors are perceived to be precise (and therefore useful (as opposed to merely noise)) the learning rate can be increased, so that updating occurs quickly. If there is a great deal of variability in past errors (and therefore, inferentially, in the environment) current, observed errors, should be inferred to be the result of noise, and therefore should not influence new predictions too strongly. In such cases the learning rate should be decremented.

Interestingly, learners can estimate the precision of their errors using a similar, hierarchically organized prediction error process. In this process, the learning rate at the initial (environmental prediction) level is based on the inferred precision of past errors at that level. The precision of past errors is learned by a second, higher order, prediction function, which uses differences in past error estimates to predict future variance in error

terms. These structures, where "errors" pass up the hierarchy, and predictions about signal/error variability flow down (and influence learning rates) can then be nested in as many hierarchical layers as are useful for a given task.

The mechanisms for precision weighted error learning, then, appear to a bridge between error driven learning and hierarchical knowledge structures. More-over this may account for the dynamics of actual non-monotonic learning, initial resistance to change (any observed change is discounted as noise) followed by eventual acceptance and change (errors become repeated and precise enough to drive a change in the learner's predictions).

Application of learning mechanisms to a non-monotonic change (re-categorization) task

To test the hypothesis that these learning mechanisms are at play in even simple learning scenarios, a hierarchical learning model was built using Hierarchical Gaussian Filtering (Mathys __), a Bayesian learning technique based on the principles of precision-weighted error driven learning, (A fuller description of the model is provided in Appendix A). This model was tested in simple set-shifting task, the re-categorization task similar to the Wisconsin Card Sort Task (Grant & Berg, 1948). The re-categorization task requires subjects to inductively learn a category definition (where category definition is given as a mapping between specific stimulus feature and category outcome) from observing (and receiving feedback on) as sequence of stimuli. Once subjects have learned the initial definition the category definition is covertly changed. Subjects must then learn the new category definition. In this task environmental variability

corresponds to probabilistic feature-category mappings, and related variability in the stimulus sequence. "Meaningful" environmental change occurs during the switch, when the category definition changes.

Methods

(Study 1: Model fitting to previous data)

Design

The simulated-participant data was compared to the performance of actual participants in a "re-categorization" task. The participant data comes from 64 participants in Ramsberg and Ohlsson's (2013) recategorization with complete feedback condition.

Materials

Materials consisted of 64 fictitious bacteria, each of which differed along 6 binary dimensions. (In the stimulus materials, dimensions were referred to as features). The features (and their possible values) were: nuclei (grey or black), headbulbs (three or none), cell membrane (single or double), tail cilia (present or absent), cytoplasm (white or grey), and ribosomes (bent or straight).

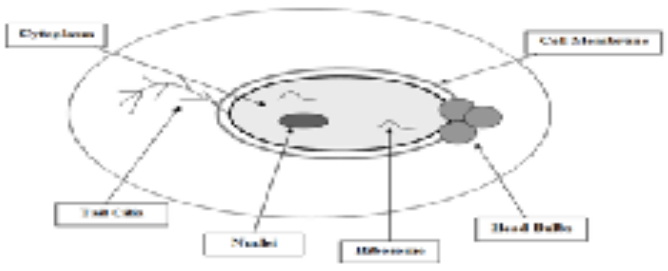

Figure __. Example Bacterium with parts labeled.

Procedure: Re-categorization Task

Participants were instructed to categorize fictitious, "martian", bacteria as to whether or not they (the bacteria) were oxygen resistant. Participants were told that some features of the bacteria might contribute to oxygen resistance. Over a period of 192 trials, participants were asked to infer which feature(s) might contribute to oxygen resistance. For the first 96 trials, a single value, on a single feature, conferred oxygen resistance to the bacteria. For the last 96 trials a different value, on a different feature conferred oxygen resistance. Participants were not informed of either defining feature/value, or the switch between features.

Trials. Each trial consisted of an image of a single bacterium and its attendant features. After viewing the bacterium, subjects were asked to indicate whether or not the bacterium was oxygen resistant. During *training trials*, participants received feedback as to the bacterium's status ("oxygen resistant" or "not oxygen resistant"). During *assessment trials* participants received no feedback after making their guess.

Useful terminology. The first 96 trials will be referred to as *misconception learning*, since participants learn a definition for oxygen resistance that is ultimately incorrect. The defining feature during misconception learning will be referred to as the *misconception*. The last 96 trials will be referred to as *target learning*, and the defining feature in these trials will be referred to as the *target*. In both cases the value of the

feature that confers oxygen resistance is referred to as the *positive value*, and the opposite value of that same feature is referred to as the *negative value* of the defining feature. (E.g. if head-bulbs are the defining feature, and three head-bulbs confers oxygen resistance, then the value "three head-bulbs" is referred to as the positive value, whereas "no head-bulbs" is referred to as the negative value). Instances that contain a positive value for the misconception and a negative value for the target will be referred to as *misconception diagnostic* instances, while instances with a negative value for the misconception and a positive value for the target are referred to as *target diagnostic instances*, instances which contain positive features for both target and misconception, or negative features for both, are referred to as non-diagnostic trials.

The period when the definition changes (after the 96th trial), will be referred to as the *switch.*

Blocks. Stimuli occurred in 16 instances each. Blocks were balanced with four non-diagnostic instances and 8 diagnostic instances (all of which were further balanced in terms of those containing the misconception positive values vs. those containing the target-positive values). An initial five misconception learning blocks were followed by a single assessment block. The switch occurred after the first five assessment blocks. After the misconception blocks and assessment, there were five target learning blocks, followed by an assessment block.

Model Simulation

The model simulated the performance of 36 learners on the recategorization task. The model implemented a hierarchical Gaussian filter (of the sort defined in Mathys et al.) and made use of the code available in the HGF Toolbox 3.0. Parameters for the model were set using empirical parameter estimates provided in Iglesias et. al ( ).

Results (Study 1)

Of the initial 64 subjects 36 met criteria to be included in the analyses. Criteria included participants responding correctly on 14 of 16 misconception learning trials for any of the misconception blocks (39 subjects), and evidence of effort on the target trials (three subjects' learning curves became flat after the switch, suggesting that they simply stopped trying and began to guess at random, these were excluded from the analyses).

The following analyses are intended to demonstrate the similarity of the learning curves produced by the simulated data to participants' actual learning curves. Broadly, this approach tries to find a curve that best fits the data (both real and simulated) and then tests to see if adding an additional categorical predictor (simulation vs. participant data) yields any significant improvement. If there is no significant improvement after adding the categorical predictor, this suggests that there is no difference between the simulated and actual learning trajectories. That is, no significant difference between the simulation and actual data suggests that the simulation accurately captures some aspect of participants' learning and behavior.

To capture the "non-monotonic" changes of interest in the learning curves, only diagnostic trials were included in the analyses. Further, to illustrate the (possibly)

separate processes of inhibiting the post-switch misconception definition while acquiring the target definition, two separate analyses were run. The first analysis was performed on simulated vs. actual data for the diagnostic misconception trials (proportion of responses consistent with the misconception), while the second analysis is performed on diagnostic, target consistent, trials (proportion of responses consistent with target).

Description of non-linear learning curve data

Both sets of learning curves exhibit the non-monotonic nature of the task (as illustrated below). For misconception responses, participants appear to initially respond at around chance, but quickly begin to learn the correct category definition (as indicated by the increase in responses consistent with the misconception defition. After the switch (block 5) responses consistent with the misconception decrease and bottom out, consistent with the idea that participants learn to inhibit the misconception. The target consistent responses exhibit an opposite pattern: the proportion of target consistent responses is initially low (suggesting that participants ignore the target feature), while after the switch, the proportion of responses consistent with the target increases, (suggesting that participant's correctly learn the target definition). As seen in figure below the simulation exhibits qualitatively similar behavior.

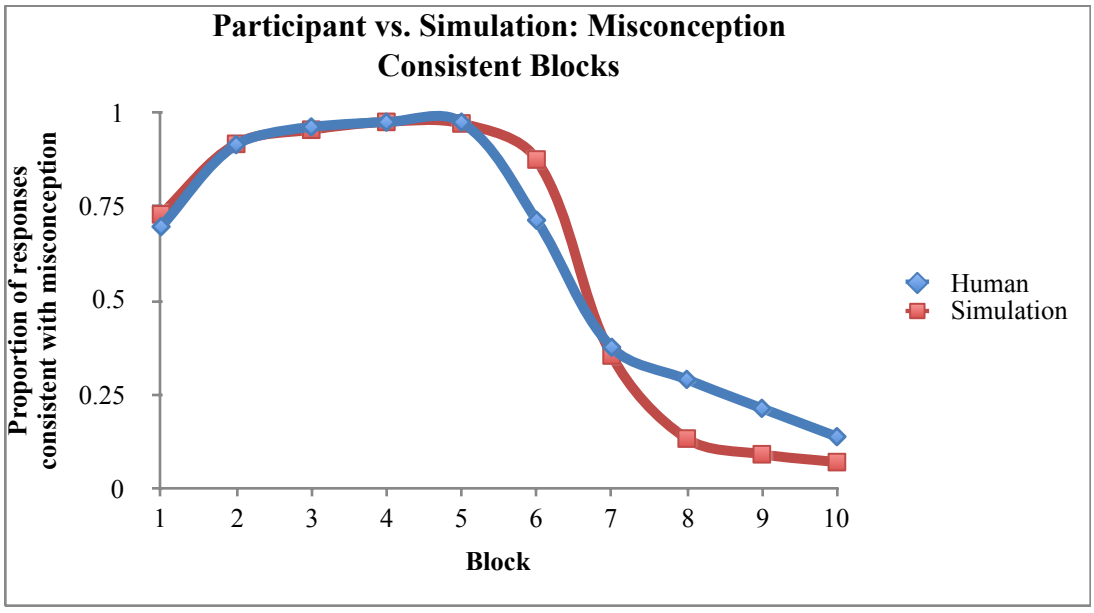

Figure 1. Proportion of responses consistent with misconception.

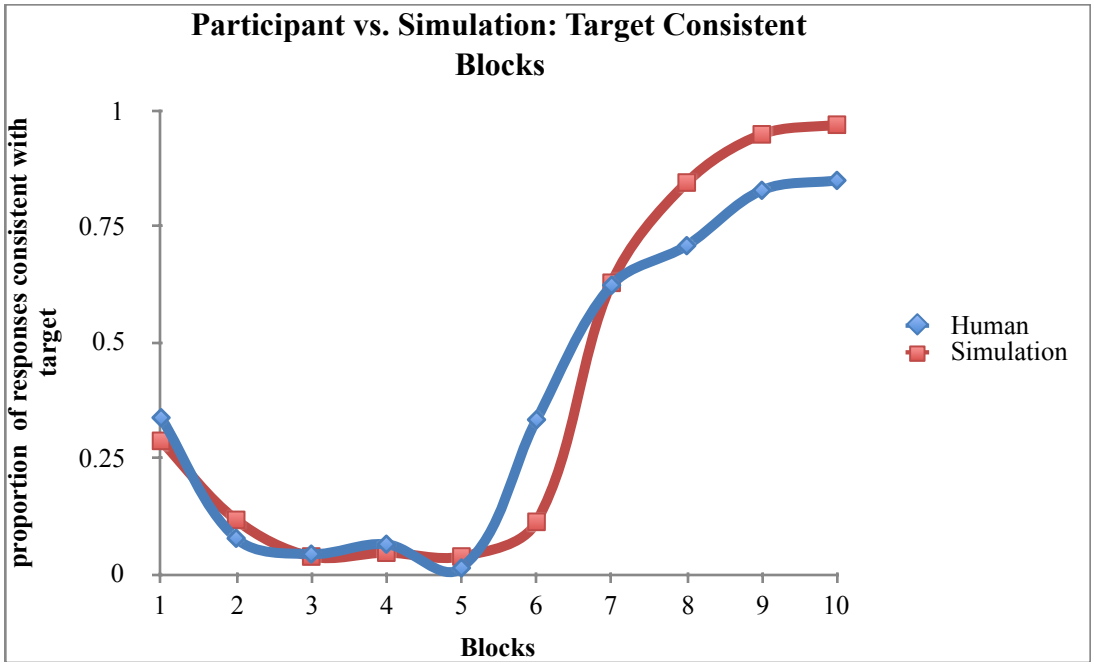

Figure 2. Proportion of responses consistent with target.

Analyses

To test the fit between the simulated and actual curves a cubic mixed model was run. The independent variable was the logit-transformed "proportion consistent [with target or misconception] score" per block. Predictors included the effects of time, time squared, time cubed, and data source (simulation vs. participant). Interactions between all of the time variables (including the quadratic and cubic terms) and the type of data source were also included. The first level of the model included terms for random intercepts, slopes, and quadratic terms, and assumed an autoregressive covariance structure. The fixed level of the model included all of these predictors plus the cubic, categorical and interaction terms.

At the fixed level of the model, a cubic regression was chosen as it seemed to best reflect the data structure (for the misconception, an initial upward sloping learning phase, followed by a decrease and plateau after the shift. For the target, a downward sloping initial phase, with an upward turn and plateau after the shift). At the random level, a quadratic structure (with the critical point of the quadratic curve corresponding to the shift in both the misconception and target scenario) and an autoregressive co-variance structure were selected to capture the underlying structure of the individual learning curves. The auto-regressive covariance structure was chosen for practical and theoretical reasons: practically, this was one of the covariance structures for which the model converged, and theoretically, because the auto-regressive structure captures the notion that, during learning performance on preceding trials influences performance on the current trial. (Note: Additional mixed models with both higher and lower order random

levels were tried. All of these other models (with the exception of the random intercepts model with an unstructured covariance matrix) either failed to converge or had a Hessian matrix that was not positive definite).

As expected for both target and misconception learning all of the initial time related predictors (linear, quadratic and cubic) were significant, while the categorical data-source predictor (simulation vs. participant) and subsequent interaction terms were not significant (see tables 1 and 2). This result suggests that there is no meaningful difference between simulated data and actual, participant data. The simulation model then, appears to describe participants' learning trajectories reasonably well.[3]

Target

---

[3] Note on the analyses. In order to fit the simulated data several analyses were tried. The most straightforward was a cubic mixed effects model, with random (linear, quadratic, and cubic) slopes and random intercepts. This model was initially run in R (in February of 2015) using the lme4 package. It converged and produced the anticipated results (no difference between the simulated and actual participants). To ensure that the effect was robust additional simulations were run and compared to bootstrapped data from the actual participants and multiple simulation runs. Strangely, however, when the analysis in R was run again in May of 2015 the same cubic mixed model did not converge. The author suspects a change in the optimization procedure used in the lme4 package (which, perhaps was auto-updated in the software at some point between May and February). In light of this, and as an additional measure to validate the simulation, several other models were run using SPSS software. The results included above are from the SPSS model. Also, the current author was not entirely sure how to assign credit for the R model (as it resulted from a collaboration between Carlos Salas, Alex Sokolovsky, and Tim Sparer) so, running a new model seemed important for reasons of academic integrity.

Parameter	*B Estimate*	*SE*
Intercept	-3.21**	0.97
Time	-10.70***	0.89
Time^2	2.64***	0.21
Time^3	-0.14***	0.01
Data Source (simulation vs. participant)	1.54	1.35
Source by Time	0.01	1.24
Source by Time^2	-0.29	0.29
Source by Time^3	-0.03	0.29

Note: **p =.001 ***p <.000, N = 72

Table 1. Mixed model for the misconception, estimated parameter values.

	Misconception	
Parameter	*B Estimate*	*SE*
Intercept	4.13***	0.99
Time	10.50***	0.91
Time^2	-2.50***	0.22
Time^3	0.13***	0.01
Data Source (simulation vs. participant)	-2.23	1.39
Source by Time	0.67	1.27
Source by Time^2	-0.03	0.30
Source by Time^3	-0.00	0.02

Note: ***p<.000, N = 72

Table 2. Mixed model for the target, estimated parameter values.

Methods (Study 2)

Design

The aim of study 2 was to derive predictions for two separate learning conditions from the simulation model, and to see if the predictions matched behavior in the two separate conditions. The two conditions were an "active learning" condition, in which the recategorization task was administered as in study 1, and a "passive" (or latent learning) condition, in which learner's did not make immediate responses after seeing a stimulus, but simply saw the correct category label.

Materials: The stimulus materials contained the same features, category, and "cover story" as in study one. For ease of presentation simple written descriptions, rather than pictures, were used (e.g. "this bacteria has THREE headbulbs, BENT ribosomes… etc.).

Subjects: A total of 29 UIC undergraduates participated in the experiment; 13 in the active condition, 16 in the passive condition.

Procedure

In both conditions, the procedure was nearly the same for study two as for study one. In the active condition the only changes were in the number and type of blocks. In the passive condition the structure of individual trials was altered as well.

Trials For the active condition the structure of the trials was the same as in experiment 1. For the passive condition the trial structure similar, but did not include a chance for participants to indicate whether or not they believed an instance was oxygen resistant. That is participants: Saw an instance, were asked to click through to the next screen, and then saw what category that the instance belonged to.

Blocks Training Blocks consisted of 8 trials each, balanced between six diagnostic trials (three misconception-present trials and 3 target-present trials) and two non-diagnostic trials (one "both misconception and target present", and one "neither target nor misconception present", trial each). Assessment blocks consisted of four trials each, including one trial of each type. (Two additional assessment blocks of 12 trials each, balanced, proportionally, similar to the training trials, were also included).

Order. The experiment consisted of a total of 16 training blocks and 16 assessment blocks. Each training block was followed by an assessment block. The switch occurred after the 8th assessment block. All of the assessment blocks, except for those immediately preceding the switch and the end of the experiment, were 4 trials long. The assessments immediately preceding the switch and the end of the experiment were 12 trials long.

Results

A one way analysis of variance revealed no significant differences between passive (16 participants) and active (13 participants conditions either pre- or post switch. Note: While all of the participants were included in the analysis, only 7 of 29 participants scored above 75% correct in either of the of the assessment conditions, suggesting either a lack of motivation or that the study was too difficult. In light of study one, it is more likely that this is simply due to motivational issues rather than task difficulty)

Mean Proportion Correct

	Misconception (Final assessment)	Target (Final assessment)

Active	.45	.46
Passive	.54	.54

*p<.05, n = 29

Discussion (Study 1)

The fit between the human and model learning curves suggests a certain "proof of concept" with regard to the model structure (hierarchically organized precision weighted prediction error learning), and provides a framework for future conceptual change modeling endeavors. For example, while the current analysis focuses on diagnostic conditions (i.e., misconception and target are treated as orthogonal), subsequent analyses should examine (non-diagnostic) conditions in which the relevant category features are either simultaneously present or simultaneously absent. Also, in the current model the Bayesian priors were set using empirical estimates from previous literature (Iglesias et al., 2013). While this avoids over-fitting, it may also overlook task specific nuances or individual differences; future research should attempt to capture individual differences by estimating the models' priors on a per subject basis. A third direction is model comparison: Testing the model against altered versions of itself (additional model layers, altered action selection functions etc.) and other models (RL, connectionist models, production rule systems, etc.). An interesting twist on this approach would be to alter the update functions of existing models to incorporate the HGF (e.g. using the HGF for the weighting function in production rule systems).

More generally this approach has implications for the type of mechanisms underlying (and therefore models used to simulate) non-monotonic learning. For example, while it has been suggested that non-monotonic learning may proceed from hierarchical structures (Ohlsson et al., 1999), a specific mechanistic account for change in such hierarchical knowledge has (to the authors' knowledge) not been well articulated. Alternatively, error driven theories of conceptual change have been offered (Kalish et al., 2005; Özdemir & Clark, 2007) but are often found wanting (Dole & Sinatra, 1998; Ohlsson, 2011; Ramsburg & Ohlsson, 2013). The HGF elegantly and normatively combines the two approaches; hierarchically coupling Bayesian error driven learning equations, and producing a nonlinear learning function as a result. While the generality and plausibility of the HGF have been remarked on here and in other sources (Iglesias et al., 2013; Mathys et al., 2011, 2014) this should not suggest that the results are limited to a single learning algorithm. The HGF algorithm belongs to a family of (Bayesian) models whose defining feature is hierarchical structure, where each level of the hierarchy attempts to predict, or "explain away", the errors of the next lowest level and serves as the source of errors for the next higher level. Theoretical work indicates that these sorts of model are likely to display chaotic and nonlinear behavior (Friston, Breakspear, & Deco, 2012). The results suggest that such models may offer a normative, mechanistic bridge between non-monotonic learning approaches that focus on the structure of knowledge and those that view change as a process of error accumulation and evidence acquisition. Finally, the results suggest the intriguing possibility that "deep" learning

may be an emergent function of the hierarchical coupling of simpler learning mechanisms.

Discussion (Study 2)

Since the model makes no predictions between the active and passive learning conditions, it was expected that there would be no difference between groups. This was in fact the case- however, given the trend in the data (and also large bodies of literature on the differences between reinforcement learning and classical conditioning [cite]) it is likely that this is more of an artifact of small sample size and lack of motivation on the task, than a true null result. Future studies should use a larger sample size, and take steps to check participants' motivation. Additionally, an model that explicitly accounts for active learning should be developed.

Works Cited

Dole, J.A., & Sinatra, G.M. (1998). Reconceptualizing change in the cognitive construction of knowledge. *Educational Psychologist, 33(2/3)*, 109–128.

Friston, K., Breakspear, M., & Deco, G. (2012). Perception and self-organized instability. *Frontiers in computational neuroscience*, *6*.

Grant, D. A., & Berg, E. A. (1948). A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card-sorting problem. *Journal of Experimental Psychology, 38,*404–411.

Iglesias, S., Mathys, C., Brodersen, K. H., Kasper, L., Piccirelli, M., den Ouden, H. E., & Stephan, K. E. (2013). Hierarchical prediction errors in midbrain and basal forebrain during sensory learning. *Neuron*, *80*(2), 519-530.

Kalish, M. L., Lewandowsky, S., & Davies, M. (2005). Error-driven knowledge restructuring in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 846.

Knoblich, G., Ohlsson, S., Haider, H., & Rhenius, D. (1999). Constraint relaxation and chunk decomposition in insight problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(6), 1534.

Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2008). A hierarchy of time-scales and the brain. *PLoS computational biology*, *4*(11), e1000209.

Knoblich, G., Ohlsson, S., Haider, H., & Rhenius, D. (1999). Constraint relaxation and chunk decomposition in insight problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(6), 1534.Ohlsson, S. (1996).

Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in human neuroscience*, *5*, 39

Ohlsson, S. (2011). *Deep learning: How the mind overrides experience*. Cambridge, UK: Cambridge University Press.

Ohlsson, S., & Cosejo, D. G. (2014). What can be learned from a laboratory model of conceptual change? Descriptive
findings and methodological issues. *Science & Education, 23,* 1485-1504.

Ohlsson, S. (1996). Learning from error and the design of task environments. *International Journal of Educational Research*, *25*(5), 419-448.

Özdemir, G., & Clark, D. B. (2007). An overview of conceptual change theories. *Eurasia Journal of Mathematics, Science & Technology Education, 3,* 351-361

Ramsburg, J. T., & Ohlsson, S. (2013). Category change in the absence of falsifying feedback. *Proceedings of the Thirty-Fifth Annual Conference of the Cognitive Science Society.*

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593-1599.

Appendix
Model: Further Elaboration

Motivation for Model

Recently, some researchers have suggested that many theories of conceptual change lack sufficient explanations as to why a given mechanism is important and under what circumstances it functions (Ohlsson, 2009; Rusanen & Lappi, 2013). In the current paper, we present a model of implicit learning that examines re-categorical change. The model utilizes hierarchical Bayesian learning and precision weighted prediction error (Mathys et al., 2011) to operationalize and normatively ground two key notions of non-monotonic learning: hierarchically organized knowledge structures and error driven feedback learning. The general form of the model and relevant theoretical components are discussed below, while specific parameter values and related concerns are addressed in the methods section.

The following observations and theoretical concerns motivate the model:

1. Conceptual change and re-categorization appear to be non-linear phenomena (Ohlsson, 2011).
2. Both generally, and in non-monotonic change, knowledge is often hierarchically organized (e.g., see Knoblich et al., 1999).
3. Past knowledge tends to interfere with knowledge being learned (Campinaro, 2002; Ohlsson, 2011).
4. Any model that addresses 1-3 should do so parsimoniously, in a manner that is consistent with known learning mechanisms, and is generalizable.

To address these concerns, we adopt a hierarchical approach for learning in re-categorization. With this approach, we presume that learners' do not just learn feature- category mappings (definitions), but also (implicitly) learn about the stability and volatility of those mappings over time. We suggest that such learning occurs in an interactive and nested manner. That is, subjects infer feature-category mappings as definitions, the relative stability of those definitions, and likely changes in the stability of category definitions. Inferring the stability and volatility of category definitions embeds prior knowledge more deeply than simply learning a category-feature mapping, and also serves to help the learner discriminate between random noise and meaningful change. As such, the nested, hierarchical structure of the model both captures the notion that prior knowledge can interfere with new learning, and suggests a reason for it (i.e., expectations about the stability of knowledge, learned over time, influence the way in which that knowledge is updated).

The hierarchical structure of the model is a consequence of the selected update function, in this case the *Hierarchical-Gaussian Filter* (HGF), a Bayesian learning algorithm (Mathys et al., 2011). In its simplest incarnation, the HGF acts like an error driven update algorithm: It makes a prediction, compares it to an observation, calculates the error between the prediction and the observation, and incorporates a weighted version of the error into the next prediction. However, unlike simpler algorithms, the HGF also makes predictions about its own errors, and updates them in like fashion. In this way, the model is hierarchical: It makes predictions about the world, but also predictions about its predictions (specifically how much error lower order predictions will have). Each "higher" level of the model tries to predict the errors of the level below, while each lower level incorporates these predictions by adjusting the weight that the lower level gives newly generated/incoming errors (In theory the HGF can have as many such levels as desired, with each new level predicting the errors of the level preceding it, and generating new errors of its own).

The structure of the HGF lends itself to the type of nested learning suggested to occur during re-categorization. Successive levels of the HGF can be interpreted in terms of the (inferred) stability or volatility of previous levels. Thus a learner using a multi-level HGF can be said to be making inferences about the underlying stability and volatility of the learning environment, as well as learning immediate, surface level relationships (such as category definitions).

Additionally, the dynamics of the HGF lend themselves to nonlinear learning. Recall that as each level of the model tries to "predict" the errors of the level below, the lower level incorporates the higher level's predictions by changing the weight it (the lower level) gives errors during updating. Since the weight given to errors changes the magnitude of the update and therefore the learning rate, this amounts to endowing the model with an error-dependent dynamic learning rate. Thus, the hierarchical coupling between model levels gives rise to nonlinear learning rates.

Finally, we note the generality of the HGF as a learning mechanism. The HGF is biologically plausible, mimics basic reinforcement-learning algorithms, is normative, and mathematically tractable (Mathys et al., 2011). The HGF requires no "extra" information beyond the normal input sequence in order to detect

deeper environmental changes, nor does it require any special processes beyond those used in error driven learning. Non-monotonic learning arises as an emergent property of hierarchical coupling of simple error driven learning mechanisms.

Specific Form of the Computational Model

General Model Form. While the learning function, in the form of the HGF, is the conceptual heart of the model, for clarity and completeness we elaborate the task-specific structures of the model here. The model can be conceptualized as a network with two layers: a stimulus representation and a category layer (depicted in Figure 1). Together they serve three main functions (categorization, action selection, and updating). The *stimulus representation layer* or attribute layer contains a node for each binary valued stimulus attributes (e.g. "has a nucleus or not"). The *category layer* has two unitary nodes, one node representing oxygen resistance and one representing oxygen intolerance. The stimulus representation (hereafter the '*attribute layer*') is connected to the category layer by links representing the estimated co-occurrence between an attribute and a category-label (All attributes are connected to all category labels, respectively).

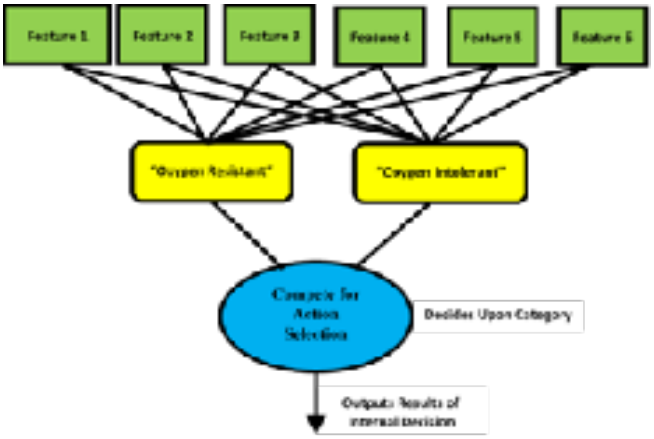

Figure 1. General Model Form.

The categorization function uses the co-occurrence estimates to calculate the likelihood (weight) that a given instance belongs to each category. The action-selection function selects the single, appropriate category-label for an instance by Gibbs sampling from the space of weighted category labels. The update function uses the hierarchical Gaussian filter described in the introduction, (see also: Iglesias et al, 2013; Mathys et al., 2011, 2014).