Alien Bacteria Found on Mars! A Model of Conceptual Change using the Re-categorization Paradigm

Tim Sparer (tspare2@uic.edu)
Jared T. Ramsburg (jramsb2@uic.edu)
Carlos Salas (salas5@uic.edu)
Stellan Ohlsson (stellan@uic.edu)

University of Illinois at Chicago
Department of Psychology (MC 285)
1007 West Harrison Street
Chicago, Illinois, 60607-7137

Abstract

Many conceptual change theories posit that change occurs due to a variety of cognitive, social, and emotional factors (Dole & Sinatra, 1998; Ohlsson, 2011), however, few theories have tested these claims via computational models of conceptual change. In this paper, we present a hierarchical Bayesian model that addresses change processes and their effects on re-categorization, a form of concept change. Human data from a study using the re-categorization paradigm (Ramsburg & Ohlsson, 2013) are compared to the computational model. The structure of the human data suggests the 'non-monotonic' nature of conceptual change (Ohlsson, 2011) as indicated by the best-fit learning curves. For several such curves, model comparisons suggest good fits between the computational simulations and human data. The nonlinear form of the model's update functions lends additional support to concept change as a non-monotonic process. The model is discussed as a "proof of concept" for future conceptual change modeling endeavors.

Keywords: Learning, recategorization, feedback, non-monotonic change, computational modeling.

Introduction

Often learners can be confronted with information that is contradictory to their pre-conceived notions of a given topic. The process of overriding an erroneous conception in favor of a new conception is referred to as conceptual change or non-monotonic change (Ohlsson, 2011). There are numerous real-world empirically investigated examples of conceptual change failures. For example, elementary school students may have trouble switching from the notion of a flat earth to a spherical earth (Vosniadou & Brewer, 1992), physics majors may still adopt naive erroneous conceptions of physics principles when faced with real-world physics problems (Halloun & Hestenes, 1985), even biologists can take more time to determine whether a plant is a living thing given their early childhood expectations that living things must be mobile (Goldberg & Thompson-Schill, 2009). Obviously, conceptual change is important because without it learning would stifle and many, if not all, of our cultural and technological advances would be absent. (Consider the caveman that thinks fire is dangerous and ought not to ever play with it).

Researchers have forwarded a variety of theories to explain why change does or does not occur (for overview see Ohlsson, 2011; Özdemir & Clark, 2007). A classic theory of conceptual change suggests that change is a process that involves learners first becoming dissatisfied with their misconceptions (Strike & Posner, 1982, 1992) due to the accumulation of errors. Similarly, theories like Theory-Theory (Gopnik & Wellman, 2012) and Category Shift Theory (Chi, 2005; Chi & Brem, 2009) also suggest that some form of dissatisfaction is needed in order to motivate revision of a concept. While still others attribute conceptual change to a combination of factors, which include dissatisfaction, but adds individual differences and contextual variables such as need for cognition, social context, personal relevance, motivation, the method of transmitting the information (Dole & Sinatra, 1998), and cognitive utility (Ohlsson, 2011).

A common theme among theories is that some form of instruction or feedback is needed in order to elicit change. For the purposes of this paper the focus will be, in part, on how feedback can be used to influence concept change. Feedback is considered to be one of the most important learning factors when a learner has already developed some knowledge within a domain (Ohlsson, 2008; Shute, 2008). Feedback that suggests that a learner's conception is false may be important for promoting the falsification process (Kalish et al., 2005). Feedback that suggests that a new conception is correct may be important for promoting revision and instilling confidence in the new conception (Butler & Wine, 1996). Taken together, inductive and falsifying feedback should produce change.

One method for investigating conceptual change in the laboratory is using a *re-categorization paradigm* (Cosejo et al., 2009; Ohlsson & Cosejo, 2014; Ramsburg & Ohlsson, 2013). The re-categorization paradigm takes learners through a categorization task. Participants learn to categorize items based on feedback about whether their decision regarding category membership for a given item is correct. After a predetermined set of training trials, the feature that is important for categorizing changes unbeknownst to the participants (this is analogous to switches with the Wisconsin Card Sort Task; Grant & Berg, 1948). The goal for participants is to learn to override their prior categorization method and adopt the new method by using corrective feedback after each training trial. The version of re-categorization that we used in this study presented stimuli that mimicked a science-learning scenario (Ramsburg & Ohlsson, 2013). Fictitious alien bacteria were

categorized with respect to their resistance to atmospheric oxygen.

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).

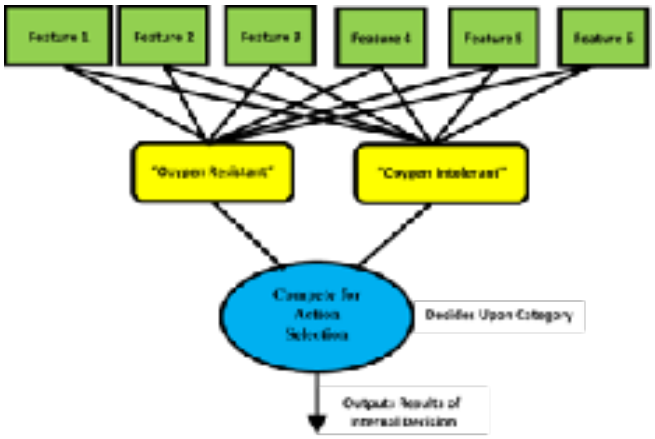

Figure 1. General Model Form.

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).

Method

Design

The data used to test the model were taken from Ramsburg and Ohlsson (2013). Their study was a between-subject comparison of two different feedback conditions. The model was fitted to the data from one of those conditions ("Complete"; N = 64).

Materials

The materials consisted of 64 fictional bacteria images (see Figure 2). The bacteria have six different parts that have different binary attributes resulting in 64 complete variants: Nuclei (grey or black), Headbulbs (three or none), Ribosomes (bent or straight), Tail Cilia (present or absent), Cell Membrane (singular or double), and Cytoplasm (white or grey).

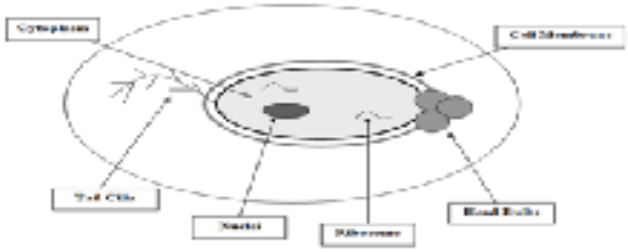

Figure 2. Example bacterium with parts labeled.

Misconception Learning. Participants first learned to categorize whether an alien bacteria was oxygen resistant based on feedback that supported the misconception feature (i.e., *black nuclei*) over the course of five training blocks of 16 trials each. Each training block was balanced to include in randomized order six images that contained the misconception, six images that contained the target, two images that contained neither, and two images that contained both the misconception and the target. After five training blocks, unbeknownst to the participants the feature that determines oxygen resistance changed to *bent ribosomes* (i.e., the target).

Target Learning. Participants had five target training blocks of 16 randomized trials to learn that bent ribosomes determined oxygen resistance. Target learning consisted of stimuli that were similar to what participants had already used for classification. Each training block was balanced to include in randomized order six images that contained the misconception, six images that contained the target, two images that contained neither, and two images that contained both the misconception and the target. All parts of the bacteria were visible on the screen allowing participants to falsify their prior categorization in favor of a new categorization supported by the computers feedback. For example, during the misconception training, the participant learned that black nuclei are responsible for oxygen resistance. During the target training, the participant was confronted with an image containing black nuclei with feedback stating that the bacterium was not oxygen resistant. This feedback should allow the learner to negate the prior conception. Moreover, when the learner is confronted with an image that does not have a black nucleus, but is shown to be oxygen resistant the learner should infer that another part of the bacteria is responsible for oxygen resistance.

Procedure

Behavioral Experiment. Participants were seated in separate cubicles. Each participant was instructed to first participate in a training session, which consisted of a series of PowerPoint slides outlining how one can sort a variety of objects into different categories. The training session ended with participants categorizing stick figures based on their features. When participants finished with the initial training activity, they were instructed to participate in the more challenging bacteria paradigm.

Participants read the instructions for the task on the computer screen and asked questions if needed. Participants were given a script stating that alien bacteria was recently discovered on a distant planet and that scientists needed to determine whether there were oxygen resistant variants of the bacteria. Participants were then asked to rate how important each feature was in determining oxygen resistance on a 7-point Likert scale from 1 (Not at all) to 7 (Extremely). After rating the features, participants went through a prompt that described the importance of determining which bacteria were oxygen resistant. Each participant was tasked with determining whether the pictured bacterium was oxygen resistant. Participants indicated their response via the keyboard. The following responses were acceptable: y= yes, n= no, d= don't know. Participants would then receive immediate feedback from the computer either stating that the bacterium was or was not oxygen resistant. Participants were instructed to make as few errors as possible. After completing all tasks, participants were debriefed.

Model Simulation. The model completed an analogous process. It was fed a series of input vectors consisting of binary strings, each representing co-occurring feature attributes. For example, the string {100101} represented a bacterium with a grey nucleus (1), no headbulbs (0), straight ribosomes (0), present tail cilia (1), double cell membrane(0), and white cytoplasm(1). Reading in this string was akin to a participant viewing a bacterium image on a given trial. The model received a set of these binary strings vectors (one per trial), generated a binary response (e.g., 1 = oxygen resistant, 0 = oxygen intolerant), and received feedback (e.g., 1 = oxygen resistant, 0 = oxygen intolerant.). Like participants in the behavioral experiment, the model also completed five misconception and five target training blocks of 16 weighted randomized trials. The model was implemented in Python 3.4.1, the code for the HGF was adapted from the HGF Toolbox 3.0 , and the Bayesian priors were set using previous empirical data (Iglesias

et al., 2013). The model ran 36 simulated learners; the priors for each were set by sampling from the empirically estimated parameter distributions provided in Iglesias et al. (2013).

Results

Thirty-nine of 64 participants met the criterion for inclusion in analyses (i.e., correctly classifying 14 of 16 alien bacteria in any of the initial five training blocks). The inclusion criterion was chosen as a way to ensure that we tested participants who were successful in learning the misconception. An additional three participants were excluded from analyses based on flat trajectories across trials, which suggested that they never inferred the category switch. We wanted to examine whether participants could override their prior conception, therefore we needed to be certain that they had learned the misconception before we could assess their target learning.

Overview of Responses by Stimulus Type

We mapped the types of responses participants gave for proportion of misconception consistent and target consistent responses per training block, averaged over 36 participants (see Figure 3). The change in the category occurred between blocks 5 and 6. Performance on the non-diagnostic trials is not reported because only the diagnostic trials are relevant to the current model.

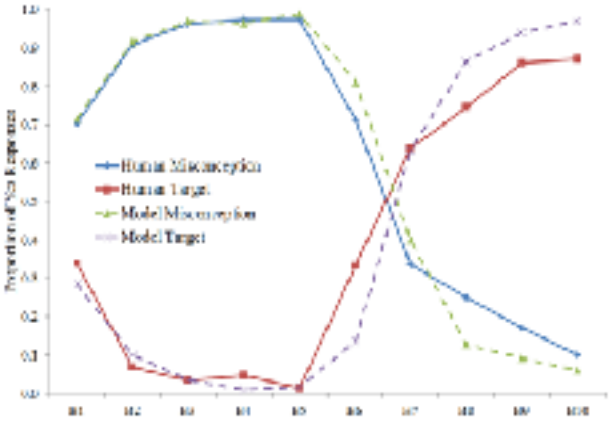

Figure 3. Proportion of misconception consistent and target consistent responses per training block.

Growth Curve Model Fitting

Mixed-effects models were used to assess whether the same polynomial growth curves reliably fit the human and simulation data (i.e., does the human and simulation data follow the same learning curves?). This family of models was chosen because, unlike traditional OLS regression, it allows one to analyze growth curves while simultaneously accounting for both individual-level effects (i.e., individual variability in intercepts and slopes) as well as group-level effects (e.g., simulation vs. human data). Polynomial growth curve models were run in R 3.1.2 (R Core Team, 2014) using maximum likelihood estimation via the *lme4* package (Bates et al., 2013).

Given the shape of the misconception and target learning curves, we parameterized the full model as containing third order polynomial fixed effects of Block (i.e., time) and a dichotomous fixed effect of Origin (i.e., simulation vs. human data). Fixed effects were entered as follows: First, a dummy coded Origin variable was entered with humans coded as the reference group. Next, linear, quadratic, and cubic Block terms were entered, representing the human coefficient terms for each slope. Finally, Origin X Block interaction terms were entered, representing the simulation versus human differences for each polynomial slope term. Following Baum's (2008) approach, the proportional outcome variable (i.e., average proportion of participants/ simulations per block responding yes) was submitted to a logit transformation prior to fitting the models.

Model Comparisons. A series of models were specified starting from a baseline Null model (i.e., containing only participant random effects) and adding slope X Origin interaction terms in single polynomial order increments. Model evaluation criteria consisted of measures that assess the probability of observing the data given each model being estimated. These included the maximized model's log likelihood (larger is preferred), and both Akaike's and Bayesian Information Criterion (AIC/BIC), which assess the relative likelihood of a given model accounting for observed data taking into account the model's goodness of fit and complexity (i.e., number of parameters). Lower AIC/BIC values represent the best empirically supported model, with BIC taxing complexity more heavily than AIC (Burnham & Anderson, 2004). Thus, a smaller set of log-likelihood values and a larger set of AIC/BIC values for the null and lower order models relative to the cubic model would suggest that the cubic model best fits observed data.

Misconception Learning Curves. Table 1 displays the model comparison tests. All polynomial models indicate a better fit than the null model based on the AIC, BIC, and log likelihood tests. Furthermore, these fit criteria unanimously favor the cubic model, suggesting that a cubic relationship best accounts for the shape of the non-monotonic learning curve.

Table 1. Model Comparisons for Misconception Condition based on Fit Statistics and Log-Likelihood Tests

Model	Model *df*	AIC	BIC	Log-Likelihood	$\chi^2(df)$
Null (Baseline)	5	5498.80	5521.80	-2744.40	
Linear	6	5356.70	5393.50	-2670.30	148.09 (3) ***
Quadratic	10	5128.30	5174.40	-2554.20	232.34 (2) ***
Cubic	12	4970.20	5025.50	-2473.10	162.16 (2) ***

Note. All polynomial models contain fixed effects of Origin and their respective "Polynomial" X Origin interaction terms. $p < .001$ ***

Table 2 displays the individual coefficient tests for this cubic model. The p-values were calculated using the Satterthwaite approximation (SAS Institute Inc., 1978) via the *lmertest* package (Kuznetsova et al., 2013). The significant intercept term and non-significant Origin term suggest that the mean human proportion of yes responses was significantly different from zero, but not different from the simulation mean. All of the polynomial slopes were significantly different from zero in the human data. Furthermore, the non-significant Origin by polynomial interaction terms suggests that these slopes were the same for human and simulation data. These converging results support the claim that the cubic fit best captures the pattern of data for both humans and the simulation.

Table 2. Coefficients for the Fixed Effects Variables in the Misconception Condition.

Factor	$M(SE)$	df	t-value
Intercept (Humans)	5.12 (.45)	55.2	9.5***
Origin	0.51 (.07)	708.0	6.7[illegible]
Linear	-23.89 (1.35)	57.5	-23.15***
Quadratic	-11.64 (0.96)	702.2	-12.1***
Cubic	10.81 (0.96)	702.2	9.5***
Linear X Origin	22.10 (1.35)	708.2	1.63
Quadratic X Origin	-11.44 (1.34)	702.2	-0.85
Cubic X Origin	-0.71 (1.34)	702.2	-0.08

Note. Coefficients estimated using Maximum Likelihood. Factor p-values were calculated based on Satterthwaite approximations of the degrees of freedom. $p < .05$*; $p < .01$**; $p < .001$***

Target Learning Curves. Fixed effects for the target were entered with the same procedure as the misconception consistent responses. Table 3 presents the model comparison tests. As with the misconception fits, all of the polynomial models indicated a better fit than the null model. The AIC, BIC, and log likelihood tests again unanimously favored the cubic model.

Table 3. Model Comparisons for Target Condition based on Fit Statistics and Log-Likelihood Tests

Model	Model df	AIC	BIC	Log-Likelihood	χ^2(df)
Null (Baseline)	3	5527.50	5550.00	-2760.75	
Linear	6	5381.20	5418.00	-2682.60	152.59 (3) ***
Quadratic	10	5062.60	5118.70	-2521.30	302.56 (4) ***
Cubic	12	4926.30	4988.40	-2452.10	138.50 (2) ***

Note. All polynomial models contain fixed effects of Origin and their respective 'Polynomial' X Origin interaction terms. $p < .001$ ***

Table 4 displays the individual coefficient tests for the target learning cubic model. The pattern of results were virtually identical to those obtained for the misconception consistent responses, albeit with flipped curves (i.e., consistent with Figure 2).

Table 4. Coefficients for the Fixed Effects Variables in the Target Condition.

Factor	$M(SE)$	df	t-value
Intercept (Humans)	10.24 (1.96)	689.98	5.23***
Origin	1.24 (2.67)	695.88	0.47
Linear	-16.09 (1.32)	674.16	-12.18***
Quadratic	3.87 (0.25)	688.98	12.32***
Cubic	-3.14 (0.01)	693.98	-31.34***
Linear X Origin	6.66 (1.04)	688.78	0.27
Quadratic X Origin	-0.35 (0.35)	693.98	-0.11
Cubic X Origin	0.13 (0.02)	688.98	1.05

Note. Coefficients estimated using Maximum Likelihood. Factor p-values were calculated based on Satterthwaite approximations of the degrees of freedom. $p < .05$*; $p < .01$**; $p < .001$***

There were no differences in the mean proportion of yes responses between the human and simulation data, as indicated by the intercept term and Origin terms (although both differed from zero). Again, all of the polynomial slopes were significantly different from zero in the human data, and the non-significant Origin by polynomial interaction terms suggest that these slopes were the same for human and simulation data. These results suggest that the cubic fit best captures the pattern of data for both humans and the simulation.

Discussion

In this paper, we examined whether participant data for the re-categorization paradigm could be modeled with a hierarchical, error driven learning model. As expected, the empirical re-categorization data exhibited non-monotonic structure in the form of cubic learning curves. The model output exhibited similar structure. The growth curve analyses suggest that the computational model reliably simulates human data for both misconception and target learning trajectories. For the misconception consistent trials, human and computational data both demonstrate a sharp initial increase in the proportion of yes responses that asymptotes until the presentation of the category switch, at which point the proportion of yes responses sharply declines. Predictably, the proportion of yes responses for the target consistent trials does not show an increase until after the category switch, but they also show a sharp rise to asymptote by the final block. For the target and misconception learning trials the computational model appears to be a good fit.

The fit between the human and model learning curves suggests a certain "proof of concept" with regard to the model structure, and provides a flexible 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 our 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 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 our 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 (see Clark, 2013 for a review). Theoretical work indicates that these sorts of model are likely to display chaotic and nonlinear behavior (Friston, Breakspear, & Deco, 2012). Our 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, our results suggest the intriguing possibility that "deep" learning may be an emergent function of the hierarchical coupling of simpler learning mechanisms.

References

Baum, C. F. (2008). Stata tip 63: Modeling proportions. *Stata Journal*, *8*, 299.

Bates, D., Maechler, M., Bolker, B. M., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4*. ArXiv e-print; submitted to Journal of Statistical Software.

Bergwerff, G., Meijering, B., Szymanik, J., Verbrugge, R., & Wierda, S. M. (2014). Computational and algorithmic models of strategies in turn-based games. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference understanding AIC and BIC in model selection. *Sociological methods & research*, *33*, 261-304.

Butler, D. L., & Wine, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, 65, 245-281.

Campanario, J. M. (2002). The parallelism between scientists' and students' resistance to new scientific ideas. *International Journal of Science Education*, *24*(10), 1095-1110.

Chi, M. T. H. (2005). Common sense conceptions of emergent processes: Why some misconceptions are robust. *Journal of the Learning Sciences, 14*, 161–199.

Chi, M. T. H., & Brem, S. K. (2009). Contrasting Ohlsson's Resubsumption Theory with Chi's Categorical Shift Theory. *Educational Psychologist, 44,* 58-63.

Cosejo, D. G., Oesterreich, J., & Ohlsson, S. (2009). Re-categorization: Restructuring in categorization In N.A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

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

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.

Goldberg, R. F., & Thompson-Schill, S. L. (2009). Developmental "roots" in mature biological knowledge. *Psychological Science*, *20*(4), 480-487.

Gopnik, A., & Wellman, H. M. (2012, May 14). Reconstructing Constructivism: Causal Models, Bayesian Learning Mechanisms, and the Theory Theory. *Psychological Bulletin.* Advance online publication. doi: 10.1037/a0028044

Halloun, I. A., & Hestenes, D. (1985). Common sense concepts about motion.*American journal of physics*, *53*(11), 1056-1065.

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.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2013). *lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package).*

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.

Mirman, D. (2014). *Growth Curve Analysis and Visualization Using R*. Chapman and Hall / CRC.

Ohlsson, S. (2008). Computational models of skill acquisition. In R. Sun (Ed.), *The Cambridge Handbook of Computational Psychology* (pp. 359-395). Cambridge, UK: Cambridge University Press.

Ohlsson, S. (2009). Resubsumption: A possible mechanism for conceptual change and belief revision. *Educational Psychologist, 44,* 20– 40.

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, G. K. S., Haider, H., & Rhenius, D. (1999). Constraint Relaxation and Chunk Decomposition in Insight Problem Solving.

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.*

R Core Team (2014). R: *A language and environment for statistical computing. R Foundation for Statistical Computing,* Vienna, Austria. URL http://www.R-project.org/.

Rusanen, A. M., & Lappi, O. (2013). What, when and how do the models of conceptual change explain. In *Proceedings of the 35th annual conference of the cognitive science society* (pp. 3331-3336).

SAS Technical Report R-101 (1978) *Tests of Hypotheses in Fixed-Effects Linear Models* Copyright (C)(SAS Institute Inc., Cary, NC, USA)

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78(1),* 153-189.

Strike, K.A., & Posner, G.J. (1982). Conceptual change and science teaching. *European Journal of Science Education, 4,* 231-240.

Strike, K.A., & Posner, G.J. (1992). A revisionist theory of conceptual change. In R. Duschl & R. Hamilton (eds.), *Philosophy of science, cognitive psychology, and educational theory and practic*e (pp. 147-176). Albany, NY: SUNY Press.

Vosniadou, S., & Brewer, W.F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology*, *24*, 535-585.