Title

Cobweb: An Incremental and Hierarchical Model of Human-Like Category Learning

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)

Authors

Lian, Xin
Varma, Sashank
MacLellan, Christopher

Publication Date

2024

Peer reviewed

Cobweb: An Incremental and Hierarchical Model of Human-Like Category Learning

Xin Lian (xinlian@gatech.edu)
Georgia Institute of Technology
Atlanta, GA 30308 USA

Sashank Varma (varma@gatech.edu)
Georgia Institute of Technology
Atlanta, GA 30308 USA

Christopher J. MacLellan (cmaclell@gatech.edu)
Georgia Institute of Technology
Atlanta, GA 30308 USA

Abstract

Cobweb, a human-like category learning system, differs from most cognitive science models in incrementally constructing hierarchically organized tree-like structures guided by the category utility measure. Prior studies have shown that Cobweb can capture psychological effects such as basic-level, typicality, and fan effects. However, a broader evaluation of Cobweb as a model of human categorization remains lacking. The current study addresses this gap. It establishes Cobweb's alignment with classical human category learning effects. It also explores Cobweb's flexibility to exhibit both exemplar- and prototype-like learning within a single framework. These findings set the stage for further research on Cobweb as a robust model of human category learning.

Keywords: categorization, concept learning, prototypes, exemplars

Introduction

Learning a *category* (or *concept*) involves inferring its structure from a set of examples (T. L. Griffiths, Sanborn, Canini, & Navarro, 2008). Various computational models of concept learning have been proposed under a number of theoretical frameworks. Some are *rule-based*, suggesting that concepts are represented as rules formulated within a compositional representation language (Kemp, 2012). These include RULEX (Nosofsky, Palmeri, & McKinley, 1994) which generates conjunctive rules and retains their exceptions, the Bayesian description of rules (Goodman, Tenenbaum, Feldman, & Griffiths, 2008), integrated mental models (Goodwin & Johnson-Laird, 2011), and the algebra of concept learning (Feldman, 2006). Other approaches are *similarity-based*, including exemplar- and prototype-based models. Representative examples are ALCOVE (Kruschke, 1992) and SUSTAIN (Love, Medin, & Gureckis, 2004), both of which are exemplar models representing concepts with the weights of connectionist networks.

Within similarity-based models, a subset involves models that employ rational analysis to learn concepts (Anderson & Matessa, 1990; Nosofsky, 1998; Ashby & Alfonso-Reese, 1995; Rosseel, 2002). These models support *incremental learning and updating* of acquired knowledge. A limitation of many rational categorization models (Anderson & Matessa, 1990; Sanborn, Griffiths, & Navarro, 2006; T. Griffiths, Canini, Sanborn, & Navarro, 2007) is that they predominantly propose flat partitions. Such representations might not fully capture important psychological effects such as the typicality effect, specifically the processing of atypical instances.

It is therefore interesting to consider *Cobweb* (Fisher, 1987), which learns concepts incrementally and hierarchically, organizing cognitive structures into hierarchical levels of partitions, making it a potentially powerful model of human-like category learning (Langley, 2022). Cobweb has a long history in artificial intelligence and is noteworthy for its incremental learning capabilities. Fisher and Langley (1990) demonstrates its ability to explain various psychological effects including basic-level, typicality, and fan effects. However, beyond these initial efforts, Cobweb remains underexplored as a model of human categorization.

In this paper, we further evaluate Cobweb's potential as a model of human category learning. We assess the alignment between predictions made from two different levels of its hierarchy (subordinate leaves and basic concepts) and the classical findings of Medin and Schaffer (1978) and Shepard, Hovland, and Jenkins (1961). Our work shows Cobweb's efficacy in accounting for human categorizations at a general level. Importantly, we observe that Cobweb does not rigidly adhere to prototype- or exemplar-like behavior. This flexibility arises from its hierarchical cognitive structure, which enables the generation of predictions that range from prototype- to exemplar-like. We demonstrate Cobweb's proficiency as an incremental learner and thus its ability to account for human category learning over multiple training blocks. Finally, we illustrate Cobweb's robust alignment with human categorization across various tasks. The versatility displayed by Cobweb underscores its potential as a comprehensive model of human categorization.

Cobweb: Human-Like Categorization

Cobweb (Fisher, 1987) takes an incremental and hierarchical approach to learning. The system forms concepts given sequentially presented instances, which are represented as discrete attribute-value pairs (e.g., `color: blue; form: triangle; size: large; number: 2`). Given such examples, Cobweb forms a probabilistic concept hierarchy. Each concept node in the hierarchy stores a probability table that tracks the frequency of each attribute value occurring in instances of the concept. The left panel of Figure 1 shows examples of Cobweb's instance and concept representations. To guide the concept formation process, Cobweb uses *Category Utility* (*CU*), which was proposed by Corter and Gluck (1992) to account for human categorization effects. This mea-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

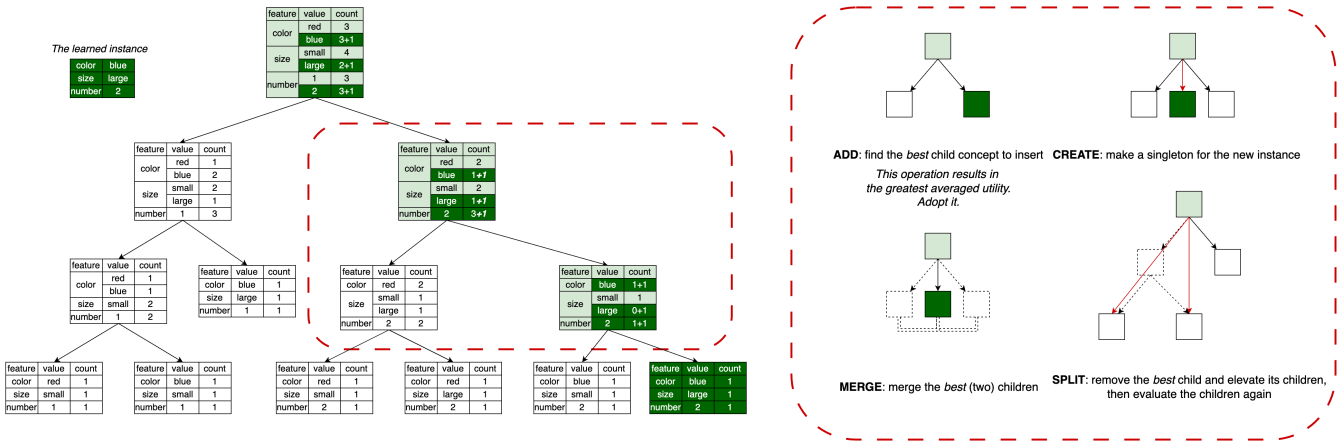

Figure 1: An illustrative example of Cobweb's learning process which involves incorporating a new instance, depicted as a dark green table, into the existing tree structure. Cobweb traverses the tree from the root to a specific leaf node, and along this path, the concept nodes (highlighted in light green) are updated to reflect the given instance. The changes resulting from fitting the new instance into the tree are denoted in bold and italics. During this traversal, Cobweb considers four available operations at each branch, *adding*, *creating*, *merging*, and *splitting*. It then proceeds with the operation that yields the highest averaged category utility within the subtree. For instance, for the branch in the red dot box, Cobweb chooses to add the instance to the "best" child because it results in the highest average utility score.

sure, which we define below, evaluates the feature predictive power of a concept C_k.

When a new instance $\boldsymbol{x}$ is introduced to be learned, Cobweb sorts it down its current categorization tree. At each node, Cobweb considers how best to incorporate the instance into the current node's children $\{C_k\}$. It evaluates four operations - *add*, *create*, *merge*, or *split* - and chooses the one that produces the highest averaged category utility:

$$\frac{\sum_{k=1}^{s} CU(C_k)}{s} \tag{1}$$

where s is the number of children at the current branch. The average category utility measure lets Cobweb compare partitions with varying numbers of concepts (Fisher, 1987).

To evaluate the operations, Cobweb starts by simulating *adding* the instance into each of the children concepts. When an instance is added to a concept, its probability table frequencies are updated to reflect the instance's attribute values. Cobweb uses these probability tables to efficiently compute the average category utility score without having to iterate over prior instances. After considering each addition, Cobweb identifies the two children C_k^1 and C_k^2 that yield the highest and second-highest average category utility.

Next, the system evaluates *merging*, *splitting*, and *creating*. To evaluate merging, Cobweb simulates the creation of a new child concept that merges the probability table counts of C_k^1 and C_k^2 (these concepts become children of the new concept) and updates the result to reflect the addition of the instance. To evaluate splitting, it simulates removing C_k^1 and promoting its children to the current level. Finally, Cobweb considers creating a new concept that reflects the instance.

After evaluating all the operations, Cobweb chooses the operation that yields the maximal average category utility. When it elects to add or merge, the tree is updated to reflect these operations and the entire process repeats recursively at the updated node (C_k^1 or the new merged node). When splitting is selected, Cobweb removes the split node, promotes the children, and then recursively repeats its evaluation at the current node. Finally, when Cobweb creates a new node, the process terminates. Figure 1 illustrates the process of Cobweb learning a new instance.

After the learning phase and the construction of the tree, Cobweb can apply its learned structure to predict the values of unobserved features of a given instance $\boldsymbol{x}^*$ with unknown feature(s). This prediction process is similar to the learning phase: the instance traverses down the tree from the root to a leaf through iterative simulations of insertion into the "best" child node within each subtree. This traversal results in a path of category nodes visited, akin to those depicted in Figure 1. However, unlike the learning process, none of the concept frequency tables are updated to reflect $\boldsymbol{x}^*$ and restructuring operations (merging, splitting, and new) are not considered. Subsequently, Cobweb predicts the unobserved feature values by using a specific node along the categorization path. The analysis of Corter and Gluck (1992) suggests the *basic-level* node, which holds the highest category utility value, should be utilized for inference. However, recent studies with Cobweb (MacLellan, Harpstead, Aleven, Koedinger, et al., 2016; MacLellan, Matsakis, & Langley, 2022; MacLellan & Thakur, 2022) have favored using the leaf node, claiming that it often yields superior predictive performance.

Cobweb is an example of a *human-like learning* system, meeting the computational "gauntlets" outlined by Langley (2022). Further, Fisher and Langley (1990) showed how Cob-

web can account for various human concept learning effects, including *basic-level* (Murphy & Smith, 1982; Hoffmann & Ziessler, 1983), *typicality* (Rosch & Mervis, 1975), and *fan* effects (Anderson, 1974). Their simulations and analyses predicted human response times in psychological studies using the notion of *category match*. This is a variation of category utility that measures the features present in a given classified observation. A notable feature of Cobweb is its direct utilization of hierarchical structure. This distinguishes it from many cognitive science and artificial intelligence models that partition observations into flat clusterings labeled by external categories. This hierarchical approach lets Cobweb capture different categorization effects on basic-level and subordinate concepts: while typical objects tend to be classified into basic-level concepts, atypical objects within a category can be classified into subordinate categories instead because of their low intra-category and high inter-category overlaps.

Despite these desirable features of Cobweb (Langley, 2022), its cognitive plausibility and potential to capture psychological effects within human categorization have been unexplored. Can Cobweb account for the human categorization data that other categorization models in cognitive science handle? Does it exhibit more prototype- or exemplar-like categorization behavior, or is it more of a hybrid model? Do the varying concept levels within the Cobweb tree make Cobweb a more flexible categorization model? This paper addresses these questions through several computational experiments focused on seminal cognitive science studies.

Experiments

To evaluate Cobweb's alignment to further aspects of human concept learning, we conducted computational experiments using the empirical paradigms developed by Medin and Schaffer (1978) and Shepard et al. (1961).

Learning and Predicting

Our experiments utilize the implementation of Cobweb developed by MacLellan et al. (2016)[1]. This implementation employs the *information-theoretic* variant of category utility (Corter & Gluck, 1992) for learning and prediction. Given a category c, its *uncertainty* (or *entropy*) is given by

$$U(c) = \sum_i P(X_i|c)U(X_i|c) \tag{2}$$

where

$$U(X_i|c) = -\sum_j P(x_{ij}|c)\log P(x_{ij}|c) \tag{3}$$

is the uncertainty of the feature X_i given the concept c. Here $P(x_{ij}|c)$ is the probability that feature X_i has value x_{ij} given c. The information-theoretic category utility is then defined as:

$$CU(c) = P(c)[U(c_p) - U(c)] \tag{4}$$

where c_p is the parent concept of c. This measure captures the informativeness of the category in terms of the expected reduction in feature value uncertainty given knowledge of the child category label over knowledge of the parent label.

Once the Cobweb tree structure is induced using training stimuli, the categorization of a test stimulus $\boldsymbol{x}$ with unobserved feature(s) occurs along a path from its root to a leaf node, defining a concept path. To determine the category for predicting the unobserved feature values, we explore two levels of the hierarchy:

`leaf` Cobweb predicts the unobserved features based on the data stored in the leaf node, which is at the subordinate level. This prediction tends to be deterministic due to the certainty of feature values stored at the lowest level. For instance, consider the leaves depicted in Figure 1, where all attributes have specific values with 100% certainty.

`basic` Cobweb predicts the unobserved features based on the data stored at the basic level, which is the node along the categorization path with the highest category utility. In general, these nodes are more superordinate than leaf nodes.

The Medin and Schaffer (1978) Experiments

Dataset and Original Study Medin and Schaffer (1978) proposed the exemplar (i.e., context) model of classification and evaluated it in an artificial category learning experiment with two sets of 16 stimuli that differ on four binary dimensions, and in particular, *color* {*red, blue*}, *form* {*triangle, circle*}, *size of each component* {*large, small*}, and *number of components* {*1, 2*} for geometric stimuli. Stimuli 4, 5, 7, 13, and 15 composed the training stimuli for Category A, and stimuli 2, 10, 12, and 14 the training stimuli for Category B. The remaining Stimuli (i.e., 1, 3, 6, 8, 9, 11, 16) were the transfer stimuli. Participants initially learned the nine training stimuli and were provided feedback. After engaging in an interpolated activity, they classified all 16 training and transfer stimuli, this time without feedback. Medin and Schaffer (1978) compared the predicted probabilities generated by the exemplar model for Stimulus 4 and 7. The purpose was to infer whether people learned more prototype-like or more exemplar-like representations. A higher predicted Category A probability for Stimulus 4 would suggest a closer alignment with the prototype model due to its greater resemblance to the prototype stimulus for Category A, Stimulus 1 (i.e., (1, 1, 1, 1)). By contrast, a higher predicted probability for Stimulus 7 would indicate greater alignment with exemplar representations as it is more similar to the individual stimuli of Category A than the individual stimuli of Category B.

Method and Hypothesis The process starts by training Cobweb with the 9 designated training stimuli, then obtaining predicted probabilities for all 16 training and transfer stimuli. We compare these probabilities with the human classification probabilities from the original study using the Pearson correlation coefficient and root mean squared deviation (RMSD) to quantify the relative and absolute fit, respectively.

[1]The codes for the experiments are available at `https://github.com/Teachable-AI-Lab/cobweb-psych`

Table 1: The left panel shows the observed classification probabilities from human subjects in the Medin and Schaffer (1978) study when using geometric stimuli and the predicted classification probabilities of Cobweb at two levels, `leaf` and `basic`, on the stimulus's respective classifications. For instance, the classification probability of Stimulus 4(A) is the probability that Stimulus 4 is classified as Category A, and the one of Stimulus 2(B) is the probability that Stimuli 2 is classified as Category B. To facilitate a direct comparison of the classification probabilities of Stimuli 4(A) and 7(A), we indicate them in shaded rows and denote the stimulus with the greater classification probability using **bold** text. The right columns show the sample standard deviation of the predicted probability of each stimulus at either the leaf or basic level.

Stimulus	Mean Probability			Sample SD	
	human	**leaf**	**basic**	**leaf**	**basic**
Training Stimuli					
4A	0.780	0.735	**0.837**	0.061	0.161
7A	**0.880**	**0.750**	0.826	0.000	0.196
15A	0.810	0.750	0.893	0.000	0.048
13A	0.880	0.750	0.854	0.000	0.114
5A	0.810	0.750	0.796	0.000	0.204
12B	0.840	0.695	0.664	0.154	0.300
2B	0.840	0.723	0.751	0.109	0.233
14B	0.880	0.750	0.839	0.000	0.156
10B	0.970	0.750	0.867	0.000	0.078
New Transfer Stimuli					
1A	0.590	0.685	0.784	0.163	0.234
6A	0.940	0.750	0.885	0.000	0.069
9A	0.500	0.350	0.220	0.206	0.239
11A	0.620	0.675	0.724	0.166	0.286
3B	0.690	0.605	0.651	0.226	0.323
8B	0.660	0.650	0.721	0.201	0.282
16B	0.840	0.750	0.828	0.000	0.146

We derive predictions using two methods, `leaf` and `basic`. To ensure the robustness and reliability of experimental outcomes, we conduct the experiments using 40 different random seeds when randomizing the stimuli learning order, and each seed is associated with 5 repeated implementations (so each stimulus is predicted 200 times). This handles stochasticity introduced because, in cases of tied expected category utility, Cobweb randomly selects the best next operation.

We expect Cobweb to exhibit a strong alignment with human data with both prediction methods. Furthermore, we have no *a priori* expectation that Cobweb will strictly adhere to either a prototype or exemplar model, so Stimuli 4 and 7 may exhibit less striking differences in predicted probability than was observed in the human data. In fact, we expect a slightly higher predicted probability for Stimulus 4, which would suggest a more prototype-like categorization by Cobweb. These expectations are rooted in the idea that Cobweb builds up a hierarchical cognitive structure of concepts, generating predictions based on a specific concept node with more or less integrated information.

Results and Discussions The observed and predicted classification probabilities for each stimulus with their sample standard deviations are listed in Table 1. The corresponding correlation coefficients and RMSD values are presented in Table 2. Considering the predicted probabilities for Stimuli 4 and 7 in Table 1, Cobweb exhibits a slightly higher probability for Stimulus 7 with the `leaf` level prediction (0.735 vs. 0.750), but a slightly higher probability for Stimulus 4 with the `basic` level prediction (0.837 vs. 0.826). Although the differences here are not very significant, they are less likely to be affected by the variance of the predicted probabilities given relatively small sample standard deviations for predicted probabilities at both levels with a sample size of 200 each. This pattern shows that Cobweb does not strictly adhere to a prototype- or exemplar-like categorization model paradigm. Instead, it appears to exhibit aspects of both. This capability aligns with the insights from Fisher and Langley (1990), particularly the typicality effects observed in Cobweb: the distributed categorization strategy employed by Cobweb allows atypical objects to be categorized into subordinate-level concepts because of low intracategory and high intercategory overlaps.

Table 2: The correlation coefficients and RMSD values between the predicted and observed classification probabilities (shown in Table 1) for the geometric stimuli in the Medin and Schaffer (1978) experiment.

Stimuli Set	**leaf**	**basic**
Correlation	0.768	0.713
RMSD	0.166	0.130

By examining the correlation scores and the corresponding RMSD values compared with predicted and observed human probabilities, both prediction levels (`leaf` and `basic`) employed by Cobweb result in a strong correlation and a modest amount of absolute error with the human data, demonstrating alignment with human concept learning.

The Shepard et al. (1961) Experiments

Dataset and Original Study In the original Shepard et al. (1961) study, there are 8 stimuli and they differ on 3 binary dimensions - *size* {*small, large*}, *color* {*white, black*}, and *form* {*square, triangle*}. Over these, six category structures I-VI are defined, wherein each category A and B span 4 stimuli each. Each structure is defined by a logical rule of increasing complexity distinguishing A from B. Type I concerns a single diagnostic dimension: the stimuli in each category just differ in color. Type II, the correlated-features task, instantiates the XOR problem along two of the dimensions. Tasks III and V are rule-plus-exception tasks, with the rule leaving an exception item that requires an additional conjunct to represent, and thus presumably additional cognitive processing to learn. Task IV mainly concerns the family resemblance—the prototypes of each category (a large black triangle for Category A and a small white square for Category B) are joined by the stimuli that share two of three features with their prototype. Type VI is arbitrary and the stimuli of each category

are special cases sharing no common structure.

Smith, Minda, and Washburn (2004) replicated the original Shepard et al. (1961) study using a more comprehensive approach to increase the robustness of the results. Each task type encompassed 6 possible stimuli arrangements (permutations), resulting in 36 distinct tasks. Each human participant engaged in six tasks, each randomly selected from the available permutations within the respective task type. For each task, participants underwent a learning phase lasting 24 blocks (or iterations), on each of which they saw all eight stimuli, equating to $24 \times 8 = 192$ trials overall.

In the original study, Shepard et al. (1961) concluded that humans do not simply follow behaviorist laws of conditioning and stimulus generalization, and instead "abstract dimensions [and] then formulate and test rules about how the values on those dimensions combine and interact to determine which classificatory response will be correct" (p. 33).

Method and Hypothesis In the replication of Smith et al. (2004), each task type is instantiated as six tasks (so there are 36 tasks in total), and each task is a permutation of the eight stimuli among two categories such that they satisfy the rule specified by the task type. We ran Cobweb on each of the 36 tasks, repeating each five times with different random seeds. For each task repetition, after training on all stimuli, Cobweb was used to predict the category of the eight trained stimuli, and we computed an average accuracy score based on these predictions.

We compared observed human and model-predicted accuracies separately on each of the six task types. For each task in its task type, Cobweb was trained across 24 blocks, each running through the randomly ordered 8 stimuli. After each block, the average accuracy over the 8 stimuli was computed. These results are averaged across the six tasks for each type to produce an average learning curve across the 24 blocks. Finally, these model-predicted learning curves are compared to the human learning curves from Smith et al. (2004), and correlation coefficients and RMSD values quantify their correspondence. Note that Smith et al. (2004) provide human accuracies for only the odd learning blocks (1, 3, 5, ..., 19, 21, 23), and so performance on these blocks is the basis of the comparison between the human and Cobweb learning curves.

We expect Cobweb to show strong alignment with human learning across the six task types: to learn the simpler Type I and II structures more rapidly and to a higher accuracy level than the more complex Type V and VI structures. We also conjectured the `leaf` predictions may be less comparable to the human learning data: Because the leaf nodes contain the homogeneous feature values only, the predictions made by these nodes are always the same. Thus, their predictions are "overly" deterministic compared to human predictions, which might make their learning curves artifactually resemble a horizontal line.

Results and Discussion Figure 2 shows the learning curves for each of the six task types – both the observed human data and the two predictor levels of Cobweb. The corresponding correlation coefficients and RMSD values are presented in Table 3. Overall, Cobweb shows promising alignment for most task types and for both prediction levels.

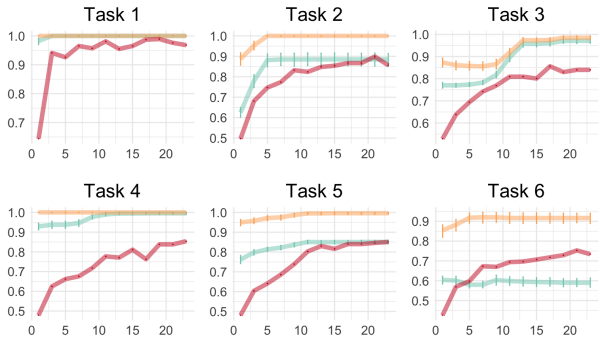

Figure 2: The learning curves for human participants (***red***), the `leaf` prediction level (***orange***), and the `basic` prediction level (***light green***) across the learning blocks $1-23$. Learning block is on the x-axis and (human and model) accuracy on the y-axis.

Table 3: Correlation coefficients and RMSD values with the Shepard et al. (1961) category structures. The values are computed by comparing the accuracy scores from human participants of the Smith et al. (2004) replication and Cobweb's `leaf` or `basic` prediction levels, respectively, across the learning blocks $1, 3, ..., 21, 23$. N/A indicates that the correlation coefficient cannot be computed because one compared set of data (the accuracy score set of Cobweb among all 11 training blocks) is constant.

Level	I	II	III	IV	V	VI
Correlation						
`leaf`	N/A	0.929	0.751	N/A	0.984	0.884
`basic`	0.981	0.932	0.841	0.910	0.984	-0.375
RMSD						
`leaf`	0.110	0.206	0.153	0.231	0.210	0.145
`basic`	0.105	0.075	0.112	0.198	0.089	0.230

In Task I (diagnostic task), both humans and Cobweb learn the categories rapidly and accurately, and the `leaf` prediction even achieves 100% accuracy after the first learning block, resulting in a horizontal learning curve across blocks. The `basic` learning curve has a high correlation with human data ($r = 0.981$). A similar outcome is observed in Task IV (family resemblance). The `leaf` prediction achieves perfect accuracy, making it challenging to compute its correlation coefficient with human performance. The `basic` predictions also exhibit a strong correlation with humans ($r = 0.910$).

For Task II (XOR), both prediction levels achieve high correlations with the human data ($r = 0.929$ and 0.932). For Type V, the correlations are again high and comparable ($r = 0.984$ for both levels). However, for Type III, another rule-plus-exception task type, the correlation coefficients are

lower compared to Task V: $r = 0.751$ for `leaf` and 0.841 for `basic`. One possible explanation for this difference is that Task V involves an exception for a single rule, whereas Task III involves an exception for two rules. This more complex scenario results in Cobweb requiring more blocks before a rapid accuracy boost. Finally, in Task VI (no family resemblance), `leaf` predictions are again strongly aligned with human performance, and overall accuracy remains high at around 0.90 (though it is less than for the other, simpler types). However, `basic` predictions underperform for Task VI, both in terms of human alignment and overall accuracy. Recall that this is the "chaotic" conceptualization, i.e., there is no distinct typical or atypical stimulus for each category. Consequently, neither the basic-level concepts nor the subordinate concepts perform well.

Note that, although Cobweb exhibits promising alignment with the human data at both prediction levels, as shown by correlation coefficients, the RMSD values diverge from 0 across most task types at both levels. Making more accurate absolute predictions is a challenge for the future development of Coweb as a model of human categorization.

Finally, we explored Cobweb's ability to predict the relative difficulty of the six task types I-VI for the humans in the Shepard et al. (1961) replication of Smith et al. (2004). Table 4 provides the observed human data after the first (1) and final (23) learning blocks. Note that there is some stability over learning, with Type I as the easiest and Type VI the hardest in both blocks. The predicted difficulty rankings of the task types for those two blocks is also shown in the table. The alignment is promising, with the basic method ranking on the first learning block and the leaf method ranking on the final learning block agreeing with the human rankings well.

Table 4: Comparison of the relative difficulty of the six task types I-VI after the first (1) and last (2) learning blocks by both humans (Smith et al., 2004) and Cobweb (leaf, basic-level nodes). Human rankings are highlighted with shaded rows, with matching task types indicated in **bold** text at corresponding ranking positions. The Spearman's rank correlation coefficients ρ between human and predicted rankings are in the right column, where tied ranks are considered.

Ranking	1	2	3	4	5	6	ρ
Block 1							
Observed	I	III	II	V	IV	VI	
`leaf`	**I**	IV	V	II	III	**VI**	0.386
`basic`	**I**	IV	III	**V**	II	**VI**	0.829
Block 23							
Observed	I	II	IV	V	III	VI	
`leaf`	**I**	**II**	**IV**	**V**	**III**	**VI**	0.857
`basic`	**I**	IV	III	II	V	**VI**	0.486

Discussion

This paper has evaluated the alignment of *Cobweb*, a classical AI model of incremental concept learning, against data from two seminal cognitive science experiments, Medin and Schaffer (1978) and Shepard et al. (1961). The promising alignment between human performance and Cobweb's predictions demonstrates its viability as a cognitive science model, adding to the evidence provided by an earlier evaluation (Fisher & Langley, 1990).

The hierarchical structure of Cobweb enables it to generate predictions at different levels. Here, we derive categorization predictions at two levels: the `leaf` (i.e., subordinate) level and the `basic` level. A notable feature is that the flexibility of Cobweb is that it can span the spectrum between prototype-like and exemplar-like representations. This flexibility may enable it to account for the transition from prototype representations early in concept acquisition to exemplar representations after extended learning (Smith & Minda, 1998).

These findings are a first step in demonstrating Cobweb's potential as a model of human categorization. It is important to note that in the experiment by Medin and Schaffer (1978), our comparison between Cobweb's predictions and observations was limited to geometric stimuli, whereas the original study covered two sets of stimuli (geometric stimuli and Brunswik faces), both share nominal representations which are simplified and artificially constructed. Indeed, many prior studies comparing exemplar and prototype models have utilized highly simplified perceptual stimuli and artificially designed category structures rather than more natural stimuli and natural category domains. One limitation of this approach is that participants typically have extensive prior experience with the categories being tested, and this learning history is not controlled in experiments (Nosofsky, Meagher, & Kumar, 2022). To bridge this gap and better understand categorization in more natural settings with more complex and high-dimensional stimuli, Battleday, Peterson, and Griffiths (2020) employed various machine learning methods on a large behavior dataset featuring natural images. Moving forward, our experiments can be extended using Cobweb/4V (Barari, Lian, & MacLellan, 2024), a derivative of Cobweb that incorporates image representations instead of low-dimensional artificial ones. This could let Cobweb account for categorization effects in studies with natural images.

Future research should also expanded the evaluation of Cobweb to other important findings on human categorization. Building on Shepard et al. (1961), can Cobweb also account for studies of structured concepts (Feldman, 2000, 2003; Hayes-Roth & Hayes-Roth, 1977)? Building on Medin and Schaffer (1978), can it account for studies of linear separability (Medin & Schwanenflugel, 1981; Levering, Conaway, & Kurtz, 2020) and correlated features (Malt & Smith, 1984; Medin, Altom, Edelson, & Freko, 1982)? And finally, what is its relationship to other "hybrid" models like RULEX (Nosofsky et al., 1994) and SUSTAIN (Love et al., 2004)?

In conclusion, we provide preliminary evidence that Cobweb can flexibly model human category learning—exhibiting both exemplar- and prototype-like behavior. We look forward to evaluating it across more study data and developing it into a robust model of human category learning.

References

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive psychology*, *6*(4), 451–474.

Anderson, J. R., & Matessa, M. (1990). A rational analysis of categorization. In *Machine learning proceedings 1990* (pp. 76–84). Elsevier.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of mathematical psychology*, *39*(2), 216–233.

Barari, N., Lian, X., & MacLellan, C. J. (2024). Avoiding catastrophic forgetting in visual classification using human concept formation. *arXiv preprint arXiv:2402.16933*.

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2020). Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature communications*, *11*(1), 5418.

Corter, J. E., & Gluck, M. A. (1992). Explaining basic categories: Feature predictability and information. *Psychological bulletin*, *111*(2), 291.

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, *407*(6804), 630–633.

Feldman, J. (2003). A catalog of boolean concepts. *Journal of Mathematical Psychology*, *47*(1), 75–89.

Feldman, J. (2006). An algebra of human concept learning. *Journal of mathematical psychology*, *50*(4), 339–368.

Fisher, D. H. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine learning*, *2*, 139–172.

Fisher, D. H., & Langley, P. (1990). The structure and formation of natural categories. *Psychology of Learning and Motivation*, *26*, 241–284.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive science*, *32*(1), 108–154.

Goodwin, G. P., & Johnson-Laird, P. (2011). Mental models of boolean concepts. *Cognitive psychology*, *63*(1), 34–59.

Griffiths, T., Canini, K., Sanborn, A., & Navarro, D. (2007). Unifying rational models of categorization via the hierarchical dirichlet process.

Griffiths, T. L., Sanborn, A. N., Canini, K. R., & Navarro, D. J. (2008). Categorization as nonparametric bayesian density estimation. *The probabilistic mind: Prospects for Bayesian cognitive science*, 303–328.

Hayes-Roth, B., & Hayes-Roth, F. (1977). Concept learning and the recognition and classification of exemplars. *Journal of Verbal Learning and Verbal Behavior*, *16*(3), 321–338.

Hoffmann, J., & Ziessler, C. (1983). Objektidentifikation in künstlichen begriffshierarchien. *Zeitschrift für Psychologie mit Zeitschrift für angewandte Psychologie*.

Kemp, C. (2012). Exploring the conceptual universe. *Psychological review*, *119*(4), 685.

Kruschke, J. K. (1992). Alcove: An exemplar-based connectionist model of category learning. In *Connectionist psychology* (pp. 107–138). Psychology Press.

Langley, P. (2022). The computational gauntlet of human-like learning. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 36, pp. 12268–12273).

Levering, K. R., Conaway, N., & Kurtz, K. J. (2020). Revisiting the linear separability constraint: New implications for theories of human category learning. *Memory & cognition*, *48*, 335–347.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). Sustain: a network model of category learning. *Psychological review*, *111*(2), 309.

MacLellan, C. J., Harpstead, E., Aleven, V., Koedinger, K. R., et al. (2016). Trestle: a model of concept formation in structured domains. *Advances in Cognitive Systems*, *4*, 131–150.

MacLellan, C. J., Matsakis, P., & Langley, P. (2022). Efficient induction of language models via probabilistic concept formation. *arXiv preprint arXiv:2212.11937*.

MacLellan, C. J., & Thakur, H. (2022). Convolutional cobweb: A model of incremental learning from 2d images. *arXiv preprint arXiv:2201.06740*.

Malt, B. C., & Smith, E. E. (1984). Correlated properties in natural categories. *Journal of verbal learning and verbal behavior*, *23*(2), 250–269.

Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*(1), 37.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological review*, *85*(3), 207.

Medin, D. L., & Schwanenflugel, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(5), 355.

Murphy, G. L., & Smith, E. E. (1982). Basic-level superiority in picture categorization. *Journal of verbal learning and verbal behavior*, *21*(1), 1–20.

Nosofsky, R. M. (1998). Optimal performance and exemplar models of classification. *Rational models of cognition*, 218–247.

Nosofsky, R. M., Meagher, B. J., & Kumar, P. (2022). Contrasting exemplar and prototype models in a natural-science category domain. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(12), 1970.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological review*, *101*(1), 53.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, *7*(4), 573–605.

Rosseel, Y. (2002). Mixture models of categorization. *Journal of Mathematical Psychology*, *46*(2), 178–210.

Sanborn, A., Griffiths, T., & Navarro, D. (2006). A more rational model of categorization.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological monographs: General and applied*, *75*(13), 1.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, memory, and cognition*, *24*(6), 1411.

Smith, J. D., Minda, J. P., & Washburn, D. A. (2004). Category learning in rhesus monkeys: a study of the shepard, hovland, and jenkins (1961) tasks. *Journal of Experimental Psychology: General*, *133*(3), 398.