CALM - A Process Model of Category Generalization, Abstraction and Structuring

René Schlegelmilch (r.schlegelmilch@psychologie.uzh.ch)
Department of Cognitive Psychology, Binzmuehlestrasse 14
Zürich, 8050 Switzerland

Andy J. Wills (andy.wills@plymouth.ac.uk)
School of Psychology, Plymouth University
Plymouth, PL4 8AA UK

Bettina von Helversen (b.vonhelversen@psychologie.uzh.ch)
Department of Cognitive Psychology, Binzmuehlestrasse 14
Zürich, 8050 Switzerland

Abstract

In this paper, we introduce CALM, a process model that is designed to abstract solutions in simple and complex category learning tasks. The model includes strong assumptions about the interaction of processes driving learning behavior, typically addressed in terms of feature attention, stimulus generalization, rule abstraction and knowledge partitioning. We present simulations of CALM, showing that the model can account both for systematic variations in Type II category difficulty, and for individual differences in extrapolation of an XOR category structure.

Keywords: category learning; process model; associative learning; abstraction; problem structuring; decision making

Introduction

In their pioneering work on category learning, Shepard, Hovland, and Jenkins (1961) provided a paradigm that became a benchmark for models that implement assumptions on the processes underlying category learning. In their tasks, participants learned to categorize stimuli with three binary features, i.e. colour (black vs. white), shape (square vs. triangle), and size (small vs. large). They learned six category structures varying in difficulty (problem Types I-VI, Fig. 1A). The authors found that the learning curves (speed and accuracy) systematically differed between the problems, such that I > II > [III, IV, V] > VI (see also Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994). These findings, especially the quick learning in Types I and II, can be predicted by existing category learning models, e.g. ALCOVE (Kruschke, 1992), assuming that people generalize from instance memory while focusing their attention on dimensions that reduce error.

However, the empirical findings regarding the performance in the Type II problem have been recently revised (Kurtz, Levering, Stanton, Romero, & Morris, 2013), challenging the current explanatory accounts. Kurtz et al. show that if the 'use of rules' is instructed beforehand, the classic findings hold. Without rule instructions, however, Type II learning curves fall together with III, IV, and V, without affecting the remaining pattern. Importantly, the overall decrease in Type II performance in the absence of rule instructions is an aggregate effect seemingly summarizing a bimodal distribution of learning success, i.e. without rule instructions some participants perform considerably worse in Type II than e.g. in Type IV, while other participants perform considerably better in Type II than in Type IV. Kurtz et al. (2013) discuss that ALCOVE (Kruschke, 1992) could predict this pattern if systematic population differences in attention learning were assumed. However, with such an assumption ALCOVE would also predict that Type I is learned more slowly than it is observed, raising unanswered theoretical questions about attention learning (see also Matsuka, & Corter, 2008). Kurtz et al. therefore suggest that these performance differences in Type II might be driven by other mechanisms, including the abstraction of 'rule-like' representations, as it was initially suggested by Shepard, Hovland and Jenkins (1961), as well.

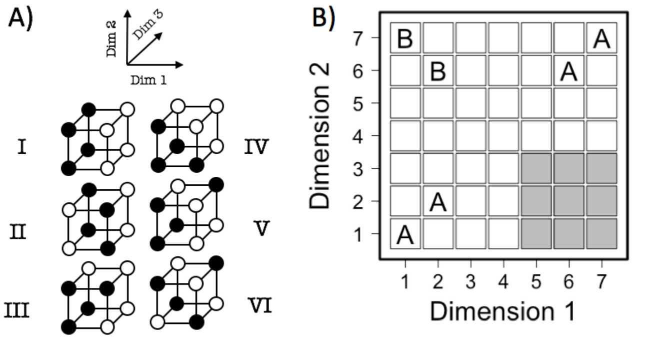

Figure 1: (A) Classic category structures Type I-VI (Shepard, Hovland, & Jenkins, 1961). Coordinates represent stimuli with three binary dimensions; black and white circles indicate categories. (B) Coordinate grid of the incomplete XOR structure as trained in Conaway & Kurtz (2017); 'A' and 'B' refer to categories of trained stimuli. Shaded cells refer to the extrapolation area for category 'B'.

Interestingly, the Type II problem could be solved in terms of a structured problem (see also Kalish, Lewandowsky, & Kruschke, 2004), i.e. as one-dimensional rule, e.g. "black objects belong to category B, and white objects to category A", that is inverted conditional on the values of a second dimension (context cue), e.g. "invert the prediction for small objects, but not for large objects". This we call contextual modulation.

In a recent study, the Type II task, also called 'Exclusive - Or' problem (XOR), was extended to explicitly test rule abstraction, or 'extrapolation' behaviour (Conaway & Kurtz, 2017). Participants were trained on a two-dimensional version of the problem (Fig. 1B). However, some stimuli were left untrained (empty cells in Fig. 1B). Crucially, the untrained stimuli were presented in a subsequent test phase, where about 31% and 45% of the participants, in Exp. 1 and 2B, respectively, still extrapolated 'B' for stimuli in the lower right quadrant (shaded area in Fig. 1B), while others responded 'A'. The response pattern of those participants who extrapolated 'B' corresponds to a complete Type II solution, which could be described in terms of contextual modulation, as well.

Despite the evident structural similarity to the classic Type II problem, XOR extrapolation can not be predicted by exemplar models such as ALCOVE (Kruschke, 1992), or SUSTAIN (Love, Medin, & Gureckis, 2004). While the model DIVA (Kurtz, 2007) has been shown to predict XOR extrapolation to some extent, it seems to be an open question whether it can account for the behavioural differences in the problem Types I-VI (see further Kurtz et al., 2013).

Here we assume that differences in contextual modulation drive the described performance differences in the incomplete XOR and the Type II problems. In particular, we assume that there is a learning process that triggers contextual modulation. We call this process 'contrasting', i.e. the tendency to abstract regularities from unobserved instances, which is further explained below.

In the following, we introduce CALM, a Category Abstraction Learning Model. CALM first learns by associating dimension values to outcomes in a Hebbian fashion (Hebb, 1949), and generates outcome expectations from these associations. To generate these associations, CALM generalizes from observed, and abstracts from unobserved instances. It monitors the success of its expectations, and reinforces learning about predictive (diagnostic) dimensions. It can detect whether its predictions are systematically violated in specific contexts and modulates these predictions accordingly.

We first give an overview of the model and its theoretical motivation, and describe its core equations. The current version is formalized for learning binary outcomes, for simplicity. We then present simulations, showing how CALM meaningfully predicts the whole pattern of performance in the described tasks.

Description of CALM

CALM represents stimuli using a separate set of ordered dimension nodes for each stimulus dimension (Fig. 2). Stimulus presentation activates the corresponding node on each dimension, which feeds forward to its own set of dimension-specific category nodes. The activation of the category nodes on a dimension m generates a dimension specific prediction of category membership (Fig. 2; r_m), which is calculated as the log odds ratio between the strengths of the two node-to-category associations. They can be positive or negative, predicting category A or B, respectively.

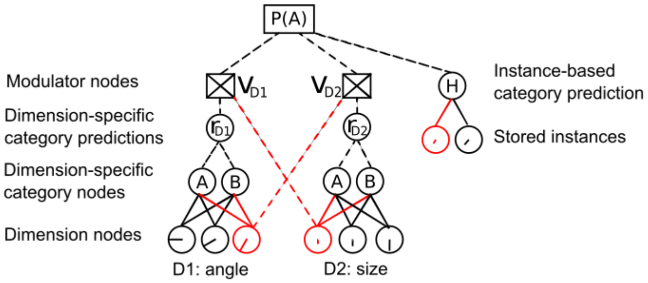
P(A)
Modulator nodes
V_{D1}
V_{D2}
Instance-based category prediction
H
Dimension-specific category predictions
r_{D1}
r_{D2}
Stored instances
Dimension-specific category nodes
A
B
A
B
Dimension nodes
D1: angle
D2: size

Figure 2: Schematic of CALM. Solid lines are modifiable-strength connections, dotted lines are fixed strength. Components in red indicate initial activity in CALM as the result of the presentation of a stimulus.

A dimension-specific prediction (r_m) can be modulated by the active node on another dimension. For example, in Fig. 2, the modulator value for the size dimension (v_2) might be negative for horizontal lines, but positive for vertical lines. Thus, multiplying r_2 and v_2 would invert the prediction of the size dimension for horizontal lines, but not for vertical lines. Passing the product of r_2 and v_2 to the decision process, thus, conveniently simplifies our assumptions about contextual modulation in terms of inversion (keep vs. inverse). Please note, CALM does not represent 'explicit rules', e.g. decision bounds, but simple associations between the cues of a dimension and the outcomes.

Stimulus presentation can also activate an instance representation (if one has been stored), generating an instance-based prediction of category membership (Fig. 2; H). We follow the assumption that people store configural stimulus representations separately from generalized category representations (e.g. RULEX, Nosofsky, Palmeri, & McKinley, 1994). Only one stored instance, most similar to the presented stimulus, contributes to this prediction by its associations to the category labels, which determine the magnitude of H. Thus, CALM differs from exemplar accounts, e.g. GCM (Nosofsky, 1986), in that it uses nearest neighbour rather than all-exemplar computations. A decision process sums the category prediction terms, H, and $r_m v_m$, and converts this sum into a prediction of category membership using an exponential ratio rule, for simplicity (Wills, Reimers, Stewart, Suret, & McLaren, 2000, for a more realistic alternative).

During learning, CALM interprets feedback, e.g. "It was category A", as saying both A is present, and B is absent. Regarding the *present* category, e.g. A, the associative strengths between the category node A and the basic nodes of a dimension are increased in a Hebbian fashion (Δw_C, Fig. 3). This additive increase is maximal for the dimension node that was activated by the stimulus, and decays with the distance from

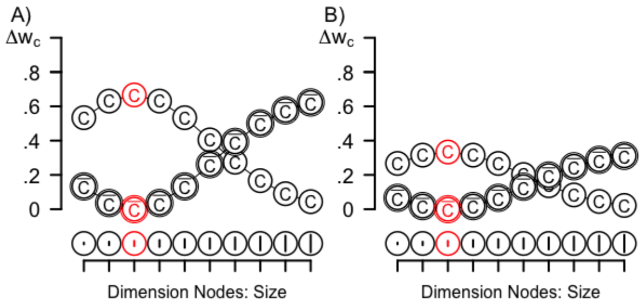

Figure 3: Update of associative strengths, exemplified for a 'size' dimension with 10 'size' nodes. Red components indicate the active stimulus input node. (A) Update on a dimension that is more diagnostic ($\alpha_m = 2/3$), and (B) that is less diagnostic ($\alpha_m = 1/3$) than other dimensions. C indicates excitatory generalization of the links to the *present* category; $\bar{C}$ indicates contrasting of the links to the *absent* category.

this node. This excitatory generalization allows CALM to respond to (novel) dimension cues that are similar to previously presented dimension cues, and was inspired by Shepard's law of generalization (Shepard, 1987). The decay is modifiable, however, the overall strength of the update depends on the model's assumption of the diagnosticity of a dimension. This means, if the nodes on a dimension already predict category membership better (more diagnostic) than the nodes on other dimensions, CALM will update the diagnostic dimension more strongly (see Fig. 3A).

Regarding the *absent* category, e.g. B, the associative strengths between the category node B and the dimension nodes are also increased ($\Delta w_{\bar{C}}$, Fig. 3). However, the update is zero for the dimension node that was activated by the stimulus, and increases with the distance from this node. This process, henceforth called 'contrasting', reflects an awareness that specific dimension cues are not observed together with specific categories. Related ideas can be found in other models (Stewart & Brown, 2005; Wagner, Brandon, Klein, & Mowrer, 1989). The contrasting decay is modifiable, and may vary independently from the decay of generalization. In the current version of CALM, however, we assume that contrasting is the inverse of excitatory generalization. Importantly, a steeper contrasting decay in CALM represents stronger abstraction, as it generates stronger 'hypotheses' about how to respond to (novel) dimension cues that are dissimilar to the previously presented dimension cues. As for generalization, the overall strength of contrasting depends on the diagnosticity of a dimension.

CALM's assumption of dimensional diagnosticity is not fixed. CALM continuously estimates the extent to which each dimension predicts different categories, using the standard deviation of the dimension's predictions along its dimension nodes. The corresponding diagnosticity values (α_m) reinforce or attenuate excitatory generalization and contrasting, which therefore could be interpreted as focused attention following associability. The use of α_m to indicate diagnosticity is a reference to the Mackintosh (1975) theory of attentional learning (see Le Pelley et al., 2016, for a recent review).

The update of the modulator nodes depends on the success of a dimension in predicting the *present* outcome. For example, in Fig. 2, if the size dimension (r_2) predicts category A when B is correct, then its modulator node (v_2) registers this failure. In this case, the modulator update is negative, otherwise positive, and recurring negative teaching signals eventually lead to contextual modulation.

The associative update is maximal for the modulator node that was activated by the stimulus, and decays with the distance from this node, following the same decay as for excitatory generalization. Thus, CALM will tend to modulate a dimension's prediction in the presence of (novel) context cues similar to observed context cues. In addition, the update is weighed by α_m which increases learning about the modulators of the most diagnostic dimensions. The update is also weighed by a similarly calculated diagnosticity value of the modulator dimension β_n. Thus, CALM's learning depends on 'focused attention' on two levels, which has been theorized previously (Matsuka & Corter, 2008).

One important prerequisite of CALM's modulation learning is the maintenance of the modulated dimension associations, instead of their correction. This requires 'conditional error discounting'. Therefore, CALM gradually suspends the Hebbian update on a dimension when its corresponding active modulator becomes increasingly negative.

Error discounting is not a new idea, the current implementation, however, deviates from previous suggestions, such as 'annealing' parameters that reduce learning rates over time, e.g. RASHNL (Kruschke & Johansen, 1999). Nonetheless, an overarching 'annealing' process is a side-effect of CALM's current architecture, due to the Hebbian learning on a log scale, i.e. the log odds predictions r_m of the dimensions are most volatile with early increments of the associative strengths, while later increments have less impact.

The final learning process in CALM concerns the creation of instance representations in memory. On each trial, CALM strengthens the association between a configural representation of the presented instance and its category label, with modifiable learning speed. However, similar but not identical to the error-based formation of exemplar clusters in SUSTAIN, (Love et al., 2004), CALM's process of storing instances is only effective if the unmodulated dimension predictions predicted the wrong category. Otherwise, the memory update is very weak, which formally defines strongly memorized instances as exceptions from the generalized category representations.

Formal Description

The presence of a stimulus S activates the associations $\mathbf{w}_{mIk}$ (initialized to .1) between the corresponding input node I and the two category nodes k on each dimension m. The vector of dimension-specific predictions $\mathbf{r}_{mIK}$ for category K represents the log odds ratios between the associative strengths.

$$\mathbf{r}_{mIK} = ln\left(\frac{\mathbf{w}_{mIK}}{\mathbf{w}_{mIk\neq K}}\right) \tag{1}$$

Zero values of $\mathbf{r}_{mIK}$ indicate equal associative strengths, positive values predict category K, and negative values predict the other category. The absolute magnitudes of $\mathbf{r}_{mIK}$ reflect the strengths of the dimensions' predictions.

The presence of the stimulus also activates the modulator associations $\mathbf{v}'_{mnJ}$ (initialized to 0) between the prediction of dimension m and the active input node J on each remaining dimension n (with $n \neq m$). However, only one modulator value $\mathbf{v}_m$ contributes to the modulation of dimension m, satisfying two constraints on cognitive complexity. In the first place, dimension m must be diagnostic, i.e. $\mathbf{v}_m$ is set to 1, if $\alpha_m \leq 1/M$, with M = number of dimensions. Second, the modulator dimension n must be the most diagnostic, i.e. $\mathbf{v_m} = \mathbf{v}'_{mnJ}$, if $\beta_n = max(\beta_{n\neq m})$. With two or more equally diagnostic modulators, the one with the stronger association is selected. In the case of same strengths, the values are averaged.[1] Both diagnosticity monitors α_m and β_n are initialized to equal weights, i.e. $1/M$.

The presence of the stimulus also activates the nearest configural instance representation Y, which is associated to the category labels k with strengths h_{Yk} (initialized to 0). Instance selection is based on the absolute distance d_y between the values σ^S_{mI} of the current stimulus and the values σ^y_{mi} of the stored instances y.

$$d_y = \sum_m |\sigma^S_{mI} - \sigma^y_{mi}| \tag{2}$$

For purposes explained below, CALM does not represent the stimulus values by their physical magnitude. Instead, the physical values x_{mi} of a dimension m at node i are standardized to σ_{mi}, with $\sigma_{mi} = {}^{(x_{mi}-\min(x_m))}\!/\!_{(\max(x_m)-\min(x_m))}$, scaling the maximum value on each dimension to 1^2. The selected instance Y satisfies $d_Y = \min(d_y)$ across all instances. If two or more instances are equally distant, the instance with the strongest category association h_{yk} is chosen.

The instance-based prediction H_K for category K is then obtained by subtraction, i.e. $H_K = \left(h_{YK} - h_{Yk\neq k}\right)\max\left(abs(\mathbf{v}_m)\right)$, which is positive when the selected instance predicts K, and negative if it predicts the other category. The difference is weighed by the maximum absolute value of the active modulators for scaling, since $\mathbf{r}_{mIK}$ can be amplified by $\mathbf{v}_m$ (see below), which may vary between tasks. H_K is then passed to an exponential ratio rule, together with the vectors of the dimension predictions $\mathbf{r}_{mIK}$, and their modulation values $\mathbf{v}_m$, calculating the probability $p(K|S)$ of choosing category K given stimulus S.

$$p(K|S) = \left(1 + e^{-H_K - \sum_m \mathbf{r}_{mIK}\mathbf{v}_m}\right)^{-1} \tag{3}$$

When feedback is provided, CALM updates the associations between the basic nodes and the category nodes on each dimension via generalization and contrasting (Fig. 3).

Excitatory generalization is applied to the associations between each node i of dimension m and the category node for the *present* outcome C. The update follows a Gaussian decay, which is maximal at the location I of the stimulus S, and decreases in strength with the distance from this node, governed by the free generalization parameter g.

$$\Delta\mathbf{w}_{miC} = \left(1 + e^{-v_m}\right)^{-1}\alpha_m\left[e^{-|\sigma^S_{mI} - \sigma_{mi}|^2/2g^2}\right] \tag{4}$$

Contrasting is applied to the associations between each dimension node and the category node for the *absent* outcome $\overline{C}$. The update is identical to excitatory generalization, despite using the inverse of the Gaussian decay. The update for the active node I is zero, increasing in strength with distance.

$$\Delta\mathbf{w}_{mi\overline{C}} = \left(1 + e^{-v_m}\right)^{-1}\alpha_m\left[1 - e^{-|\sigma^S_{mI} - \sigma_{mi}|^2/2g^2}\right] \tag{5}$$

Both updates are weighed by the diagnosticity α_m of the dimension, and the conditional error discounting term $\left(1 + e^{-v_m}\right)^{-1}$, which approximates 0 with negative values on the active dimension modulator $\mathbf{v}_m$. I.e. if a dimension is currently strongly modulated, the updates are suspended, and the current associations are maintained.

The updates of the modulator associations $\mathbf{v}'_{mnj}$ for the j nodes of each dimension n modulating the predictions of dimension m follow the same decay as used for excitatory generalization. The update is strongest for the active modulator node J and decays with distance from this node.

$$\Delta\mathbf{v}_{mnJ} = sign(\mathbf{r}_{mIC})\alpha_m\beta_n\left[e^{-|\sigma^v_{nJ} - \sigma_{nj}|^2/2g^2}\right] \tag{6}$$

The update is weighed by the diagnosticity α_m of the modulated dimension m, and the diagnosticity β_n of the modulator dimension n. The current success of dimension m in predicting the correct category C, i.e. $sign(\mathbf{r}_{mIC})$, determines the direction of the update, which is -1 if incorrect, and +1 if correct. The strengths of the modulator nodes are reset to +-5 if they become stronger, which limits the response amplification when taking the product in Equation 3.

A memory update further strengthens the association h_{SC} between the instance representation S and the correct category C, where the free parameter λ reflects the learning rate.

$$\Delta h_{SC} = \lambda(1 + e^{\sum_\mathbf{m} \mathbf{r}_{mIC}})^{-1} \tag{7}$$

If the correct category C was strongly predicted by the unmodulated dimension nodes, as indicated by the ratio rule, the update will be nearly 0, regardless of the learning rate λ, thus, efficiently storing exceptions only. The *unmodulated* predictions are chosen, because we think of modulation itself not as a representation of category space, but an awareness of when predictions require 'adjustment', which may also result in configural memory representations.

[1] Different active modulator associations having the same strength might arise with correlated dimensions.

[2] E.g. a binary dimension [white, black] is coded as [0,1], while equally spaced shadings [white, lightgrey, grey, darkgrey, black] are coded as [0,.25,.5,.75,1].

The final updates concern the diagnosticity values α_m and β_n. Dimension diagnosticity α_m is formalized as the extent of variation in the predictions of a dimension m along its nodes i, by taking the standard deviations SD_m of the corresponding vector of dimension predictions r_{mi}.

$$\Delta\alpha_m = \frac{SD_m(r_{mi})}{\sum_m SD_m(r_{mi})} \tag{8}$$

The estimates of $\Delta\alpha_m$ are normalized across dimensions, and then averaged with the previous values of α_m. In case of zero variances, α_m is reset to 1/M.

The update of the modulator diagnosticity β_n is calculated similarly, by taking the standard deviations of the associations $\mathbf{v}'_{mnj}$ along the modulator dimension nodes j. However, only their sign is taken into account for neglecting variations unrelated to contextual modulation.

$$\Delta\beta_n = \frac{\sum_{m\neq n} SD_n(sign(\mathbf{v}'_{mnj}))}{\sum_n \sum_{m\neq n} SD_n(sign(\mathbf{v}'_{mnj}))} \tag{9}$$

The standard deviation is taken separately for the associations between the nodes of a modulator dimension n and each remaining dimension m (with $m \neq n$), before they are summed across the m dimensions, and normalized across the modulator dimensions. Thus, β_n reflects overall modulator diagnosticity. The update is then averaged with the previous values of β_n. In case of zero variances, β_n is reset to 1/M.

In its current version, CALM has 2 free parameters, i.e. the exception learning rate λ, and the generalization decay g, which is inversely related to the contrasting decay. Large values of g indicate strong generalization and weak contrasting, and vice versa for low values of g. Due to CALM's standardized dimension values, i.e. a dimension's maximum is always 1, the impact of g is scaled and g estimates can be directly compared between tasks that include stimuli with differently scaled dimensions.

Simulations

In the following, we simulate CALM's performance for the six standard category learning tasks, and the incomplete XOR task (see Fig. 1). To address the initially described differences in task performance, we assume that rule instructions increase the tendency to abstract category membership, i.e. increase contrasting (low g) in CALM. This increases the occurrence of prediction errors in specific contexts, which frequently triggers contextual modulation. We consequently assume that the absence of rule instructions leaves contrasting at a moderate level (high g), such that contextual modulation is triggered less frequently. Conaway and Kurtz (2017) point out that they did not instruct the presence of rules in their incomplete XOR task, hence, we again assume moderate contrasting (high g) in this task.

We presented the problem Types I-VI (Fig. 1A) to CALM, each in 8 blocks of training, as done in Kurtz, et al. (2013). For simulating the classic pattern, i.e. with rule instructions, 500 values of g were sampled from a truncated Gaussian distribution with a low mean and a small standard deviation, i.e. $g \sim Gaussian_{(0,3)}(.3,.1)$. For simulating the revised pattern, i.e. without rule instructions, 500 values of g were sampled from a distribution with a high mean and a large standard deviation, i.e. $g \sim Gaussian_{(0,3)}(.7,.5)$, which was also done for simulating CALM's performance in the incomplete XOR task, where we presented the problem (Fig. 1B) in 12 training blocks, followed by a test block including all possible stimuli, as done in Conaway and Kurtz (2017). The exception learning rate was always sampled with $\lambda \sim Gaussian_{(0,\infty)}(.15,.05)$. Figure 4 depicts the results.

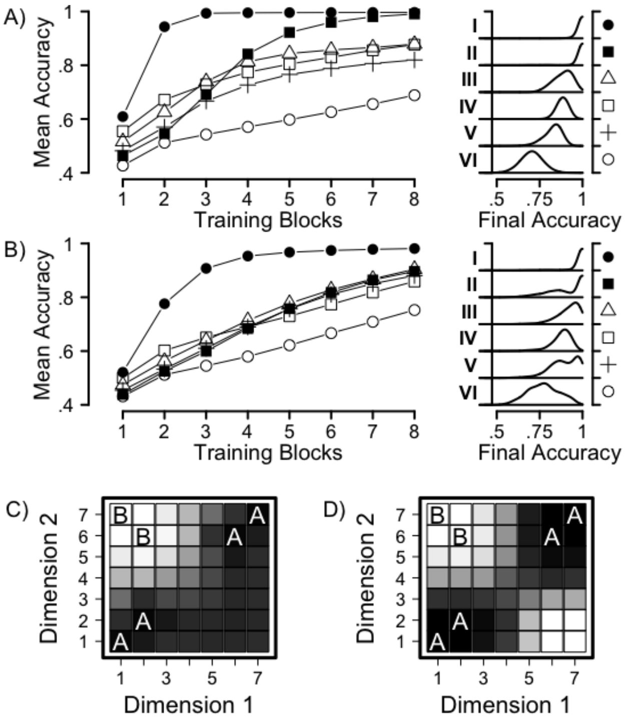

Figure 4: CALM simulation results. (A) and (B) depict CALM's average learning curves (left) for the six problem Types I-VI (see Fig. 1A), with strong contrasting in (A), and weak contrasting in (B). The right columns depict the corresponding distributions of accuracy in the final training block. (C) and (D) depict the predicted average response gradients in the incomplete XOR task (test phase) for two sub-groups of samples. Letters indicate categories of trained items. Response probability is indicated by shading; white=100% 'B', black=0% 'B'. See text for further details.

As can be seen, CALM with strong contrasting (Fig. 4A) predicts the classic ordinal pattern of learning. CALM with weak contrasting (Fig. 4B) predicts the revised pattern, including a bimodal Type II distribution, without affecting the remaining pattern, well accounting for the described phenomena, by assuming variations in the strength of contrasting.

The simulation results for the incomplete XOR task can be seen in Figures 4C and 4D. The two plots represent two groups of sub-samples, i.e. the 500 samples were split by the mean probability of responding 'B' in the lower right quadrant (shaded area in Fig. 1B) being smaller (4C) or larger than 50% (4D). The depicted sub-groups contain 69% and 31% of the 500 samples, respectively. As can be seen, CALM predicts the pattern of performance as described (Conaway & Kurtz, 2017), by assuming individual variations in the strength of contrasting.

Discussion

We introduced CALM, a Category Abstraction Learning Model, combining generalization, abstraction and problem structuring. CALM successfully accounts for the observed behavioural pattern in the classic benchmark categorization problems Type I-VI (Shepard, Hovland, & Jenkins, 1961), as well as for the systematic influence of rule instructions on Type II performance (Kurtz et al., 2013). This is achieved by assuming systematic variations in the tendency to abstract category representations, described by a contrasting mechanism that generalizes from unobserved instances and triggers the contextual modulation of dimension-specific predictions. In addition, we have shown that individual differences in this process might also explain the observed patterns in the incomplete Exclusive-Or problem (Conaway & Kurtz, 2017).

The results of our model simulations suggest that contrasting (inverse generalization) might be a key process in category learning. Importantly, CALM further implements constant monitoring of the network state, for reinforcing learning about diagnostic dimensions, for selecting potential modulators, and for maintaining associations of modulated dimensions (conditional error discounting).

We thereby put strong assumptions into the model, which have yet to be tested. CALM provides various predictions for sophisticated experimental designs, including variations in learning order, task switching, or the statistical and structural properties of categories. In addition, CALM provides a framework that includes attention parameters on two cognitive levels, which has been theorized before (Matsuka, & Corter, 2008).

As a next step, we seek to validate the model in further studies. We believe, that CALM has the promising potential to provide an alternative approach to human learning behaviour in a wide array of tasks.

Acknowledgments

The work on this project was funded by the Swiss National Foundation (SNF).

References

Conaway, N., & Kurtz, K. J. (2017). Similar to the category, but not the exemplars: A study of generalization. *Psychonomic bulletin & review*, *24*(4), 1312–1323.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory.* New York: Wiley.

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of linear experts: Knowledge partitioning and function learning. *Psychological Review*, *111*(4), 1072.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological review*, *99*(1), 22.

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(5), 1083.

Kurtz, K. J. (2007). The divergent autoencoder (DIVA) model of category learning. *Psychonomic Bulletin & Review*, *14*(4), 560–576.

Kurtz, K. J., Levering, K. R., Stanton, R. D., Romero, J., & Morris, S. N. (2013). Human learning of elemental category structures: Revising the classic result of Shepard, Hovland, and Jenkins (1961). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 552.

Le Pelley, M. E., Mitchell, C. J., Beesley, T., George, D. N., & Wills, A. J. (2016). Attention and associative learning in humans: An integrative review. *Psychological Bulletin*, *142*(10), 1111.

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

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological review*, *82*(4), 276.

Matsuka, T., & Corter, J. E. (2008). Observed attention allocation processes in category learning. *Quarterly Journal of Experimental Psychology*, *61*(7), 1067–1097.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, *115*(1), 39.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing modes of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & cognition*, *22*(3), 352–369.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

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

Stewart, N., & Brown, G. D. (2005). Similarity and dissimilarity as evidence in perceptual categorization. *Journal of Mathematical Psychology*, *49*(5), 403–409.

Wagner, A. R., Brandon, S. E., Klein, S., & Mowrer, R. (1989). Evolution of a structured connectionist model of Pavlovian conditioning (AESOP). *Contemporary learning theories: Pavlovian conditioning and the status of traditional learning theory*, 149–189.

Wills, A., Reimers, S., Stewart, N., Suret, M., & McLaren, I. (2000). Tests of the ratio rule in categorization. *The Quarterly Journal of Experimental Psychology: Section A*, *53*(4), 983–1011.