Category Representation in Neural Networks with Dynamic Learning Rates   
matches Human Learning Performance

A Connectionist View of Common Coding in Category Learning

# Abstract (431 words)

Agents tend to group together physically distinct stimulus exemplars that have never directly been paired together but that share predictive value with respect to some outcome. This “common coding” phenomena was examined with a category learning task. We conducted two experiments, where human participants learned to classify arbitrarily chosen exemplars (different fractal images) into two distinct categories. Each participant completed two tasks that involved reversing the exemplar-category assignments after the initial learning phase. In the total reversal task, all exemplar-category assignments were reversed together, while in the partial reversal task half the exemplar-category relations were reversed with the remaining ones maintained. We predicted that participants would learn total reversals faster than partial reversals to the extent that within-category grouping took place. Experiment 1a trained participants with 4 distinct exemplars mapping onto 2 separate categories, while Experiment 1b trained participants with 8 distinct exemplars. In a third experiment, we evaluated whether the baseline perceptual distinctiveness of the exemplars within- and between-categories differed. We employed Multidimensional Scaling and Euclidean distance measures to analyse perceived similarities among exemplars both within and between categories in a 'psychological' space. Altogether, results showed that participants learned faster the total reversal task and displayed interference not only with reversed items but also non-reversed items in the partial reversal task; also, there were no significant differences in the perceived similarities among exemplars used within or between categories, suggesting that the learning effects we observed were not due to different baseline perceptual similarities among our exemplars. To explain the results, we used a standard 3-layered neural network to model category learning with an error backpropagation learning algorithm to adjust the weights between units. In these models, categories are represented by distinct unit activation vectors at the hidden layer, and different within-category exemplars are “grouped together” by activating the same hidden layer category representations. However, these models also suffer from “catastrophic interference” under reversal conditions. In order to preserve hidden layer category representations to enable more rapid total than partial reversal learning an additional mechanism would be required to produce slower weight changes in the input-hidden layer and, correspondingly, faster weight changes in the hidden-output layer. We incorporated the different dynamic learning rate rules inspired by Mackintosh (1975) and Pearce and Hall (1980) from the animal learning literature to implement this additional mechanism, and the model faithfully reproduced our empirical results. Overall, these findings underscore the significance of associative mechanisms in category formation and reversal learning, and they demonstrate how catastrophic interference in neural nets can be avoided with different dynamic learning rate adjustments made at different layers of the network.

This is The Manuscript

# Introduction (484 words)

The New York Yankees have several skilled hitters. When two different players hit many home runs over the year, we often label those players as 'good hitters' or ‘home run hitters.’ This exemplifies the concept of common coding, where two cues (players) share the same outcome (hitting home runs), and this results in the formation of a common code (home run hitter) by which we relate the different players. During the off-season, for example, general managers looking to obtain a ‘home run hitter’ may then categorize different players interchangeably.

More broadly, common coding is a cognitive phenomenon in which an agent learns that one stimulus generalizes more to another when they have previously been associated with similar outcomes (Delamater & Joseph, 2000). Common coding, also known as acquired equivalence, is often recognized as important for category formation, although its underlying mechanism is not well understood (Delamater, 1998; Goldstone, 1994; Honey & Hall, 1989; Miller & Dollard, 1941). In this paper, we establish a common coding effect using a category reversal learning task (see also Zentall et al., 1991, 1992) and offer a mechanistic explanation of category learning using an artificial neural network model of associative learning (Castiello et al., 2021; Delamater, 2012).

Common coding effects have been observed across various settings and species. It has been documented using instrumental conditioning procedures in pigeons (Urcuioli et al., 1995), in Pavlovian conditioning tasks with rats (Johns & Williams, 1998), and in matching to sample tasks with humans, even when the paradigm requires common consequences and common antecedents (Delamater & Joseph, 2000). Regardless of species or conditioning type, common coding may share similar underlying mechanisms.

An experimental approach to test whether common coding develops between stimuli categorized similarly is to reverse the category assignments either totally or partially following an initial training phase. If participants initially learn to group one set of exemplars in Category A and another set in Category B, then there is an opportunity for distinct common codes to develop among the exemplars within each category. During a total reversal, all the exemplar-category assignments are switched. If common coding occurs, then learning the total reversal would simply involve reassigning the internal common codes attached to each exemplar with the updated category assignment and this should be relatively easy. In contrast, in the partial reversal condition only half the original exemplar-category assignments are reversed. This requires a complete remapping of the exemplar-category common codes since exemplars from each original category are now treated differently than before. Faster total than partial reversal learning has been observed in pigeons, rodents, and humans in related tasks (Delamater & Joseph, 2000; Urcuioli et al., 1995). The rationale for the present experiment is to investigate this total/partial reversal effect using a more standard human categorization task than has been reported in the literature. In Experiment 1a we trained participants on a 4-exemplar and in Experiment 1b an 8-exemplar categorization task.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 1.** | | | | | | | | | |
| Categorical Reversal Learning Experimental Design (easy) | | | | | | | | | |
| **Category Total Reversal** | | | | | **Category Partial Reversal** | | | | |
| Phase 1 | | Phase 2 | |  | Phase 1 | | Phase 2 | |  |
| Learning | | Reversal | |  | Learning | | Reversal | |  |
| Cues | Cat. | Cues | Cat. |  | Cues | Cat. | Cues | Cat. | Condition |
| A | N← | A | →S |  | A | N← | A | →S | reversal |
| B | N← | B | →S |  | B | N← | B | N← | non-reversal |
| C | →S | C | N← |  | C | →S | C | N← | reversal |
| D | →S | D | N← |  | D | →S | D | →S | non-reversal |
| *Note*: A, B, C, and D are cues in form of fractals; S← and →N are two distinct response categories (Cat.). | | | | | | | | | |

# Experiment 1a – Easy Categorization Task

Our categorization task consisted of exposing participants on a given training trial to one of a set of exemplars and asking them to learn to categorize it appropriately based on corrective feedback for each trial. A set of fractal images were used as exemplars that participants were to group into two distinct categories. Following an initial training phase long enough for participants to learn the task, the exemplar-category assignments were reversed using either the partial or total reversal protocols described above. Each participant performed in a partial reversal task with one set of exemplars and in a total reversal task with a second set of exemplars, with the order in which they experienced the two and their identities counterbalanced across participants.

**Table 1** presents the experimental design in more detail. Participants were led to believe that each fractal image represented the molecular structure of an object in the world, and it was their task to determine if the object came from the Northern (N) or Southern (S) hemisphere. Corrective feedback was given on each trial. 4 exemplars (cues A-D) were used in Experiment 1a in which participants first learned to classify 2 exemplars (A, B) as belonging to the Northern and 2 (C, D) to the Southern hemisphere. During the Total Reversal phase, the two exemplars that previously were classified as belonging to the Northern hemisphere now were correctly classified as belonging to the Southern hemisphere, and vice versa. During the Partial Reversal phase one of the exemplars initially classified as belonging to the Northern hemisphere continued to be classified in that manner with the other one now being classified as belonging to the Southern hemisphere, and vice versa. During the reversal phases we were interested in determining if the Total Reversal condition would result in more rapid reversal learning relative to the reversed exemplars in the Partial Reversal condition, and whether categorization performance among non-reversed exemplars in the Partial Reversal condition might deteriorate as the new categorizations were learned. Both these effects would be expected if within-category exemplars developed a “common code” during the initial classification learning phase.

## Methods

### Participants

Thirty-two participants were recruited through the Prolific platform ([www.prolific.com](http://www.prolific.com)), with the constraints that they were at least 18 yr old, performed the experiment on a computer, and resided in either the UK or USA. Their ages ranged between 20 and 63 with a mean of 36.29 and standard deviation (SD) of 10.86. Twenty participants were male, 7 female, one non-binary, and 4 did not respond. The experiment lasted approximately 12 minutes and participants were paid $2.50 for their participation. All procedures posed no risk to participants and were approved by ARD’s CUNY IRB protocol.

We exclude participants that did not show adequate learning in the first 96 trials of the learning phase (before reversal, see below). We define adequate learning when the probability of correct categorical classification, averaged over all phase 1 training trials, exceeded chance performance using a binomial test where chance = 0.5. Thus, participants must have at least 57 correct trials over 96 trials in phase 1. Two participants were failed to meet this criterion and were excluded.

### Task and Procedures

The experiment consisted of two tasks: Total and Partial Reversal. Every participant performed in each one with the order of tasks counterbalanced. Each task, in turn, included two training phases: learning and reversal. The tasks had 6, 8-trial blocks of training trials in each phase. In each block there were 2 training trials with each one of the 4 fractal images, and the trials were fully randomized. This gave rise to a total of 6 (blocks) \* 8 (trials/block) \* 2 (phases) = 96 trials for each of the two tasks.

Participants were given a cover story indicating that fractal images represented the molecular structure of different objects in the world and were asked to classify the four fractals into two categories – objects coming from the “Northern Hemisphere” (N) or “Southern Hemisphere” (S; see instructions in the **Appendix**). In every trial, participants saw one fractal exemplar and selected which category it belonged by pressing the left or the right arrow keys (**Figure 1A**). Left was linked with the N (*i.e.*, N←; see **Table 1**) and right with S (*i.e.*, →S). Between trials there was an Inter-Trial Interval (ITI) with a white fixation cross centered on a grey background. The ITI was a random variable distributed uniformly between .5 and 1.5 seconds rounded to the first decimal place. The exact instructions are provided in the **Appendix**. When a participant completed one task they were given the option of starting the second task after a short break.

### Software and Open Science

The task was programmed in PsychoPy® (Peirce et al., 2019) and the experiment was run online via Pavlovia through Prolific. All the statistical analysis were conducted in R (R Core Team, 2021). The PsychoPy code and R scripts are available in GitHub, along with the analysis scripts and previous drafts of the paper: <https://github.com/santiagocdo/categoricalReversalLerning>

All the fractal images used as stimuli in these experiments were kindly provided by the artist Cory Ench (<https://www.enchgallery.com/index.htm>). We chose 2 sets of 4 images to use in Experiment 1a and 2 sets of 8 images to use in Experiment 1b. There were two counterbalancing schemes used in which each unique image was assigned a role, across participants, to a total and partial reversal task.

### Analysis

To test the hypothesis, we used only the experimental data for the second phase, *i.e.*, post-reversal. We modelled correct classification (correct=1, incorrect=0) with a Logistic Mixed Model (LogMM):

, (*eq. 1*)

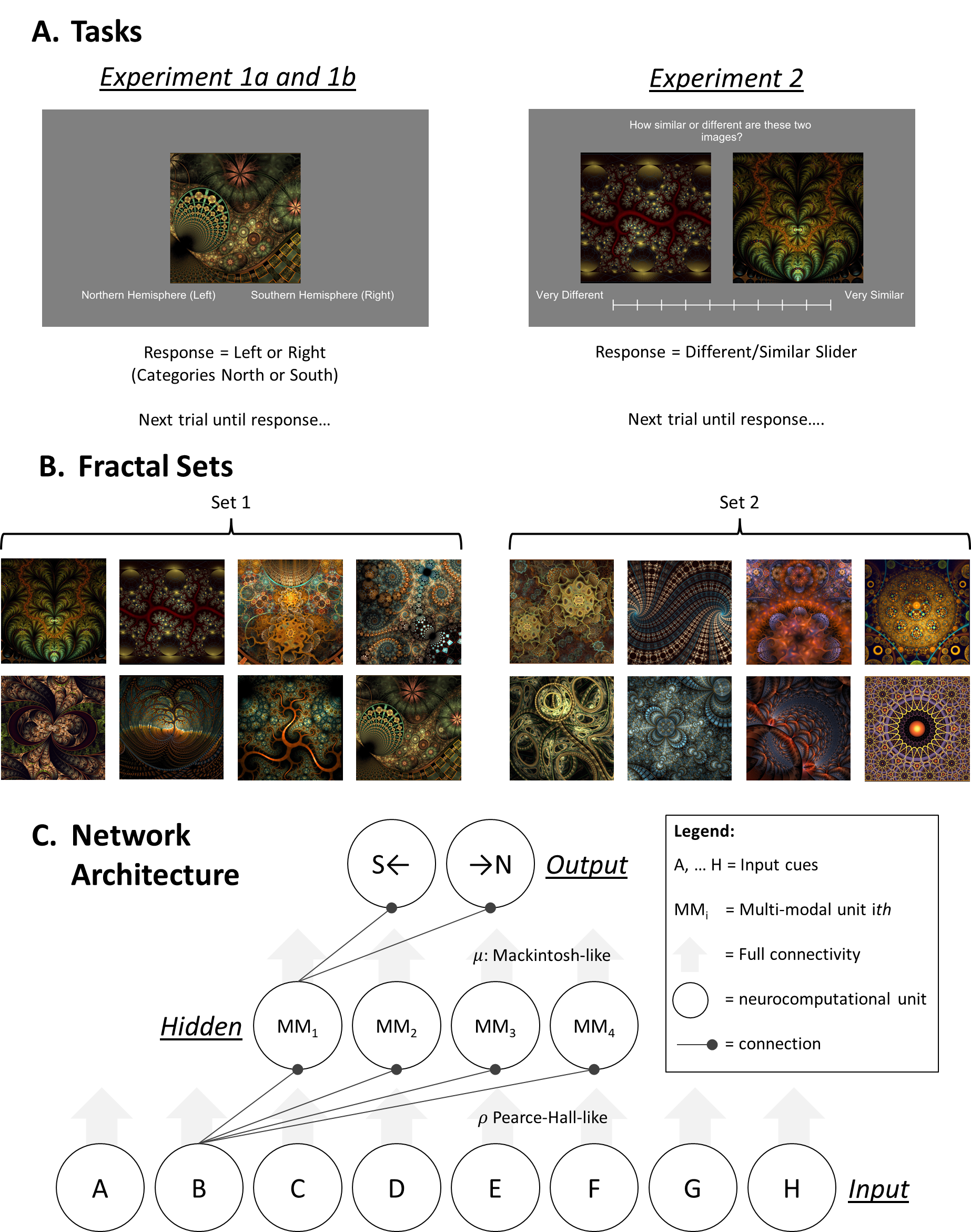
The model displayed in *eq. 1* had Blocks as random slope and participants as random intercept. To confirm whether the effect is due to the reversal, we ran another LogMM with only data from the first block in the second phase.

, (*eq. 2*)

No random slope was used, and participants were incorporated as random intercept. Lastly, we modelled Reaction Time (RT) with a Linear Mixed Model (LMM), equivalent to *eq. 1*. Although the model residuals from the raw data were not normally distributed, nor with log nor square root transformations, we decided to run the model with that limitation—we also conducted a sensitivity analysis with the log transformation, see Supplementary Results—. The model had the following form:

, (*eq. 3*)

Same random structure as *eq. 1*. The significance alpha for all models’ parameters was set to 0.05.



**Figure 1**. ***A****. present one trial for both tasks used in Experiments 1a and 1b, and 2.* ***B****. Sets of fractals used for each of the two tasks.* ***C****. Architecture used with all models: BP (Castiello et al., 2021; Delamater, 2012). 𝜌 control 𝛼 changes from Input-Hidden, and 𝜇 controls 𝛼 changes between Hidden-Outputs. Hidden layer always consisted in four multimodal units (i.e., fully connected). Fractal stimuli usage was authorized by artist Cory Ench.*

## Results

### *Superior Learning After the Total Reversal*

The mean probabilities of correct classification for all blocks in both tasks are presented in **Figure 2A**. In red the Total Reversal task averages performance across all exemplars from both categories experienced in each training phase. On the other hand, the partial task has two conditions during the reversal phase: reversed (blue) and nonreversed (green) trials. Although these distinctions did not apply during phase 1 training, that data are segregated in this manner as well. The critical comparisons are to compare in phase 2 the reversed exemplars between the total and the partial task (**Figure 2C**), and to assess performance on nonreversed trials in the partial task.

The LogMM revealed a positive main effect for total vs partial [], suggesting that in phase 2, participants had higher correct scores in the Total than Partial Reversal task. There is also a positive main effect of Blocks [], which suggest that participants improved across reversal trials. However, the interaction between task and Blocks was not significant []. Just to corroborate that partial reversal impaired reacquisition of the categories, we tested the first block between Total vs Partial, and we found a significant and large effect []. We also compared Total Reversal versus Partial Non-reversal conditions and found no differences (see **Supplementary Information**).

### No Differences in Reaction Time After Reversal

We also tested the RT for the second phase with an LMM. We found that the RT was not distinct between tasks (Total versus Partial) [], nor was there an effect of Blocks [], or an interaction []. We also ran the same model but after transforming the RT data with the natural logarithm and the root squared. The results were similar (see **Supplementary Information**).

### No Differences in Phase 1 Acquisition

As a control measure, we corroborated that learning in Phase 1 was not distinct between tasks (Total versus Partial Reversal). We fit another LMM with only data from the Phase 1. We found a positive significant effect of Blocks [] but no main effect of Task [] or an interaction []. This is evidence that participants similarly learn both tasks and conditions in Phase 1.

A group of graphs showing different types of data

AI-generated content may be incorrect.

**Figure 2**. *Left panels Experiment 1a, right panels Experiment 1b.* ***A*** *and* ***D*** *probability of correct as a function of blocks.* ***B*** *and* ***E*** *mean of median reaction times as a function of blocks.* ***C*** *and* ***F*** *same as* ***A*** *and* ***D*** *but only plotted for the reversal part of the task, after the vertical line between bocks 6th (easy task) and 8th (hard task). We also removed the partial-nonreversed (green colour) conditions. Error bars are standard errors of the means.*

# Experiment 1b – Hard Categorization Task

Experiment 1b was run in an identical manner to Experiment 1a except we required participants to categorize 8 exemplars into two categories of 4 exemplars each. This 8-exemplar task made it more difficult for participants to learn, and so the initial training phase was extended to 8 training blocks and since most of the effects in Experiment 1a appeared during the early reversal blocks we decreased the number of reversal blocks to 4.

## Methods

### Participants

Thirty-two participants were recruited through Prolific ([www.prolific.com](http://www.prolific.com)). The age ranged between 22 and 65 with a mean of 40.16 (SD 13.80). Eleven participants were male, 15 female, one non-binary, and 5 did not replied. The present experiment lasted approximately 20 minutes and participants were paid $4.00 for their participation. All procedures posed no risk to participants and were approved by ARD’s CUNY IRB protocol.

We excluded participants using the same criterion as in Experiment 1a. Thus, participants’ inclusion was based on a binomial test in which the probability of correct phase 1 performance had to be higher than random accuracy (0.5). This implies at least 142 correct trials over 256 trials in phase 1. Six participants were excluded for not reaching this criterion.

### Task and Procedures

The task structure is the same as Experiment 1a. The only difference in this experiment is the difficulty which was manipulated by adding more stimuli (fractals; **Table 2**). Also, the number of blocks per phase varied, where Phase 1 included 8 training blocks and 4 reversal blocks in Phase 2. Furthermore, each block had 16 trials (2 presentations of each exemplar). Within each block the trials were fully randomized for a total of 8 (phase1: blocks) \* 16 (trials) + 4 (phase 2: blocks) \* 16 (trials) = 192 trials for each of the two tasks (total and partial reversal).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 2.** | | | | | | | | | |
| Categorical Reversal Learning Experimental Design (hard) | | | | | | | | | |
| **Category Total Reversal** | | | | | **Category Partial Reversal** | | | | |
| Phase 1 | | Phase 2 | |  | Phase 1 | | Phase 2 | |  |
| Learning | | Reversal | |  | Learning | | Reversal | |  |
| Cues | Cat. | Cues | Cat. |  | Cues | Cat. | Cues | Cat. | Condition |
| A | S← | A | →N |  | A | S← | A | →N | reversal |
| B | S← | B | →N |  | B | S← | B | →N | reversal |
| C | S← | C | →N |  | C | S← | C | S← | non-reversal |
| D | S← | D | →N |  | D | S← | D | S← | non-reversal |
| E | →N | E | S← |  | E | →N | E | S← | reversal |
| F | →N | F | S← |  | F | →N | F | S← | reversal |
| G | →N | G | S← |  | G | →N | G | →N | non-reversal |
| H | →N | H | S← |  | H | →N | H | →N | non-reversal |
| *Note*: A, B, C, D, E, F, G, and H are cues in form of fractals; S← and →N are two distinct correct response categories (Cat.). | | | | | | | | | |

### Analysis

The same analysis plan was used as in Experiment 1a.

## Results

### Faster and Better Learning After the Total Reversal

**Figure 2D** shows the mean probability correct category classifications for each training and reversal blocks in the total and partial reversal tasks. The total reversal task is again represented by one line (red) collapsing across the two categories, but the partial task is represented separately for reversed exemplars (blue line) and non-reversed exemplars (green). The critical comparisons involve comparing Phase 2 data for reversed exemplars in the total and partial tasks (**Figure 2F**) as well as non-reversed trials in the partial task relative to reversed trials in the total task.

We fit a LogMM to the reversed exemplars in Phase 2, and the model revealed two significant main effects, total vs partial [] and Blocks []. The former suggests that, in Phase 2, participants were more accurate in the total than in the partial task; the latter effect suggests that overall participants improved over the reversal phase. In contrast to the previous experiment, the interaction between task and Blocks was significant [], this suggests that the rate of readjusting to the new phase (*i.e.*, the slopes difference between reversed tasks as a function of Blocks) was different. We also tested the first block between total vs partial, and we found a significant and large effect []. This suggests that total reversal was acquired quickly than partial reversal. We also compared total-reversal versus partial-nonreversed conditions and found no differences (see **Supplementary Information**).

### Larger Reaction Time in the Total Reversal

In contrast to the previous experiment, we found RT differences between tasks during the reversal phase. Overall the RTs in the total task were higher than in the partial task [], but there was no main effect of Blocks [] suggesting that people did not significatively decrease their RTs as they learned the new exemplar-category assignments. No significant interaction was found []. The results are similar when we used the log-transformed and root-squared RT (see **Supplementary Information**).

### No Differences in Phase 1 Acquisition

As in the previous experiment, it is important to test that the learning trajectories in Phase 1 were not different between the two tasks. We ran a similar model with Blocks, Task, and its interaction. We only found a significant effect of Blocks suggesting adequate learning in Phase 1[], but no Task as main effect [] or an interaction []. This is evidence in favour of similar learning in Phase 1.

# Experiment 2 – Stimuli Control

In Experiments 1a and 1b we assigned fractal images to the two categories using two sets of exemplars and made sure that across the total and partial reversal conditions the same physical set of exemplars underwent reversal training (or non-reversal training in the partial task). However, we did not randomize the assignments of individual exemplars across the two categories. Because there was no obvious way to categorize the various fractal images based on any particular set of perceptual features (e.g., animate/inanimate, spiky/stubby, red/green, circular/triangular shapes, etc), we did not attempt to randomize our assignments. Instead, we chose assignments that appeared to us to cut across any perceptual dimension we could think of. Therefore, we thought it would be important to collect normative data on an independent set of participants to test whether our fractal exemplar-to-category assignments resulted in different inherent similarities within vs between the two categories. We suspected there would be none, but, nevertheless, thought it important to establish that fact.

To accomplish this goal, we asked new participants to judge the similarities between every pair of images used in Experiments 1a and 1b. We then subjected this similarity rating data to multidimensional scaling procedures and obtained an overall distance measure within multidimensional space among those exemplars that appeared within and between categories. As will be seen below, our intuitions about the stimuli were supported by this multidimensional analysis.

## Methods

### Participants

Sixteen participants were recruited through Prolific ([www.prolific.com](http://www.prolific.com)). The age ranged between 28 and 61 with a mean of 42.69 (SD 9.91). Six participants were male and 7 were female. Again, all procedures were in accordance with ARD’s IRB protocol. The task took approximately 10 min and participants were paid $2.50 for their participation.

### Task and Procedures

Participants rated each pair of exemplars (**Table 3**) on a 10-point sliding scale varying from 'very similar' to 'very different'. Each trial consisted of a fixation cross followed by the rating screen. The fixation duration, serving as the inter-trial interval (ITI), was sampled from a uniform distribution ranging between 0.5 and 1.5 seconds. After the ITI, participants saw the rating screen (see **Figure 1A**), which remained until they provided their similarity or difference rating. The slider appeared 1 second after participants viewed the two exemplars. This delay ensured that participants had adequate visual exposure to both exemplars before responding automatically. They chose anywhere on the line (i.e., on or in between markings) and the computer registered their rating numerically as a real number between 1 and 10.

In experiments 1a and 1b, participants completed two tasks, each involving distinct set of eight exemplars (A–H). In this experiment participants rated all possible comparisons per set of exemplars. Thus 64 possible pairings (8 left positions × 8 right positions) minus the 8 self-comparisons resulted in 56 unique comparisons per task (**Table 3**). This number of trials ensured that each comparison was made twice, with left and right positions counterbalanced. Consequently, participants were expected to rate a total of 112 trials (56 trials per task \* 2 tasks). However, due to a programming error, the final comparison for each task (left H versus right G) was omitted.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 3.** | | | | | | | | | | |
| Categories Perceptual Similarity: exemplar comparisons | | | | | | | | | | |
|  |  | Right | | | | | | | | |
|  |  | A | B | C | D | E | F | G | H |
| Left | A |  | x | x | x | x | x | x | x |
| B | x |  | x | x | x | x | x | x |
| C | x | x |  | x | x | x | x | x |
| D | x | x | x |  | x | x | x | x |
| E | x | x | x | x |  | x | x | x |
| F | x | x | x | x | x |  | x | x |
| G | x | x | x | x | x | x |  | x |
| H | x | x | x | x | x | x |  |  |
| *Note:* 56 fractals similarity trials and ratings. 8 x 8 = 64, but 64 - 8 = 56, given the 8 stimuli diagonal, where it does not make sense to compare the same stimulus against itself. However, the comparison Left H and Right G was not tested due to a programming mistake. | | | | | | | | | | |

### Analysis

The intuition behind the analysis is that the perceived similarity between exemplars within one category should not differ from the perceived similarity within the other category. Similarly, perceived similarity within categories should not differ from the perceived similarity between categories. The perceived similarity was obtained by the distance between each pair of exemplars in in a 'psychological' space (describe below).

We first averaged the similarity ratings for each exemplar comparison. For example, we combined the similarity ratings for Left A vs. Right B with those for Left B vs. Right A. For each participant and each task, we used the 56 averaged similarity ratings between pairs of exemplars and implemented Multidimensional Scaling (MDS) with 2 dimensions (*k*=2). We choose two dimensions due to previous research in visual perception (Bracci & De Beeck, 2023). This provided us with 2 coordinates in “psychological” space for each exemplar. This procedure was repeated for each task and each participant. We then used the Euclidean distance to compare each exemplar against the others in the psychological space. For each participant, this yielded 28 distances, resulting from the combination of 2 elements in a set of 8, per task. We conducted a sensitivity analysis with three dimensions (MDS with *k*=3) and results are the same (see **Supplementary Information**)

We conducted two relevant comparisons: (1) the within-category distance for both categories used in Experiment 1, by comparing exemplars A to D versus exemplars E to H; and (2) between-category versus within-category distances. As the distance distributions did not pass the Kolmogorov-Smirnov test for normality, we used the Wilcoxon rank-sum test with continuity correction for the comparisons. The Type I error rate, alpha, was set to 0.05.

## Results & Discussion

The distributions of distances are shown in **Figure 3**. When we compare the distances between the categories (A to D versus E to H) used in Experiment 1, we did not find any differences in task 1, (**Figure 3A**), nor in task 2, (**Figure 3B**). In addition, we found no differences between within-category distances and between-category distances for task 1, (**Figure 3C**), nor for task 2, (**Figure 3D**).

These results suggest that participants did not perceive one set of category exemplars as more alike than the other. This provides evidence that the effects observed in Experiment 1 are not due to perceptual resemblance of the exemplars but due to the underlying learning of those categories.

A group of graphs showing different types of objects

AI-generated content may be incorrect.

**Figure 3**. *Category Similarity Analysis, using two psychological dimensions (multidimensional scaling with k=2) in the psychological space.* *Left panels (****A*** *and* ***C****) represent Experiment 1a results for the exemplars used in Task 1 (could be either total or partial, they were counterbalanced), and the right panels (****B*** *and* ***D****) Experiment 1b.*

# General Discussion

The results of our experiments demonstrate that category learning results in a grouping process whereby the cognitive machinery enhances the functional similarity among exemplars within a category and differentiates them from exemplars assigned to different categories. This effect is reminiscent of so-called acquired equivalence and distinctiveness of cue effects, common coding effects, and equivalence class effects investigated in the animal learning literature for some years (Sidman, 2000). We provided evidence for such an effect using a total/partial reversal learning paradigm in which reversed exemplar-category assignments were not only learned more rapidly in a total than partial reversal task, but performance was superior in total reversals compared to non-reversed exemplars in the partial task. Apparently, once the category task is learned, new learning to any one exemplar in the category can generalize more effectively to other within-category exemplars. This would result in both an enhancement of category reversal learning in the total reversal task, as well as impaired performance on reversed and non-reversed exemplars in the partial reversal task until new category structures emerge. In addition, our results in Experiment 2 suggest that there was very little basis for the two arbitrarily chosen set of exemplars other than their functional relationships defining their category assignments (i.e., objects belonging to the “northern” or “southern hemisphere”).

Our findings also are consistent with other results that have used a transfer procedure. Once two sets of exemplars have been categorized appropriately, researchers have shown that learning new responses transfer more readily among exemplars within than between categories. Both total/partial reversal and transfer effects, then, can be taken to support the basic finding of greater functional similarity among exemplars within and reduced similarity among exemplars between categories.

Others have commented on the possibility that associative processes might play a role in category learning (Delamater, 1998; Goldstone, 1994; Honey & Hall, 1989; Miller & Dollard, 1941). However, our results demonstrate that any such effort would require an associative framework that goes beyond the formation of simple binary exemplar-category associations. For example, if 8 distinct exemplar-category associations had been established, then there is no direct way in which total reversals should be learned more rapidly than partial reversals, for the simple reason that separate binary associations should not generalize any more to exemplars within versus between categories. What is needed is an associative framework that provides a mechanism whereby greater within- than between-category generalization can occur.

## Simulation

We have explored this idea by considering connectionist models that include a “hidden layer” in between an input layer that codes for individual exemplars and an output layer that codes for the different category responses used in the task (Castiello et al., 2021; Delamater, 2012). Within such a network, learning consists of weight changes between the input-hidden and hidden-output layers in such a way that exemplars come to activate the appropriate output layer category response. Importantly, such networks accomplish this by forming strong connections among all the exemplars within a category and the same set of hidden units that, in turn, come to form strong connections with the appropriate category output unit. This hidden layer convergence within-category and between-category divergence is exactly the type of mechanism that would result in “common coding” amongst within-category exemplars and “distinctive coding” amongst between-category exemplars. This can readily be seen by considering the input-hidden (or first layer, k=1) and hidden-output (or last layer, k=2) weight matrices (see **Figure 4**) that commonly emerges in a simple 8-4-2 connectionist network (see **Figure 1C)**.

The weights matrices in **Figure 4C** and **4D** illustrates a very typical solution to the category learning problem when 8 distinct exemplars are categorized into two categories. In essence, through the application of a standard backpropagation learning algorithm, each within-category exemplar forms strong excitatory connections to hidden layer units that themselves form strong connections with one of the category output units. In addition, exemplars within each category develop inhibitory connections to hidden units that participate excite the other category output unit. In essence, all within-category exemplars are “commonly coded” at the hidden layer which develops category representations that come to activate the appropriate output category response.

This sort of connectionist modelling approach has the potential to reveal the sorts of empirical results we report here. If within-category items are commonly coded by activating a similar set of hidden layer units, then learning a total reversal should be relatively easy because it merely would involve altering connections between the “category representation” and the output response. In contrast, a partial reversal task can only be solved once the hidden layer category representations have been realigned, since they no longer apply during the reversal phase. That deconstruction of the internal task representation should, intuitively, be more difficult to achieve.

To our surprise, this was not the case. Our simulations with the standard model (Delamater, 2012) revealed more rapid partial than total reversal learning (**Figure 4A**). Compare the weight matrices of the first later after learning and after reversals, these categories were completely mapped again for both tasks, total (**Figure 4C**) and partial (**Figure 4D**). Given that remapping the categories in the partial task require less matrix changes (only half of the exemplars) learning this task is quicker for the network (**Figure 4A**). However, particularly for the total task, this is not efficient because the network has already learned the categories after learning (**Figure 4C**). So, there should be no need to remap the categories in the first layer, but instead only change the output matrix layer preserving the already formed common coding. To achieve this, we implemented an additional mechanism to the networks.

One reason for faster remapping categories in the first layer (k=1) instead of the second (k=2) is that learning rates are the same for every layer within the network. Thus, when a category is already formed in first layer weights, it is easier to remap the weights in the first layer than only changing the hidden to output layer. To face this difference between networks and human participants, we modify the algorithms and incorporated two new rules in different layers. One rule Mackintosh-like controlled by and the second Pearce-Hall-like controlled by (**Figure 1C**). These parameters operate in the prediction error (), which is the average of absolutes prediction errors from the output layer (*i.e.*, k=2, hidden to output). Both parameters ( and ) control how the learning rate () changes each trial (Kaye & Pearce, 1984; Pearce et al., 1984). For the case of the first layer (k=1, input to hidden) changes:

(Eq. 1)

Thus, if , the higher the prediction error the higher . Thus, this rule implicitly encodes the Pearce-Hall intuition, where is larger when surprise is larger. In the case of the second layer (hidden to output), changes in the following way:

(Eq. 2)

where incorporates the Mackintoshian intuition (Mackintosh, 1975), where higher s is given to trials with less error, *i.e.*,when the world is quite predictive no surprise or small . By adding these dynamic learning rates controlled by different parameters (, and ; **Figure 1C**), then we can control the learning rates changes in each layer. In the case of the presented simulation (**Figure 4A**), the output later accumulates a higher given the small and predictable learning phase. Then when the reversal occurs the networks with a high in the hidden to output layer change faster than the weights than the ones in the input to hidden layer. This produced the results shown in **Figure 4E**, where the category representation is held for the total task, hence promotes faster learning. These results are dependent on the chosen parameters, particularly when . When both parameters are equal, the network learn faster the partial reversal condition. An interesting feature of this model is that a representation made in the first layer can be forced to be kept as a function of how quickly weights change in the last layer.

A screenshot of a graph

AI-generated content may be incorrect.

**Figure 4***. Simulation of the two conditions in the current Category Learning task.* ***A****. Model 1: backpropagation as is shown here cannot emulate findings with living organisms.* ***B*** *Model 2: almost the same algorithm but with changing learning rates. This shows faster learning to the total reversal condition.* ***C****,* ***D****,* ***E****, and* ***F*** *are the weight matrices at the end of each phase (learning or reversal), for both models (model 1:* ***C*** *and* ***D****; model 2 with dynamic learning rates:* ***E*** *and* ***F****) and for both conditions (total:* ***C*** *and* ***E****; partial:* ***D*** *and* ***F****).*

## Conclusions

Our study also contributes to the broader understanding of common coding as a general phenomenon in humans and neural networks. The findings from our experiments provide compelling evidence for a common coding phenomenon that emerges in category learning. This was evident given that participants learnt faster the task where all outcomes were inverted persevering the same mapping, in comparison the partial reversal task where half of the outcomes were reversed in a second phase forcing participant to remap the underlying mapping of the learnt categories. Across both the easy (Experiment 1a) and hard (Experiment 1b) version of our tasks, participants consistently showed better learning and faster reacquisition in total reversal conditions compared to partial reversals.

Our analysis through Multidimensional Scaling and psychological space measures revealed no significant differences in perceived similarities within and between categories used in experiments 1a and 1b. This suggests that the learning effects observed in experiments 1a and 1b were rooted in the associative mechanisms such as common coding rather than perceptual resemblances of the exemplars. The multidimensional space approach allowed us to map out and quantify the 'psychological' distances between exemplars, providing a robust framework to understand the cognitive processes underlying category learning.

The computational model we employed further supports our experimental findings. By incorporating dynamic learning rates inspired by the Macintosh (Mackintosh, 1975) and Pearce-Hall (Pearce et al., 1984) models, our artificial neural network was able to mimic the total reversal advantage seen in living organisms.

In conclusion, the research presented here advances our knowledge of category learning and common coding. By integrating behavioural experiments with computational modelling, we provide a comprehensive perspective on how associative mechanisms drive cognitive processes such as formation of categories. Future research could build on these findings by exploring the neural correlates of common coding.

# References

Bracci, S., & De Beeck, H. P. O. (2023). Understanding Human Object Vision: A Picture Is Worth a Thousand Representations. *Annual Review of Psychology*, *74*(Volume 74, 2023), 113–135. https://doi.org/10.1146/ANNUREV-PSYCH-032720-041031/CITE/REFWORKS

Castiello, S., Zhang, W., & Delamater, A. R. (2021). The retrosplenial cortex as a possible “sensory integration” area: A neural network modeling approach of the differential outcomes effect in negative patterning. *Neurobiology of Learning and Memory*, *185*, 107527. https://doi.org/10.1016/J.NLM.2021.107527

Delamater, A. R. (1998). Associative mediational processes in the acquired equivalence and distinctiveness of cues. *Journal of Experimental Psychology: Animal Behavior Processes*, *24*(4), 467–482. https://doi.org/10.1037/0097-7403.24.4.467

Delamater, A. R. (2012). On the nature of CS and US representations in Pavlovian learning. *Learning & Behavior*, *40*(1), 1–23. https://doi.org/10.3758/s13420-011-0036-4

Delamater, A. R., & Joseph, P. (2000). Common coding in symbolic matching tasks with humans: Training with a common consequence or antecedent. *Quarterly Journal of Experimental Psychology Section B: Comparative and Physiological Psychology*, *53*(3), 255–273. https://doi.org/10.1080/027249900411182,

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, *123*(2), 178–200. https://doi.org/10.1037//0096-3445.123.2.178,

Honey, R. C., & Hall, G. (1989). Acquired Equivalence and Distinctiveness of Cues. *Journal of Experimental Psychology: Animal Behavior Processes*, *15*(4), 338–346. https://doi.org/10.1037/0097-7403.15.4.338

Johns, K. W., & Williams, D. A. (1998). Acquired equivalence learning with antecedent and consequent unconditioned stimuli. *Journal of Experimental Psychology: Animal Behavior Processes*, *24*(1), 3–14. https://doi.org/10.1037/0097-7403.24.1.292

Kaye, H., & Pearce, J. M. (1984). The strength of the orienting response during Pavlovian conditioning. *Journal of Experimental Psychology: Animal Behavior Processes*, *10*(1), 90–109. https://doi.org/10.1037/0097-7403.10.1.90

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, *82*(4), 276–298. https://doi.org/10.1037/h0076778

Miller, N. E., & Dollard, J. (1941). *Social learning and imitation*. Yale University Press.

Pearce, J. M., Kaye, H., & Hall, G. (1984). Predictive accuracy and stimulus associability: Development of a model for Pavlovian learning. In M. Commons, R. Herrnstein, & A. R. Wagner (Eds.), *Quantitative analysis of behavior* (Vol. 3). Ballinger.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/S13428-018-01193-Y/FIGURES/3

R Core Team. (2021). *R A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. httpswww.R-project.org

Sidman, M. (2000). Equivalence relations and the reinforcement contingency. *Journal of the Experimental Analysis of Behavior*, *74*(1), 127. https://doi.org/10.1901/JEAB.2000.74-127

Urcuioli, P. J., Zentall, T. R., & DeMarse, T. (1995). Transfer to Derived Sample-comparison Relations by Pigeons Following Many-to-one versus One-to-many Matching with Identical Training Relations. *The Quarterly Journal of Experimental Psychology Section B*, *48*(2), 158–178. https://doi.org/10.1080/14640749508401445

Zentall, T. R., Sherburne, L. M., Steirn, J. N., Randall, C. K., Roper, K. L., & Urcuioli, P. J. (1992). Common coding in pigeons: Partial versus total reversals of one-to-many conditional discriminations. *Animal Learning & Behavior*, *20*(4), 373–381. https://doi.org/10.3758/BF03197960

Zentall, T. R., Steirn, J. N., Sherburne, L. M., & Urcuioli, P. J. (1991). Common Coding in Pigeons Assessed Through Partial Versus Total Reversals of Many-to-One Conditional and Simple Discriminations. *Journal of Experimental Psychology: Animal Behavior Processes*, *17*(2), 194–201. https://doi.org/10.1037/0097-7403.17.2.194

# Appendix

## *Connectionist Model*

The model was implemented with code written in R and is publicly available at the following website: <https://github.com/santiagocdo/ALANN> (check folder *training\_files*: *category\_total\_reversal.csv* and *category\_partial\_reversal.csv*). The connectionist model consists of a three-layer feedforward neural network, where activations go bottom-up (from CS to US representations) and learning top-down. The layers are: input (L = 0), hidden (L = 1) and output (L = 2). This model can be described with two rules: the activation rule (that calculates unit activations) and the learning rule (that determines changes in weights or connections between units).

The connectivity constraint between the input and the hidden layer given the visual, multimodal, and auditory pathways (Delamater, 2012) is encoded in the **C** matrix (size: number of input units x number of hidden units), where 1 represents an existent connection and 0 a non-existent connection; hence, connections between visual input units to auditory hidden layer pathway units will be equal to 0, and similarly for auditory inputs and visual pathways.

The model computes weight matrices for layers 1 and 2, i.e., between input and hidden (L = 1), and between hidden and output (L = 2), at every trial *t.* The values of at *t* = 0 are random variables from a uniform distribution between -0.5 and 0.5. In order to maintain the partial connectivity between input and hidden  is multiplied (via Hadamard product or elementwise multiplication) with **C**. On the one hand, maps the connections’ weights between input and hidden units, thereby, the matrix size is the number of inputs x number of hidden units. On the other hand, represents the weights between hidden and output units, hence, the matrix size is the number of hiddenx number of outputs. At every *t*, the CS information is encoded in the activation of the input layer, , where this vector contains 0s and 1s, representing presences and absence of CSs). The USs are encoded in a vector (size is equal to the number of USs)[[1]](#footnote-1).

*Activation Rule*. The first rule of this model specifies how units are activated. The activation of the hidden and output layers is given by a sigmoidal function that receives the net input (**n**) vector for the current layer L, i.e., :

(1).

The activations are a real number between 0 and 1. We subtract 2.2. from to shift the sigmoidal curve to the right and obtain low activation levels when elements is 0. The net input for the hidden layer is:

(2),

where is a matrix multiplication between a transpose () column vector (i.e., a row vector), and the matrix.

*Learning rule*. The second rule of the model specifies how changes at every *t*. This rule is an adaptation of backpropagation (Rumelhart, Hinton, & Williams, 1986). In general terms, the weights are the sum of the current weights and the change of those weights at *t*. Hence:

(3),

Where is the Hadamard product or elementwise multiplication, and the partial connectivity only applies to input layer (L = 0) to hidden layer (L = 1) connections within the weight matrix . This ensures that weights between units with impermissible connections (e.g., auditory inputs to visual pathway hidden units) are fixed at 0. Changes in weights are determined by:

(4),

where and are free parameters, representing the learning rate (assumed to be 0.3) and the momentum decay (assumed to be 0.9), respectively. Finally, is a vector of delta values, and encodes the prediction errors that are back-propagated throughout the network, following the next equation:

(5),

where is a vector which encodes with 0s (absent) and 1s (present) the USs at every *t*.

## Instructions

### Experiment 1a and 1b

In this experiment, you will be presented with a series of abstract images that represent the molecular structure of various natural objects. Your task will be to learn from which of two regions in the world these objects come (Northern Hemisphere or Southern Hemisphere).

More specifically, you will see 1 of 8 different abstract images at a time and be asked to indicate whether you think that image reflects an object taken from the Northern or Southern Hemisphere. Choose the Left Arrow Key for Northern or the Right Arrow Key for Southern Hemisphere. At first, you will need to guess, but you will be provided with feedback after your answer to help you learn which objects come from Northern or Southern Hemispheres.

Your response times are also important. Please make your response choices as quickly, but also as accurately, as you can. Your feedback will display the time (in sec) that it took for you to reach your decision, and also if your choice was correct (with a high pitch sound) or not (low pitch sound).

There will be a break halfway through.

Press the space bar when you are ready to begin.

### Experiment 2

[categorySimilarity]

In this experiment, you will be shown pairs of abstract images. Your task is to rate how similar or different the two images appear to you.

The experiment consists of two parts. In the first part, you will see and rate 56 pairs of images, and in the second part, another 56 pairs. Below each pair, there will be a slider ranging from "very different" to "very similar." Use your mouse to select a position on the slider that best represents how similar or different you find the two images are relative to one another.

There will be a break between both parts.

Press the space bar when you are ready to begin.

[fractal\_similarity]

In this experiment, you will be shown pairs of abstract images. Your task is to rate how similar or different each pair appears to you.

The experiment consists of two parts. In the first part, you will see 56 pairs of images, and in the second part, another 56 pairs, individual images may repeat between pairs. Below each pair, there will be a slider ranging from "very different" to "very similar." Use your mouse to select a position on the slider that best represents how similar or dissimilar you find the images.

There will be a break between both parts.

Press the space bar when you are ready to begin.

## Informed Consent

This research investigates the psychological processes used when people learn to identify objects in the world. You are being asked to participate in this research study because you are a normal healthy adult, and we wish to better understand basic learning processes in your population. The purpose of this research is to gain more knowledge about the cognitive processes involved in simple forms of associative learning.

If you agree to participate, we will ask you to perform in a simple computer task that will last approximately 12 minutes. In this task, you will see a series of abstract images presented individually on the screen and your task will be to learn to choose one of two response options for each image. Also, you will be asked to respond quickly and accurately on your computer keyboard when the image appears.

• Risks/Discomforts: There are no risks for participating in this study beyond those associated with normal computer use over a 15 min period.

• Benefits: This research is not designed to directly benefit you, but your help with this study will advance basic science on the cognitive processes involved in predictive learning in normal healthy individuals. Ultimately, this research could lead to a better understanding of some of the associative learning processes that are negatively impacted by various psychological conditions (such as aging, dementia, etc).

• Confidentiality: This study does not collect identifying information, and all data collected will remain anonymous. We will ask about your gender, age, and nationality, and we will record your performance in the task itself. However, this information will not be linked directly to any individual participant. The data we obtain will be stored indefinitely and may be shared publicly via online repositories for findings that are ultimately published.

Your participation in this research is completely voluntary, and you will be able to stop at any time without penalty. If you have any questions, you can contact: Andrew R. Delamater (andrewd@brooklyn.cuny.edu). If you have any questions about your rights as a research participant or if you would like to talk to someone other than the researcher, you can contact CUNY Research Compliance Administrator at 646-664-8918 or HRPP@cuny.edu.

If you wish to participate in the study, please press the spacebar for additional instructions.

# Supplementary Information

## Experiment 1a: Total Reversal versus Partial Non-reversal

We ran the same LogMM with condition (total reversal versus partial nonreversal), Blocks, and its interaction. We found no main effects in condition [] nor in the Interaction []. But a positive and significant effect on Blocks [].

## Experiment 1a: Sensitivity Analysis of the Reaction Time

*RT transformed with Natural Logarithm*. We ran the same LMM with condition (total versus partial), Blocks, and its interaction. We found no main effects in Total-Condition [] nor Blocks [], and no Interaction [].

*RT transformed with Root Square*. We ran the same LMM with condition (total versus partial), Blocks, and its interaction. We found no main effects in Total-Condition [] nor Blocks [], and no Interaction [].

## Experiment 1b: Total Reversal versus Partial Non-reversal

We ran the same LogMM with condition (total reversal versus partial nonreversal), Blocks, and its interaction. We found no main effects in condition [] nor in the Interaction []. But a positive and significant effect on Blocks [].

## Experiment 1b: Sensitivity Analysis of the Reaction Time

*RT transformed with Natural Logarithm*. We ran the same LMM with condition (total versus partial), Blocks, and its interaction. We found the same main significant effects in Total-Condition [] nor Blocks [], and no Interaction [].

*RT transformed with Root Square*. We ran the same LMM with condition (total versus partial), Blocks, and its interaction. We found no main effects in Total-Condition [] nor Blocks [], and no Interaction [].

A group of graphs showing different types of distance

AI-generated content may be incorrect.

**Figure S1**. *Category Similarity Analysis, using three psychological dimensions (multidimensional scaling with k=3) in the psychological space.* *Left panels (****A*** *and* ***C****) represent Experiment 1a results for the exemplars used in Task 1 (could be either total or partial, they were counterbalanced), and the right panels (****B*** *and* ***D****) Experiment 1b.*

1. All vectors are considered to be column vectors. Hence to get a row vector we transpose them, e.g. **x** (column vector), **x**T (row vector). [↑](#footnote-ref-1)