COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 49 (2025) e70070

ISSN: 1551-6709 online
DOI: 10.1111/cogs.70070

Examining the Relationship Between Early Experience, Selective Attention, and the Formation of Learning Traps

Yanjun Liu, Ben R. Newell, Jaimie E. Lee, Brett K. Hayes

School of Psychology, University of New South Wales

Received 5 October 2024; received in revised form 30 April 2025; accepted 12 May 2025

Abstract

A simple-rule learning trap occurs when people show suboptimal category learning due to insufficient exploration of the learning environment. By combining experimental methods and computational modeling, the current study investigated the impact of two key factors believed to play essential roles in the development of a simple-rule learning trap: early learning experience and selective attention. Our results showed that, in a learning environment where the true category mapping was determined by conjunctions of two predictive dimensions, the likelihood of falling into a single-dimensional learning trap increased when early learning experience involved a large loss that could be predicted from a single feature dimension. In addition, using a model-based measurement of attention bias, we observed that early experience affected trap formation by narrowing the distribution of attention to exemplar features. These findings provide the first direct empirical evidence of how early learning experience shapes the formation of a simple-rule learning trap, as well as a more granular understanding of the role of selective attention and its interaction with early learning experience in trap formation.

Keywords: Selective attention; Category learning; Exemplar models

1. Introduction

From everyday decisions like grocery shopping to significant financial investments, we often rely on beliefs acquired through prior experience to guide our choices. For instance,

Correspondence should be sent to Yanjun Liu or Brett Hayes, School of Psychology, University of New South Wales, Kensington, NSW 2052, Australia. E-mail: yanjun031130@gmail.com; b.hayes@unsw.edu.au

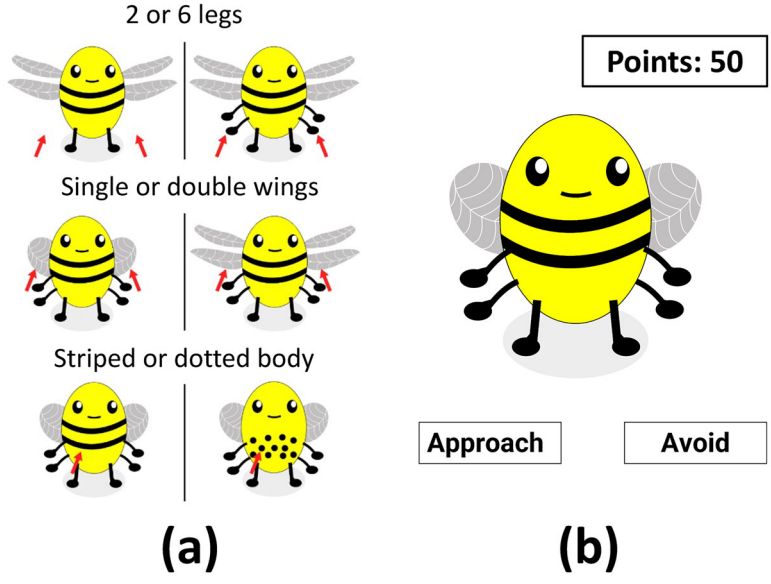

Fig. 1. Stimulus design. Panel (a): Exemplar binary-valued bee features, including numbers of legs (two vs. six), numbers of wing pairs (single vs. double), and body patterns (striped vs. dotted). Two feature dimensions were randomly selected as relevant features for category membership for each participant, and one was irrelevant. Panel (b): An example choice screenshot where participants make an approach/avoid decision on the present bee by clicking on the respective alternate buttons at the bottom of the screen. The total amount of accumulated points is displayed at the top-right of the screen.

when engaging in stock trading, we may reflect on our past experiences with certain types of stocks to determine if we would like to buy them again. Regardless of our desire to know everything everywhere all at once, decision-makers are inevitably constrained by the exploration-exploitation dilemma (Sutton & Barto, 2018) when making decisions from experience (Hertwig, Barron, Weber, & Erev, 2004; Hills, 2006; Mehlhorn et al., 2015). To maximize overall rewards, we need to balance the potential gain from learning new knowledge with the benefits from using existing knowledge.

Selective attention plays an essential role in addressing the exploration-exploitation dilemma (Niv et al., 2015). It allows decision-makers to focus on relevant information while filtering out irrelevant details (Broadbent, 2013). Nevertheless, this seemingly efficient mechanism can sometimes result in suboptimal decisions. Rich and Gureckis (2018) documented a *simple-rule learning trap* in environments where stimuli (e.g., cartoon bees) were composed of multiple binary-valued feature dimensions (see examples in Fig. 1). A conjunction of features on two of these dimensions determined the stimulus category (i.e., safe or dangerous bees in Fig. 2a), and different categories were associated with different reward structures (i.e., approaching safe bees yielding a positive return, while approaching dangerous bees yielding a loss; avoiding either yielding zero return).

When the category membership of the stimulus was informed after each response (i.e., full-feedback category learning), participants readily learned the conjunctive category

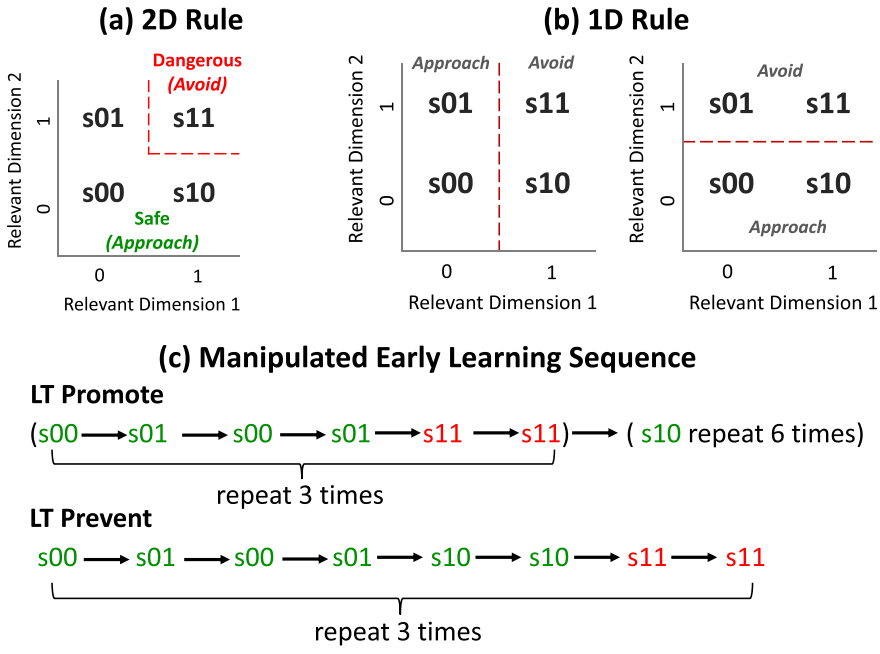

Fig. 2. Category rules and key manipulations of early learning experience. Panel (a): Schematic optimal two-dimensional category rule of different bee types (i.e., s00, s01, s10, s11) based on relevant dimensions. 0 denotes a safe feature and 1 denotes a potentially dangerous feature. Combinations of these features determine four bee types. The red dashed line notates the conjunctive category bound. The bee type composed of two potentially dangerous features (i.e., s11) is dangerous and should be avoided. All other bee types are safe to approach. (b): Schematic one-dimensional category rule on the basis of either relevant dimension 1 (left plot) or 2 (right plot). The dark-red lines notate the suboptimal one-dimensional category bound. (c): The manipulated order of early learning sequence (i.e., bee items shown in the first 24 trials of the experiment). In the Learning Trap (LT) Promote condition, participants encounter only one type of the ambiguously safe bee types (i.e., s01) before encountering a dangerous bee. In the LT Prevent condition, participants encounter all safe bee types before encountering a dangerous bee.

structure. However, when corrective feedback was available only after a stimulus was approached (i.e., contingent-feedback), many learners ended up falling into a simple-rule trap; they classified stimuli based on a single predictive dimension (1D rule, see Fig. 2b) rather than both dimensions (2D rule, see Fig. 2a). The use of a 1D rule resulted in a suboptimal total return because learners missed opportunities of exploring options that could have yielded rewards.

Outside the laboratory, feedback is often decision-contingent—available only when a decision-option is chosen but not when it is avoided (Denrell, 2005; Elwin, Juslin, Olsson, & Enkvist, 2007). Hence, the simple-rule trap has been shown to arise in a wide range of environments, such as the development of negative stereotypes of out-groups (Denrell, 2005) and in psychopathologies like depression, both are characterized by avoidance of potentially rewarding experiences (Teodorescu & Erev, 2014). Moreover, the trap persists even when

learners have many opportunities to explore choice options that would disconfirm the one-dimensional rule, leading to a substantial reduction in earned rewards (Lee, Li, Lee, & Hayes, 2024; Liu, Newell, Lee, & Hayes, 2024; Li, Gureckis, & Hayes, 2021; Rich & Gureckis, 2018). Additionally, learners who fall into the simple-rule learning trap are also often blind to dynamic changes in the reward structure (Blanco, Turner, & Sloutsky, 2023; Lee et al., 2024). Understanding the processes that drive trap formation can help us to identify remedies for the negative costs it incurs.

Rich and Gureckis (2018) outline a descriptive account of how a simple-rule learning trap could emerge. According to their account, early positive outcomes following the approach of a subset of stimuli with the salient feature may lead to the false belief that values on this single dimension are predictive of all choice outcomes. This preference for simple over complex rules may reflect a general inductive bias (e.g., Chater & Vitányi, 2003; Feldman, 2016; Pothos & Close, 2008). For instance, let us assume that "safe" bees are identified by a rule involving values on two dimensions: body pattern and leg number. If learners receive positive feedback from approaching bees with a predictive body feature early in their learning, they may falsely believe that body pattern is sufficient to predict the outcomes for all bee types. This could result in increasing attention to body features and decreased attention to other predictive features (e.g., leg number in this example). Consequently, learners fall into a simple-rule learning trap by avoiding stimuli without the salient body feature in subsequent trials, even though some of them are rewarding.

1.1. Testing the descriptive account of learning traps

With a novel empirical manipulation and a new model-based analysis, the current work aimed to test key assumptions of Rich and Gureckis (2018)'s account concerning: (1) the role of early experience in the development of a simple-rule learning trap, and (2) how such experience interacts with selective attention to stimulus features in trap formation.

To this end, we conducted experiments with a "honey harvesting" task similar to that used by Rich and Gureckis (2018), where bee categories were predicted by conjunctions of two relevant features. The key empirical manipulation to test the first assumption is the early learning sequence that determines when participants first encounter stimuli yielding a large loss. We hypothesized that when a dangerous bee had appeared before learners experienced all types of safe bees (i.e., Learning Trap [LT] Promote condition, see top row in Fig. 2c), it would bias learners' attention toward a single dimension and result in a higher likelihood of falling into a simple-rule learning trap. In contrast, when all types of "safe" conjunctions appeared before a large loss (LT Prevent condition, see bottom row in Fig. 2c), this might reduce the prevalence of 1D-rule users because learners were more likely to receive corrective category feedback that disconfirmed a single-dimension rule.

The design of our LT promote sequence was inspired, in part, by previous work showing that "blocked" presentation of instances from contrasting categories promotes learning of the features that differentiate those categories (e.g., Carvalho & Goldstone, 2017, 2022). In our case, the LT promote sequence was intended to draw attention to a single feature dimension that differentiated safe from dangerous bees. In a similar vein, blocked presentation of

instances that conform to the same logical rule (e.g., a 1D dimensional rule in our LT promote sequence) can promote rule learning (Mathy & Feldman, 2009). By contrast, our LT Prevent sequence explicitly avoided blocking of stimuli based on a single dimension. The early trials in this sequence were intended to show that the features of safe bees could vary across two dimensions.

Our predictions about early learning experience apply primarily to learning environments with contingent feedback, where corrective category feedback only appears after approaching decisions. If feedback is provided on every learning trial (i.e., full feedback), learning traps are less likely to form. This is because full feedback could provide immediate disconfirmation of simple 1D rules (Lee et al., 2024; Li et al., 2021; Rich & Gureckis, 2018), which would help people escape from simple-rule learning traps, regardless of the early learning sequence. We conducted experiments with both full and contingent feedback to test these predictions.

The second key assumption of Rich and Gureckis (2018)'s account is that early learning experience may change attention distribution among feature dimensions during category learning. Preliminary evidence has been shown to support this assumption. In a dynamic category learning environment where an unannounced switch of predictive feature dimensions occurred, Blanco et al. (2023) found that children who distributed attention among stimulus features broadly early in learning were less likely to miss this important switch and fell into learning traps, compared to adults whose early attention distribution was much narrower during learning.

The current work goes beyond Blanco et al. (2023) by testing the interaction between early experience and selective attention with explicit manipulation of initial learning sequences. In addition, building on Rich and Gureckis (2018)'s modeling work, we derived a model-based measurement to assess changes in attention distribution with manipulated early experience. We predicted that our LT Promote condition would strengthen attentional bias toward a single relevant feature, while the LT Prevent condition should reduce this bias.

The data collected for all of the experiments are available on the Open Science Framework at: https://osf.io/pbf8y/. All of the experiments were conducted under human research ethics #3333 approved by the institutional review board of the University of New South Wales.

2. Method

2.1. Participants

For each of the six conditions (three early learning sequences with either full or contingent feedback), we intended to recruit 75[1] participants. All participants were recruited through the Prolific online platform, with inclusion criteria of (1) fluent in English, (2) aged between 18 and 80, and (3) have completed at least 10 prolific tasks with an approval rate of at least 90% or higher. Informed consent was obtained for each participant. They were paid with a £2 base rate/15 min plus a performance-based bonus ranging from £0 to £1.34. The amount of the bonus was determined by the points participants accumulated throughout the experiment. Each point was worth £0.01.

No participant was excluded from the data analyses. In total, with full feedback, we had 75 participants in the baseline (26 women, 48 men, 1 nonbinary; age: $M = 35.72$, $SD = 12.38$), 80 in the LT Prevent condition (35 women, 45 men; age: $M = 38.73$, $SD = 13.19$), and 75 in the LT Promote condition (26 women, 45 men, and 4 nonbinary; age: $M = 35.72$, $SD = 11.82$). With contingent feedback, we had 75 participants in the baseline condition (26 women, 48 men, 1 nonbinary; age: $M = 35.72$, $SD = 12.38$), 76 in the LT Prevent condition (24 women, 51 men, 1 nonbinary; age: $M = 35.11$, $SD = 13.23$), and 75 in the LT Promote condition (24 women, 49 men, 2 nonbinary; age: $M = 37.32$, $SD = 12.80$).

2.2. Materials and procedure

Factorial combination of three binary-valued bee feature dimensions (Fig. 1a), including numbers of legs (two vs. six), numbers of wing pairs (single vs. double), and body pattern (striped vs. dotted), generated eight unique stimuli. At the beginning of the experiment, two out of these three dimensions were randomly assigned as relevant to predicting category membership with the third dimension irrelevant for prediction. For each relevant dimension, one feature value was randomly assigned as the potentially dangerous feature (denoted by 1), and the other feature value was assigned as the safe feature (denoted by 0).

Bee category membership was determined by the conjunction of feature values on the two relevant dimensions (Fig. 2a). The stimuli with at least one safe feature belonged to the safe category (i.e., s00, s01, and s10). The stimuli with both potentially dangerous features (s11) belonged to the dangerous category. For example, if the two potentially dangerous features were "two legs" and "dotted body" (in this case, the relevant dimensions are numbers of leg and body pattern, the irrelevant dimension is number of wings), then the bees with six legs and striped body (s00), the bees with six legs and dotted body (s01), and the bees with two legs and striped body (s10) were safe. The bees with two legs and dotted body (s11) were dangerous. Since half of those bees have single wings and the rest have double wings, number of wings is not predictive of stimulus category.

The experiment followed a 2 (feedback scheme: full, contingent) $\times$ 3 (early learning sequence: baseline, LT Prevent, LT Promote) between-subjects design. Participants were randomly assigned to one of six conditions. They were given the role of beekeeper with the goal of maximizing the honey harvested from bees (quantified as points). Some bees were safe to approach, yielding honey valued at $+1$ points; while some bees were dangerous, causing a large loss of -3 points if approached. Avoiding a bee resulted in no gain or loss (0 points).

During the instruction session, all three binary-valued feature dimensions were shown to participants as depicted in Fig. 1a. Participants were informed that perfect prediction of bee categories was possible based on feature combinations and that the features associated with each type would not change during the task. To ensure participants' understanding of the task, they had to achieve a perfect score on a comprehension survey that queried: (1) all three feature dimensions that would vary; (2) the task goals; and (3) the consequences of avoiding a bee, before proceeding to the experimental session.

Participants then completed six blocks of 16 learning trials followed by a single block of 16 test trials.[2] In each learning trial, a bee stimulus appeared on the screen for participants to

make an approach/avoid decision without a time limit, using on-screen buttons (Fig. 1b). If a learner clicked on the approach button, they were informed the true category of the bee and received the appropriate payoff. Corrective category feedback also appeared after an avoid response in the full-feedback conditions, but was omitted in the contingent-feedback conditions. After each response, a feedback screen was displayed for 2 s followed by a 500-ms interval between trials. Test trials were identical to learning trials, except that no feedback was provided after a response.

The order of the first 24 learning trials determined three early learning sequence conditions (Fig. 2c). In the LT Promote condition, the sequence of stimuli across the first 24 trials was $s00 \rightarrow s01 \rightarrow s00 \rightarrow s01 \rightarrow s11 \rightarrow s11$ for three times, and then followed by $s10$ for six times. In the LT Prevent condition, the corresponding sequence was $s00 \rightarrow s01 \rightarrow s00 \rightarrow s01 \rightarrow s10 \rightarrow s10 \rightarrow s11 \rightarrow s11$ repeated three times. In other words, participants in the LT Promote condition had the opportunity to experience only one type of the ambiguously safe bee types (i.e., s01) before encountering a dangerous bee so that a single dimension appeared predictive of stimulus category in early learning sequences. In contrast, participants in the LT Prevent condition had the opportunity to experience all safe bee types before encountering a dangerous bee. After the first 24 trials, the eight unique stimuli appeared once for the following eight trials in random order. Stimulus order was pseudo-randomized for the rest of the experiment so that each unique bee stimulus appeared twice in every block. To assess the impact of early experience on trap formation, we also conducted a baseline condition where the presentation order of stimuli within blocks was by-subject pseudo-randomized throughout the experiment.

Participants commenced the task with an endowment of 50 points. Points were accumulated across trials and shown as a tally at the top-right of the screen in learning trials. In test trials, points continued to be earned but the tally was not shown. On completion, participants received a bonus payment based on their final tally.

The experiment was programmed in jsPsych (de Leeuw, 2015). R (R Core Team, 2023) was used for model fitting and Jamovi (jamovi, 2025) for statistical analyses.

3. Results

3.1. Learning of category strategy

To examine the effect of early experience on trap formation, we assessed changes in the prevalence of different category rules across the six learning and one test block as a function of manipulated early learning sequences and feedback. We identified participants' category rules used in each block based on their patterns of approach and avoid decisions (cf. Rich & Gureckis, 2018). Hypothetically, if participants based their decisions on the optimal 2D rule (Fig. 2a), they should approach all safe bees s00, s01, and s10, and avoid the dangerous bees s11. In contrast, if participants relied on 1D rules (Fig. 2b), they would still approach s00 and avoid s11, but incorrectly avoid one of the ambiguous bee types (either s01 or s10).

A participant was deemed to use a particular category rule within a block if their choice patterns were consistent with the rule prediction on at least 15 out of 16 trials (Rich & Gureckis,

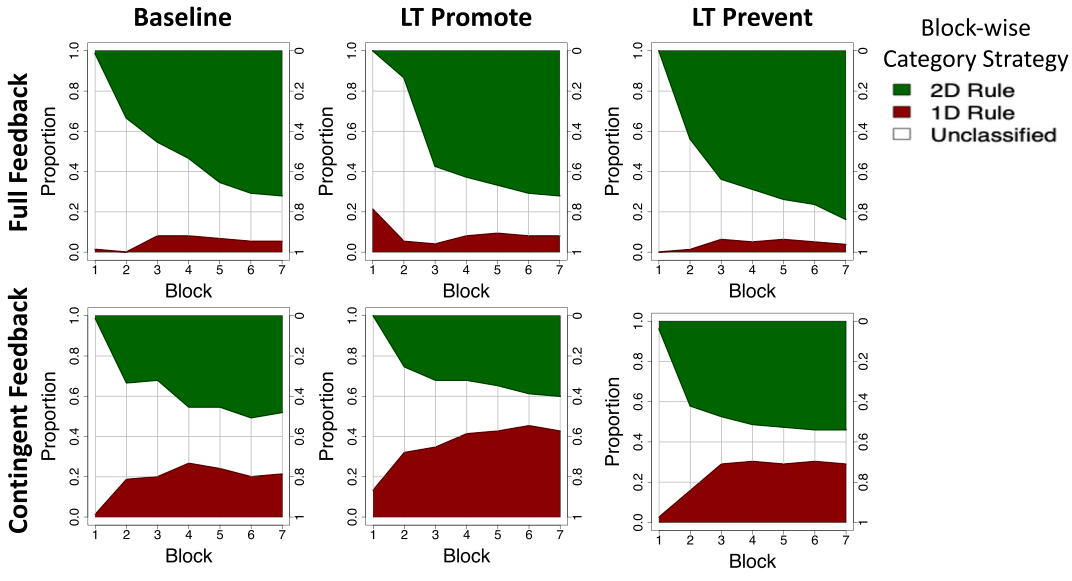

Fig. 3. Prevalence of different rule users within each block. Top Row: Proportions of different category-strategy users across blocks in each early experience condition with full category feedback. Bottom Row: Proportions of different category-strategy users across blocks in early experience conditions with contingent category feedback. Blocks 1–6 are learning blocks, and Block 7 is the test block. The left y-axis denotes the proportion for 1D-rule users, and the right y-axis denotes the proportion for 2D-rule users.

2018). The use of 1D rules based on either relevant dimension 1 or 2 (Fig. 2b) were combined for analyses. Responses to bees composed of the same relevant features but different irrelevant features were pooled for analyses. If participants' choice behavior within a block was not consistent with either a 1D or 2D rule, they were deemed to be using an unclassified rule. Given our interest in the formation of learning traps, we mainly focused on results for 1D- and 2D-rule users. We also created three subgroups of learners (1D-rule learners, 2D-rule learners, unclassified rule learners) based on the type of rule that they used most frequently across blocks. An assessment of accumulated points confirmed that the 2D rule learners earned more points (with full feedback: $M = 122$, $SD = 7.54$; with contingent feedback: $M = 123$, $SD = 6.85$) than those using a 1D (with full feedback: $M = 102$, $SD = 6.64$; with contingent feedback: $M = 97.8$, $SD = 5.30$) or an unclassified rule (with full feedback: $M = 84.5$, $SD = 18$; with contingent feedback: $M = 84.6$, $SD = 17.7$), $F\ (2, 450) = 478.1$, $p < .001$.

The top row of Fig. 3 shows that, when full feedback was provided, most participants learned the optimal 2D rule in all the early experience conditions. Only a few ($< 10\%$) fell into the trap of using a 1D rule. In comparison, the prevalence of 1D-rule users increased, while that of 2D-rule users decreased in the contingent-feedback conditions (the bottom row in Fig. 3). In contrast, under contingent feedback, the prevalence of different rule users differed across early experience conditions. Compared to the baseline condition, the use of a 1D rule in test trials (i.e., Block 7) increased in the LT Promote condition (see Table 1), along with a

Table 1
Prevalence of different rule users in Test block (Block 7) across Feedback × Early Sequence Order conditions

Feedback	Early Sequence Order	2D	1D	Unclassified
Full	Baseline	72%	5.3%	22.7%
	LT Promote	72%	8%	20%
	LT Prevent	83.8%	3.7%	12.5%
Contingent	Baseline	48%	21.3%	30.7%
	LT Promote	40%	42.7%	17.3%
	LT Prevent	54%	28.9%	17.1%

Table 2
Results of the omnibus tests for the most parsimonious multinomial logistic regression predicting odds of using different strategies (1D, 2D, or unclassified) by Feedback (Full, Contingent), Early Sequence Order (baseline, LT Promote, or LT Prevent), and Block (ordinal, 1–7)

Predictor	$\chi 2$	df	p-value
Feedback	1.430	2	.489
Early Sequence Order	40.180	4	$< .001$
Block	217.790	12	$< .001$
Early Sequence Order × Block	62.200	24	$< .001$
Feedback × Block	21.120	12	.049

Note. Estimated coefficients are summarized in the Supplementary Materials (Table S1).

decrease in the use of a 2D rule. However, the prevalence of both 1D- and 2D-rule users in Block 7 remained at a similar level between the LT Prevent condition and the baseline.

To statistically test the effects of our group manipulations on the change in likelihoods of using different category rules (i.e., 1D, 2D, unclassified) across blocks (as depicted in Fig. 3), we conducted a multinomial logistic regression analysis. We used a backward selection based on likelihood ratio tests (Li et al., 2021; Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017) to determine the most parsimonious multinomial logistic regression predicting the odds of using different rules by feedback conditions, orders of sequence in early learning, and blocks. The results of omnibus tests for each term included in the most parsimonious model ($R^2_{McF}{}^3 = .202$) are summarized in Table 2, corroborating the significance of the main effects of early experience and blocks, their interaction term, and the interaction between feedback and block on the odds of learning different category strategies. The estimated coefficients are summarized in the Supplementary Materials (see Table S1), which confirmed that, compared to the baseline condition, the odds of using a 1D rule were higher, while the odds of using the 2D rule were lower across blocks in the LT Promote condition. On the other hand, the odds of using either 1D or 2D rule remained the same between the baseline and the LT Prevent conditions. These results indicate that early learning experience plays an essential role in the development of a 1D learning trap.

We also assessed changes in choice patterns of different rule users across blocks (see the Supplementary Materials for details). The results confirmed that the manipulated early

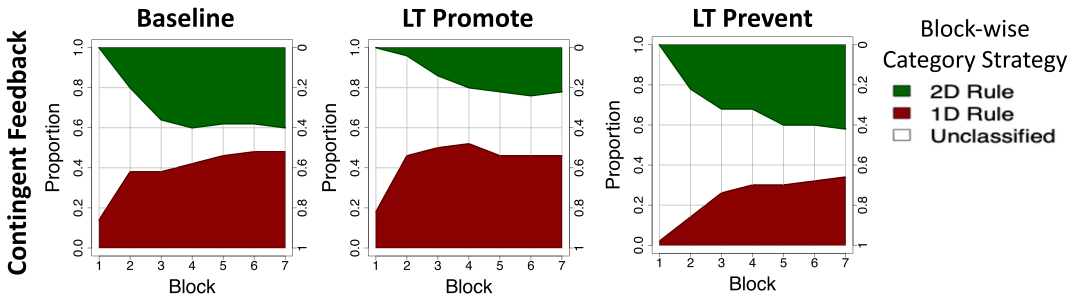

Fig. 4. Prevalence of different rule users across blocks in the replication study. Proportions of different category-strategy users across blocks in early experience conditions (left: LT Promote; right: LT Prevent) with contingent category feedback. Blocks 1–6 are learning blocks, and Block 7 is the test block. The left y-axis denotes the proportion for 1D-rule users, and the right y-axis denotes the proportion for 2D-rule users.

learning sequence in the LT Promote condition increased the odds of using a *particular* 1D rule relying on the single relevant dimension that was meant to be biased toward (i.e., relevant dimension 1, as s11 appeared before s10 in early learning trials of the LT Promote condition; also see Fig. 2b,c). This suggests that early experience could shape how a learning trap is formed.

To examine the robustness of these results, we conducted a conceptual replication study of the three early experience conditions with contingent feedback, using a set of more abstract visual stimuli (see the Supplementary Materials) with the same underlying category structure as our original study (Fig. 1a). Experimental procedures were identical to the original contingent feedback conditions. Participants found the new set of features more difficult to learn compared to those used in our original honey harvester task. The baseline condition in the conceptual replication showed an increase in 1D rule use and a decrease in 2D rule use compared to the original baseline. Nevertheless, the key qualitative patterns in the early learning sequence conditions were replicated (Fig. 4). The LT Promote manipulation once again encouraged trap formation, as evidenced by an increase in 1D rule use and a decrease in 2D rule use compared to the baseline. The LT Prevent condition again failed to increase the use of the optimal 2D rule. However, unlike the original study, a reduction in 1D rule use was observed in the LT Prevent condition.

Contrary to expectations, the early learning sequence in the LT Prevent conditions of the main study and conceptual replication did not increase the prevalence of 2D-rule use. One possible reason for this result is that the early sequence manipulated in this condition might not be optimal for drawing attention to the two *relevant* dimensions for categorization. As suggested in previous studies (Lejarraga, Schulte-Mecklenbeck, Pachur, & Hertwig, 2019), people tend to deploy more attentional resources in the presence of losses than gains. In the LT Prevent condition, the initial items in the early learning sequence (i.e., $s00 \rightarrow s01 \rightarrow s00 \rightarrow s01$) always yielded positive rewards for approach responses, which might not induce the deployment of sufficient attention resources to learn the optimal category rule. Relatedly, the absence of instances from the contrast category (i.e., dangerous bees) in early trials may have

made it more difficult for participants to learn the features that differentiated the categories (cf. Carvalho & Goldstone, 2017).

Hence, we trialed an alternative version of the LT Prevent condition where participants still had substantial opportunities to discover all safe conjunctions early in learning, yet dangerous bees (s11) appeared more frequently (see details of the follow-up LT Prevent condition in the Supplementary Materials). The observed prevalence of 2D-rule learners in this follow-up study remained similar to that of the baseline condition, suggesting that stronger interventions might be needed to redirect those who are initially prone to be trapped by the simple rule in the current multi-feature category learning paradigm.

3.2. The relationship between selective attention and category learning

Building on Rich and Gureckis (2018)'s modeling work, we investigated how early experience interacted with the dimension-wise distribution of attention in the formation of learning traps. Rich and Gureckis (2018) extended Kruschke (1992)'s well-known exemplar-based category learning model (ALCOVE) to account for reinforcement learning from response-contingent feedback in multi-feature categorization (such as the contingent-feedback conditions in the current study). This ALCOVE-Reinforcement Learning (RL) model was able to simulate the development of a simple-rule learning trap by imposing uneven attention weights among feature dimensions. We built upon this modeling work in two ways. First, we went beyond simulation by fitting ALCOVE-RL to participant learning data. Second, we used ALCOVE-RL as a measurement model to derive trial-by-trial model-based estimates of selective attention for individual learners and examined how these estimates were affected by our learning sequence manipulation.

Like many other category-learning models (e.g., Galdo, Weichart, Sloutsky, & Turner, 2022; Kruschke, 1992; Love, Medin, & Gureckis, 2004; Nosofsky, 1986; Weichart, Galdo, Sloutsky, & Turner, 2022), ALCOVE-RL instantiates the contribution of stimulus input to activation of an exemplar in psychological space through dimensional attention weights (e.g., α in Eq. 1 in the Supplementary Materials). Uneven attention weights reflect selective attention to different input dimensions.[4] To examine changes in attention bias with manipulated early learning sequences and feedback, we derived trial-by-trial attention weights using the best-fitted ALCOVE-RL parameters and quantified the degree of *selective attention to the relevant dimensions* by taking the absolute difference in attention weights for relevant dimensions (i.e., normalized $|\alpha_1 - \alpha_2|$; see detailed explanation in the Supplementary Materials). A larger difference indicates a stronger attention bias toward a single dimension.

Below, we highlight the key findings of this model-fitting. Details of model description, model-fitting procedures, and results are in the Supplementary Materials. Qualitatively, the model captured the distinctive choice patterns of 1D- and 2D-rule learners (i.e., 2D-rule learners were similarly likely to approach all safe bees, while 1D-rule learners were less likely to approach s01 or s10 as compared to s00). Quantitatively, the choice proportions predicted by the model were positively correlated to the observed data, $r\ (905) = .590$, $p < .001$ (see Supplementary Materials for a detailed discussion on model-fitting evaluation). We view this fit as sufficient to use ALCOVE-RL as a measurement model for the

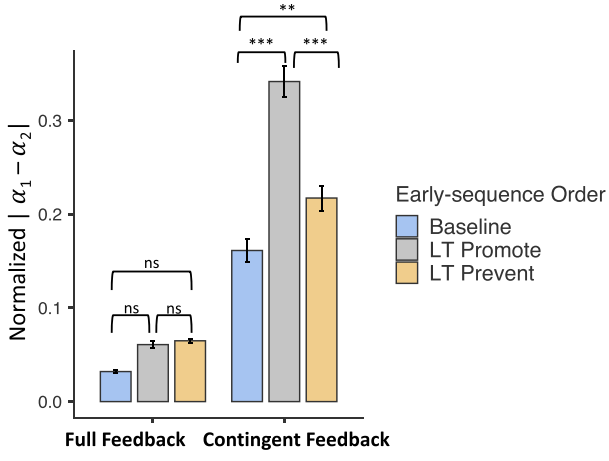

Fig. 5. Degree of attention bias in each condition. Selective attention to relevant dimensions for each individual was measured by averaging the absolute difference in normalized attention weights between relevant dimensions across trials. α_1 and α_2 are the respective attention weights for relevant dimensions 1 and 2. Since ALCOVE-RL does not impose any bounds for attention weights, we normalized attention weights by dividing them by the total sum of weights in each trial for each participant for comparison. Error bars denote the standard error of the mean. Brackets denote post-hoc pairwise comparisons between early sequence conditions with either full or contingent feedback, using Tukey-adjusted p-values. ***: $p < .001$, **: $.001 < p \leq .01$, ns: not significant.

investigation of the interaction between early experience and attention distribution at the trap formation.

Our key question addressed by model-fitting analyses was how early experience affected selective attention. Fig. 5 plots the degree of attentional bias in each condition and shows that the attentional bias to a single dimension was much higher in the LT Promote condition than either the baseline or the LT Prevent condition with contingent feedback. In contrast, this bias remained at a similar level across early experience conditions with full feedback.

A 2(feedback) × 3(early learning) analysis of the attention-bias measure confirmed that the bias was stronger in contingent-feedback than in full-feedback conditions, F (1, 3186) = 537.1, $p < .001$, $\omega^2 = 0.137$. There was also a significant main effect of early learning sequence, F (2, 3186) = 55.4, $p < .001$, $\omega^2 = 0.028$, and an interaction between learning sequence and feedback, F (2, 3186) = 33.7, $p < .001$, $\omega^2 = 0.017$. Fig. 5 indicates that this interaction was primarily driven by the accentuated attention bias in the LT Promote condition with contingent feedback.

4. General discussion

The current work examined the role of early experience and its interaction with selective attention in the development of a simple-rule category learning trap. Our results confirmed Rich and Gureckis's (2018) hypothesis that the experience of a loss early in a learning

sequence would promote trap formation. When a single feature dimension appeared to be predictive of category instances associated with losses or rewards in early learning (LT Promote condition), participants were more likely to focus on that dimension and less likely to discover a more complex, optimal category rule. Those who fell into the simple-rule trap avoided rewarding stimuli, earning substantially less over the course of learning than those who discovered the optimal rule. Once established, the trap persisted despite further opportunities to explore other stimuli. Our model-based analysis revealed that early experience affected participants' attention as well as their behavior. Learners were more likely to restrict their attention to a single relevant dimension in the LT Promote condition than in other conditions. These results were replicated in a follow-up study with a new set of visual stimuli.

The current work reinforces the view of learning traps as emerging from a self-perpetuating cycle of exploration, attention, and learning. Early learning leads to a false belief about the features of the environment that predict rewards and losses. This false belief limits attention to a subset of relevant features and leads to avoidance of rewarding stimuli. In environments where feedback is only received when an item is approached, there is little opportunity for the false belief to be corrected and the trap persists.

Consistent with this view, and with previous findings, the simple-rule trap was rarely observed under full-feedback conditions (Lee et al., 2024; Li et al., 2021; Rich & Gureckis, 2018). Notably, even when the manipulation of early learning order favored the development of a one-dimensional rule in the LT Promote condition, participants quickly escaped from the trap when they received corrective feedback after each response (see the top-middle plot in Fig. 3). This highlights that immediate disconfirmation of an overly simple rule is of one way of escaping from learning traps.

Although our manipulation of early learning succeeded in potentiating trap formation, we had less success in trap prevention. Across the primary experiment and two follow-up studies, we only observed a reduction in learning trap in the LT Prevent condition once. This suggests that it may be harder to *expand* attention across relevant dimensions than to *restrict* attention to a subset of dimensions by manipulating early experience. One possible explanation for this pattern is that, even though the experience of a large loss was delayed in LT Prevent conditions, some participants could still have encountered a large loss at an early stage of learning (i.e., within the first eight trials), and this was sufficient to lead them into the trap. Previous work by Li et al. (2021) evidenced the impact of large losses as opposed to large gains on the formation of learning traps. They found that reversing the payoff schedule so that only small losses were experienced (i.e., approaching s00, s01, and s10 resulted in a small loss of -1 point, but approaching s11 led to a large gain of $+3$ points) dramatically reduced trap formation. Building on this result, future work could examine if the experience of small losses early in learning reduces trap prevalence.

On reflection, the differing results for our LT Promote and LT Prevent conditions may not seem so surprising. The LT Promote condition encouraged more people to learn a simple category rule that had some benefits to the learner (e.g., allowed them to avoid losses). The goal of LT Prevent, however, was more challenging—to encourage people to continue exploring the category stimuli until they had discovered a more complex rule that avoided losses and optimized rewards.

Building on the modeling efforts of Rich and Gureckis (2018), we used ALCOVE-RL as a measurement model to probe the interaction between early experience and attention distribution in trap formation. Although the model performed reasonably well in capturing the observed key choice patterns, we note that it tended to underestimate the proportion of approach choices for safe bees and overestimate that for dangerous ones (see the Supplementary Materials). This is likely because the trial-by-trial network updating in ALCOVE-RL might underestimate the speed with which learners shift their attention between dimensions (e.g., Rehder & Hoffman, 2005).

In addition, ALCOVE-RL is constrained by the assumption that experienced exemplars are perfectly encoded and stored in memory throughout the task (e.g., Griffiths & Mitchell, 2008), which might not always be true in reality. It is important for future work to investigate the extent to which memory precision may affect the formation of learning traps, possibly by joining the modeling efforts of ALCOVE-RL with other candidate models (e.g., AARM, Weichart et al., 2022). Likewise, future work could utilize eye tracking (e.g., Blanco et al., 2023; Watson et al., 2024) to provide a more direct test of ALCOVE-RL's predictions about the effects of early experience on selective attention.

The current work focused mainly on behavioral and quantitative analyses for those who learned a one-dimensional or two-dimensional category rule. However, our model-fitting results also revealed some intriguing findings for the unclassified learners whose behavior was inconsistent with either rule. We present detailed analyses for this subgroup in the Supplementary Materials. In short, estimated attention weights indicated that the attention distribution to relevant features of this subgroup was actually more similar to that of two-dimensional rule users than one-dimensional rule learners. This shows that a broad attention distribution across stimulus features does not guarantee optimal category learning. As well as attending to relevant features, it is necessary to learn the correct mapping of these features to category outcomes. Inspection of the ALCOVE-RL exemplar/outcome learning parameters (l_ω; see the Supplementary Materials) suggests that unclassified learners often struggled to learn this mapping. Consequently, the unclassified group accumulated markedly lower earnings throughout the learning phase compared to both 1D and 2D rule users.

To summarize, the current study investigated the impact of early experience and its interaction with selective attention on the formation of a simple-rule learning trap. Our findings confirm the important role played by early learning experience in shaping the development of learning traps, as well as providing nuanced insights into its interaction with selective attention in trap formation in multi-feature environments. As well as providing a deeper understanding of the cognitive mechanisms that underpin learning traps, our results have implications for learning outside the laboratory. They reinforce that, under the common environmental constraint of contingent feedback, traps are common and hard to prevent. Nevertheless, our results suggest how training environments should be structured if we want to avoid trap potentiation.

Acknowledgments

This work was supported by the Australian Research Council Discovery Grant DP220101592 to BKH and BRN.

Open access publishing facilitated by University of New South Wales, as part of the Wiley - University of New South Wales agreement via the Council of Australian University Librarians.

Data availability statement

The data collected for all of the experiments are available on the Open Science Framework at: https://osf.io/pbf8y/. All of the experiments were conducted under human research ethics #3333 approved by the institutional review board of the University of New South Wales. Some of the ideas and data appearing in this paper were presented in preliminary form at the 46th Annual Meeting of the Cognitive Science Society (Liu, Newell, Lee, & Hayes, 2024).

Notes

1 A power analysis for balanced one-way analysis of variance tests for six groups with a medium effect size (= 0.25), significance level of 0.05, and power of 0.9 requires at least 45 participants within each group.

2 This division guarantees that each bee type appears four times within a block, except for Blocks 1 and 2 in LT Promote conditions.

3 R^2_{McF} denotes McFadden's R^2 (McFadden, 1974), which is a measure of improvement in log-likelihood of models like logistic regressions as compared to the null model ($R^2_{McF} = 1 - \frac{LL_{Full}}{LL_{Null}}$). Due to its formulation, R^2_{McF} tends to yield lower values than traditional R^2, with values between .2 and .4 generally considered to indicate an excellent fit (McFadden, 2021).

4 A preliminary inspection of these normalized attention weights affirmed an overall effect of learning the relevant dimensions for predicting category membership in all conditions (as the magnitude of attention weights for both relevant dimensions increased, weights for the irrelevant dimension decreased; see Fig. S6).

References

Blanco, N. J., Turner, B. M., & Sloutsky, V. M. (2023). The benefits of immature cognitive control: How distributed attention guards against learning traps. *Journal of Experimental Child Psychology*, *226*, 105548.

Broadbent, D. E. (2013). *Perception and communication*. Elsevier.

Carvalho, P. F., & Goldstone, R. L. (2017). The sequence of study changes what information is attended to, encoded, and remembered during category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(11), 1699.

Carvalho, P. F., & Goldstone, R. L. (2022). A computational model of context-dependent encodings during category learning. *Cognitive Science*, *46*(4), e13128.

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*(1), 19–22.

de Leeuw, J. R. (2015). Jspsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*, (1), 1–12.

Denrell, J. (2005). Why most people disapprove of me: Experience sampling in impression formation. *Psychological Review*, *112*, (4), 951.

Elwin, E., Juslin, P., Olsson, H., & Enkvist, T. (2007). Constructivist coding: Learning from selective feedback. *Psychological Science*, *18*, (2), 105–110.

Feldman, J. (2016). The simplicity principle in perception and cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *7*, (5), 330–340.

Galdo, M., Weichart, E. R., Sloutsky, V. M., & Turner, B. M. (2022). The quest for simplicity in human learning: Identifying the constraints on attention. *Cognitive Psychology*, *138*, 101508.

Griffiths, O., & Mitchell, C. J. (2008). Selective attention in human associative learning and recognition memory. *Journal of Experimental Psychology: General*, *137*, (4), 626.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*, (8), 534–539.

Hills, T. T. (2006). Animal foraging and the evolution of goal-directed cognition. *Cognitive Science*, *30*, (1), 3–41.

jamovi. (2025). *The jamovi project* (Version 2.6)[Computer Software]. Retrieved from https://www.jamovi.org

Kruschke, J. K. (1992). Alcove: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, (1), 22–44.

Lee, W. J., Li, A. X., Lee, J. E., & Hayes, B. K. (2024). Learning traps and change blindness in dynamic environments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 50(9), 1345.*

Lejarraga, T., Schulte-Mecklenbeck, M., Pachur, T., & Hertwig, R. (2019). The attention–aversion gap: How allocation of attention relates to loss aversion. *Evolution and Human Behavior*, *40*, (5), 457–469.

Li, A. X., Gureckis, T. M., & Hayes, B. (2021). Can losses help attenuate learning traps? In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Liu, Y., Newell, B., Lee, J. E., & Hayes, B. (2024). Examining the relationship between selective attention and the formation of learning traps. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

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

Mathy, F., & Feldman, J. (2009). A rule-based presentation order facilitates category learning. *Psychonomic Bulletin & Review*, *16*, (6), 1050–1057.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior.

McFadden, D. (2021). Quantitative methods for analysing travel behaviour of individuals: Some recent developments. In Hensher, D. A., & Stopher, P. R. (Eds), *Behavioural travel modelling* (pp. 279–318). Routledge.

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., Hausmann, D., Fiedler, K., & Gonzalez, C. (2015). Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision*, *2*, (3), 191.

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, *35*, (21), 8145–8157.

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

Pothos, E. M., & Close, J. (2008). One or two dimensions in spontaneous classification: A simplicity approach. *Cognition*, *107*, (2), 581–602.

R Core Team. (2023). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rehder, B., & Hoffman, A. B. (2005). Thirty-something categorization results explained: Selective attention, eyetracking, and models of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, (5), 811.

Rich, A. S., & Gureckis, T. M. (2018). The limits of learning: Exploration, generalization, and the development of learning traps. *Journal of Experimental Psychology: General*, *147*, (11), 1553.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press.

Teodorescu, K., & Erev, I. (2014). On the decision to explore new alternatives: The coexistence of under- and over-exploration. *Journal of Behavioral Decision Making*, *27*, (2), 109–123.

Watson, P., Lee, W. J., Lee, J. E., Liu, Y., Newell, B. R., & Hayes, B. K. (2024). Trapped by selective attention: The role of attentional processes in the emergence and prevention of learning traps. *PsyArXiv*. Retrieved from https://osf.io/preprints/psyarxiv/7fx4p

Weichart, E. R., Galdo, M., Sloutsky, V. M., & Turner, B. M. (2022). As within, so without, as above, so below: Common mechanisms can support between- and within-trial category learning dynamics. *Psychological Review*, *129*, (5), 1104.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supplementary Material