Examining the Relationship Between Selective Attention and the Formation of Learning Traps

Yanjun Liu (yanjun031130@gmail.com)
School of Psychology, University of New South Wales
Sydney, NSW, 2052, Australia

Ben R. Newell (Ben Newell (ben.newell@unsw.edu.au)
School of Psychology, University of New South Wales
Sydney, NSW, 2052, Australia

Jaimie Lee(jaimie.lee@unsw.edu.au)
School of Psychology, University of New South Wales
Sydney, NSW, 2052, Australia

Brett K. Hayes(b.hayes@unsw.edu.au)
School of Psychology, University of New South Wales
Sydney, NSW, 2052, Australia

Abstract

Selective attention to predictive cues is often considered an efficient way to address the exploration-exploitation dilemma in decision-making. Yet in some circumstances, it can also lead to sub-optimal decision-making due to false beliefs about the environment acquired early in learning - a learning trap. In this study, we examined the relationship between attention selectivity and the emergence of a one-dimensional learning trap in a multidimensional categorization learning task. Combining empirical work (N=75) and computational modeling, we find that more selective attention, especially in the early phase of learning, increases the likelihood that an individual will fall into a learning trap. This finding sheds light on the causal role of attentional biases in the way that individuals explore and learn about choice-options.

Keywords: selective attention, decision-making, categorization, category learning, exemplar models

Introduction

From everyday decisions like grocery shopping to significant investments such as buying a house, we often rely on beliefs learned from prior experience to guide our choices. For instance, when engaging in a real estate transaction, we may reflect on our past experiences with an agent to determine if we would like to hire an agent this time. 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 utilizing existing knowledge.

Selective attention plays an essential role in addressing the exploration-exploitation dilemma (Nosofsky, 1986; Kruschke, 1992; Niv et al., 2015). It allows decision makers to focus on relevant information while filtering out irrelevant details (Desimone & Duncan, 1995; Soto, Hodsoll, Rotshtein, & Humphreys, 2008). Nevertheless, this seemingly efficient mechanism can sometimes result in sub-optimal decisions (Hoffman & Rehder, 2010; Rich & Gureckis, 2018; Blanco, Turner, & Sloutsky, 2023; Kruschke & Blair, 2000). A negative experience with a tech stock, may lead an investor to believe that all tech stocks are bad and avoid trades with those stocks, regardless of the potential for some of them to yield a good return. This is an example of a "*learning trap*" - where early negative experience leads the decision maker to form a false belief about the reward structure of the environment, causing them to avoid exploring choice options that would disconfirm the false belief. Such traps can lead people to miss out on potential rewards (Rich & Gureckis, 2018) and fail to learn causal relations (Liquin & Gopnik, 2022), which have been implicated in the development of negative stereotypes of out-groups (Denrell, 2005) and the maintenance of psychopathologies such as depression (Teodorescu & Erev, 2014).

Rich and Gureckis (2018) suggested that selective attention plays a key role in the formation of learning traps. In our investment example, the trap emerges because the investor focuses on a single salient feature (i.e., whether the stock is tech-related) to guide subsequent decisions, while ignoring other key indicators (e.g., the relative strength index). Some existing work supports this relationship between selective attention and trap formation in learning (Rich & Gureckis, 2018; Blanco et al., 2023). Blanco et al. (2023) identified a group-level distinction in attention distribution during learning between adults and children using eye-tracking data. They found that children who distributed their attention across stimulus features more broadly during learning were less likely to fall into a learning trap than adults who tended to focus on a small number of predictive features.

The current work aimed to carry out an in-depth examination of the relationship between selective attention and learning trap formation with a joint approach of experimental methods and quantitative model fitting. We examined

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

trap formation in a task where different categories (i.e., types of cartoon bees) were associated with rewards or losses. A conjunctive rule involving two feature dimensions could perfectly predict the category bound. On each learning trial, participants chose to approach or avoid a bee. Approaching the bee led to the associated outcome, while avoidance meant that no gain or loss was incurred. Using a similar task, Rich and Gureckis (2018) found that when learners only received outcome feedback if they approached a stimulus (but not when they chose to avoid), many fell into the trap of using an overly simplistic one-dimensional category rule. That is, they avoided losses but also earned fewer rewards than those who learned the optimal two-dimensional rule.

Unpacking the relationship between selective attention, representation of the environment, and exploratory behavior is challenging (Turner & Sloutsky, 2024). Causal relations between these processes are likely to run in both directions. For instance, a belief that a single feature dimension predicts category membership could lead to selective attention to that feature and less exploration of alternative features, which then enhances the original belief.

In the current work, we probed the relationship between selective attention and the learning trap with a novel approach of assessing the link between a model-based measurement for selective attention and the likelihood of falling into a trap. Rich and Gureckis (2018) showed that the ALCOVE-RL model, which extends Kruschke's (1992) well-known ALCOVE model of category learning to incorporate aspects of reinforcement learning, could generate learning trap formation. The current work goes beyond simulations. Using parameters estimated from model fitting of ALCOVE-RL to individual learning data, we derived a model-based measurement that quantified learners' trial-by-trial degree of attention selectivity. This allowed a granular-level test of the relationship between selective attention and the emergence of a one-dimensional trap over the course of learning. In particular, we examined whether individuals who showed more attentional biases toward a subset of predictive features early in learning were more at risk of falling into a trap, as opposed to those who distributed their attention more broadly.

Methods

Participants

A total of 75 participants (26 women, 48 men, 1 non-binary; age: M = 35.72 , SD = 12.38), recruited from the Prolific online platform, completed the study. The sample size were determined prior to data collection, based on a preliminary power analysis (Rich & Gureckis, 2018).

All participants who completed the study were paid with a £2 base rate and a performance-based bonus ranging from £0 to £1.70. 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.

Materials

We adapted the bee stimuli used by Rich and Gureckis (2018), generating eight unique stimuli by factorially combining three binary-valued bee dimensions (Figure 1a), including numbers of legs (two vs. six), numbers of wing pairs (single vs. double), and body pattern (striped vs. dotted).

For each participant, two of these binary-valued dimensions were randomly selected as relevant for categorization. Two feature values on relevant dimensions were randomly assigned as potentially dangerous (denoted by 1), and the other two were assigned as friendly features (denoted by 0). The category membership of each stimulus was determined by the conjunction of features on relevant dimensions (Figure 1b). The stimuli having friendly features on both relevant dimensions (s00), and the stimuli having a single friendly feature (s01 and s10) were members of the friendly category. The stimuli incorporating both potentially dangerous features (s11) belonged to the dangerous category. For example, when the relevant dimensions were legs and body pattern, a bee with the two potentially dangerous features of two legs and dotted body (s11) was dangerous. The bee types having six legs and striped body (s00), six legs and dotted body (s01), and having two legs and striped body (s10) were friendly. Feature values on the third, irrelevant dimension (wings in this example) were not predictive of category membership.

The experiment was programmed in jsPsych (de Leeuw, 2015) and run online via the Prolific platform. Model fitting was conducted in R (R Core Team, 2023) and statistical analyses were conducted in Jamovi (jamovi, 2021).

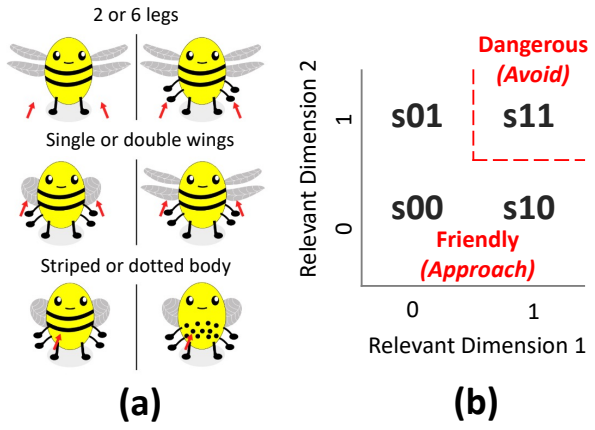

Figure 1: Exemplar bee features and categories used in the experiment. Panel (a): Exemplar binary-valued bee features. Two dimensions were relevant for predicting category membership and one was irrelevant. Panel (b): Schematic category structure (friendly vs. dangerous) of different bee types (i.e., s00, s01, s10, s11) based on relevant dimensions. The bee type composed of two potentially dangerous features (i.e., s11) is dangerous and should be avoided. All other bee types are friendly and should be approached.

Procedure

Participants were given the role of beekeeper in a virtual game, with the goal of maximizing the honey harvested from bees, quantified as accumulated points. They were informed that some bees were friendly - approaching these bees would yield honey valued at +1 points; while some bees were dangerous - approaching these bees would lead to a loss of 3 points. During the instruction session, participants saw the features of all three binary-valued dimensions (see Figure 1a). They were not informed of the category rule, but were told that perfect prediction on bee categories was possible. Before learning commenced, participants had to achieve a perfect score on a comprehension survey that queried: (1) the three dimensions that would vary across stimuli, (2) the task goals, and (3) the consequences of avoiding a bee.

There were 112 trials in the experimental session, divided into six learning blocks and one test block. Transition between blocks was not signaled. In each block, 12 friendly bees and 4 dangerous bees were presented in random order. Each unique bee stimulus appeared twice in a block. On each trial, a bee stimulus appeared on screen and participants had to make an approach/avoid decision. Responses were made by clicking on on-screen buttons. In learning blocks, after an action was chosen, the stimulus was replaced by a feedback screen for 2s, giving the associated outcomes. The test block was similar except that no feedback was given on any trials. Points were still accumulated throughout the test block.

Participants commenced the task with an endowment of 50 points. Accumulated points appeared on the top-right of the screen on each trial. On completion, participants received payment based on their final points tally.

Results

Learning of Category Rules

We assessed the categorization rules that learners acquired by examining their approach and avoidance patterns in the training and test blocks, using a scoring method similar to Rich and Gureckis (2018). Those who learned the correct two-dimensional (2D) rule should approach all three types of friendly bee types (s00, s01 and s10 in Figure 1b) and avoid the dangerous bee type (s11). Those using a sub-optimal one-dimensional (1D) rule would only use values on a single relevant dimension to guide decisions. Given the stimulus structure, there are two versions of the 1D rule, either approaching s00 and s01 but avoiding s11 and s10, or approaching s00 and s10 but avoiding s11 and s01. Those using either version were classified together as "1D rule learners". A participant was deemed to use a category rule within a block if their approach/avoid decisions were consistent with that rule on at least 15 out of 16 trials. Participants whose choices were inconsistent with both rules were labeled as unclassified learners. We analysed and modeled the responses of all participants. However, for brevity, we focus on results of 1D- and 2D-rule learners.

As shown in Figure 2, the proportion of participants employing either a 1D or 2D rule increased over the course of learning. The proportion of 2D-rule learners increased from 0.013 in the first block to 0.480 in the last block ($\chi^2(6) = 74.9,\ p < 0.001$), and the proportion of 1D-rule learners increased from 0.013 to 0.213 ($\chi^2(6) = 28.0,\ p < 0.001$). Hence, similar to what was found in previous studies (Rich & Gureckis, 2018; Blanco et al., 2023), a substantial minority of learners fell into the trap of using an overly simplistic one-dimensional rule to guide approach and avoid decisions. To assess the progress of category learning across blocks,

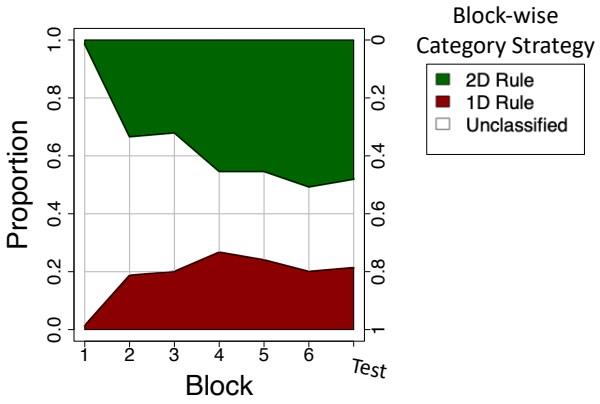

Figure 2: Proportions of participants employing different dimensional category strategies in each block. Note that the proportions of 1D learners combine different versions of this rule - focusing on either of the two relevant dimensions.

we classified participants into one of two sub-groups based on the category rule they employed most frequently across blocks. Approach responses to different bee types for each sub-group are shown in Figure 3. Stimuli that were unambiguously "friendly" (s00) were approached and stimuli that were unambiguously "dangerous" (s11) were avoided by both 2D and 1D rule users. The groups diverged in their approach of stimuli (s01, s10) that were actually friendly but would have been believed dangerous if one focused only on features from a single relevant dimension. Notably, this divergence in approach behavior was evident at an early stage of learning (i.e., in the first two blocks). This suggests that an early selective attention bias may have contributed to the formation of a persistent learning trap. The results of a mixed effects linear regression model (Table 1) supported the significance of the observed patterns in Figure 3.

Modeling the Relationship Between Selective Attention, Category Learning and Learning Traps

The goal of our modeling was to (1) examine the capability of ALCOVE-RL model in capturing choice patterns of different category strategies, and (2) more importantly, to probe how selective attention to a subset of relevant features early in learning contributes to the one-dimensional learning trap.

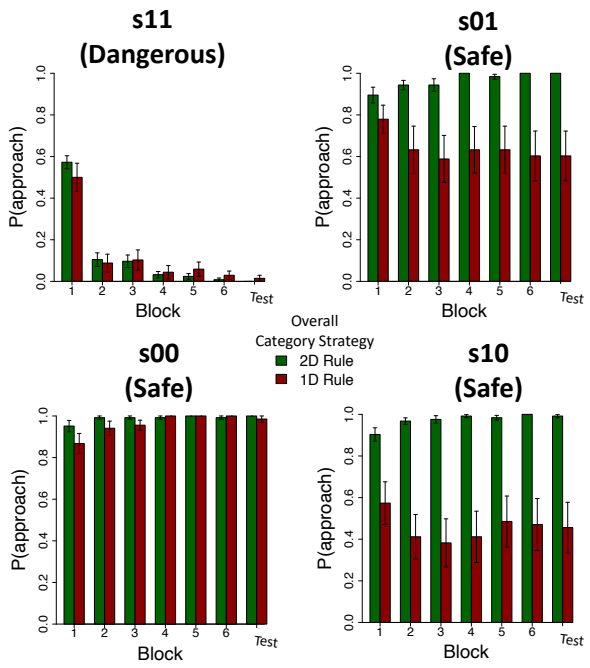

Figure 3: Choice proportions for approaching bees of different types, by subgroups using different category rules. Error bars denote the standard error of the mean.

Table 1: The results of omnibus tests for the fixed effects in the mixed regression model predicting P(approach) as a function of Stimulus Type, Dimensional Rule and Block, after accounting for by-subject random intercepts.

Fixed Effect	F	$df_{between}$	df_{within}	p-value
Stimulus Type	1498.13	3	2632	< 0.001
Dimensional Rule	519.77	1	2632	< 0.001
Block	12.96	6	2632	< 0.001
Stimulus Type $\times$ *DimensionalRule*	169.33	3	2632	< 0.001
Stimulus Type $\times$ *Block*	14.87	18	2632	< 0.001
Dimensional Rule $\times$ *Block*	1.53	6	2632	0.165
Stimulus Type $\times$ *DimensionalRule* $\times$ *Block*	2.06	18	2632	< 0.001

A Brief Description of ALCOVE-RL Model ALCOVE-RL belongs to the class of connectionist models of exemplar-based category learning (Nosofsky, Palmeri, & McKinley, 1994; Eiser, Fazio, Stafford, & Prescott, 2003). Following the structure of the original ALCOVE model (Kruschke, 1992), ALCOVE-RL describes the underlying categorization mechanism as a network connecting a hidden layer of exemplar nodes to a hidden layer of action nodes. The exemplar layer captures the mapping of dimensional inputs (a_i^{in}) from stimuli to an exemplar-based representation (h_{ji}) through a similarity-based function (Equation 1).

$$a_j^{hid} = exp[-c(\sum_i \alpha_i |h_{ji} - a_i^{in}|)], \quad j = \text{exemplar index},\ i = \text{index of input dimension}. \tag{1}$$

It implements a selective attention mechanism through the incorporation of attention weights for feature dimensions (α_i in Equation 1). Same values of attention weights on dimensions indicate that each dimensional input is contributing equally to the activation of exemplars for categorization decisions; while asymmetrical attention weights suggest an attention bias to ward certain dimensions when encoding stimulus. These attention weights are our key measure in assessing the role of selective attention in learning trap formation. c is a specificity parameter that reflects overall psychological discriminability when activating exemplars.

The layer of action nodes captures the cognitive mapping of exemplar nodes to action outputs (Equation 2), which is regulated by learnable association strength ω.

$$a_k^{out} = \sum_j \omega_{kj} a_j^{hid} / \sum_j a_j^{hid},\ k = \text{action index} \tag{2}$$

The choice probability of taking an action is computed using a classic probabilistic choice model (Equation 3, Luce, 1959), regulated by a parameter ϕ that reflects the overall degree of determination of taking an action

$$P(K) = \frac{exp(\phi a_k^{out})}{\sum_k exp(\phi a_k^{out})} \tag{3}$$

ALCOVE-RL involves a reinforcement learning algorithm to improve its prediction on reward outputs with outcome sampling. Specifically, the association weights (i.e., α and ω) of the model are updated based on the difference between the predicted reward of the model and the actual reward of a chosen action on each trial (Equation 4). Prediction improvement does not take place for the action that is not taken by the learner (see Rich & Gureckis, 2018, for a detailed model description). The updating speed is regulated by the learning rates (l_α and l_ω).

$$\begin{aligned} \Delta\omega_{kj} &= l_\omega (t_k - a_k^{out}) a_j^{hid} \\ \Delta\alpha_i &= -l_\alpha \sum_j [\sum_k (t_k - a_k^{out}) \omega_{kj}] a_j^{hid} c |h_{ji} - a_i^{in}| \end{aligned} \tag{4}$$

Categorization Choices with Different Degrees of Attention Bias We fitted ALCOVE-RL to each participant's data and obtained the maximum-likelihood estimations of the model parameters, including a specificity constant c (M = 5.43, 95% CI = [3.96, 6.91]) in the activation function for exemplars (Equation 1), a deterministic parameter ϕ (M = 2.72, 95% CI = [2.38, 3.06]) in the probabilistic choice function (Equation 3), and two learning rates l_α (M = 0.179, 95% CI = [0.169, 0.188]) and l_ω (M = 0.162, 95% CI = [0.148, 0.176]) in the error-driven learning function (Equation 4).

The attention weights for each feature dimension were then derived for each trial based on Equation 4. Initial weights were set to be identical for all three dimensions at the beginning of learning. As depicted in Figure 4a, with the advance of experiment, the estimated attention weights on both relevant dimensions (a1 and a2) increased while those on the irrelevant dimension (a3) decreased, suggesting that participants directed their attention to relevant dimensions as learning progressed.

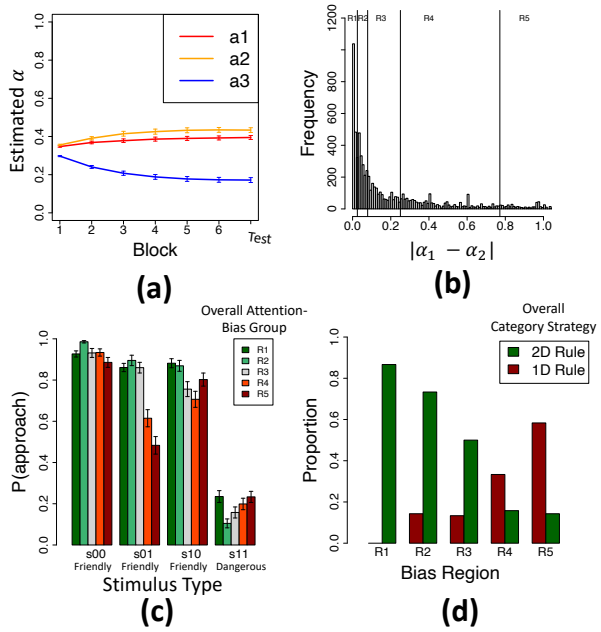

Figure 4: Panel (a): Estimated attention weights on each binary-valued dimension across blocks. a1 and a2 are weights for the relevant dimensions, a3 is the weight between the irrelevant dimension. Panel (b): The distribution of absolute differences in attention weights for relevant dimensions, divided into five regions with evenly-spaced quantiles. Panel(c): Approach proportions for each bee type, estimated from participants with different levels of attention bias. Error bars denote the standard error of the mean. Panel (d): Proportions of category-rule users in each quantile region of bias.

We quantified the degree of attention bias by the absolute difference between normalized attention weights on the two relevant dimensions. For instance, if the estimated a1, a2, and a3 were 0.8, 1.6, 0.1, respectively, the absolute difference in attention weights between relevant dimensions was $\frac{|0.8-1.6|}{0.8+1.6+0.1} = 0.32$. Figure 4b summarizes the full distribution of absolute attention-weight differences estimated from trials of all participants. Four quantiles (20%, 40%, 60%, 80%) divided the values of weight differences into five regions, reflecting different levels of dimensional attention bias. Weight-difference values in R1 reflected the null or marginal attention bias, while those in R5 reflected the relatively strongest bias toward a single relevant dimension.

Figure 4c shows the choice proportions of approach decisions for participants in different attention-bias groups. The null/marginal attention-bias group (R1) demonstrated a choice pattern consistent with use of a 2D rule (green bars in Figure 2b). They tended to always approach the friendly bees (s00, s01 and s10) and avoid the dangerous bees. In contrast, learners with high levels of bias (R4-R5) demonstrated a choice pattern consistent with use of a 1D-rule; that is, they were less likely to approach the ambiguous friendly bees (s01 and s10). The patterns observed in Figure 4c were supported by the results of mixed effects regression model ($R^2_{Conditional} = 0.593$), confirming the significance of main effects of both Stimulus Type ($F(3,2010) = 814.65,\ p < 0.001$), and Overall Attention Bias ($F(4,70) = 4.20,\ p = 0.004$), as well as their interaction ($F(12,2010) = 15.31,\ p < 0.001$) on the choice proportion of approaching bees.

In a similar vein, Figure 4d shows how levels of attention bias quantified by model-based measurement were related to use of different category rules inferred from approach/avoid choice patterns. There was a significant association between these measures, $\chi^2(4) = 20.2,\ p < 0.001$. Participants with the highest levels of attention bias were most likely to be classified as falling into the 1D learning trap,

So far, these results reveal that observed differences in use of one- or two-dimensional rules for categorization decisions are reflected in the attention bias estimates derived from the ALCOVE-RL model. In the next section, we turn to the more interesting question of whether model-based estimates of attention bias early in learning can predict the later emergence of a one-dimensional learning trap.

Early Attention Bias and the Emergence of Learning Traps Figure 5a plots the values of absolute attention-weight differences as a function of blocks for different category-rule learners, reflecting how attention selectivity changed with learning for different learners. The absolute differences of the 2D-rule learners started near zero and remained at a consistent level across blocks, suggesting an even distribution of attention on stimulus dimensions during category learning. In contrast, the absolute differences of the 1D-rule learners started with a positive value and increased with the advance of blocks. This suggests that 1D-rule learners initiated learning with an uneven distribution of attention, and the attention biases toward a single dimension became stronger as learning progressed. The observed patterns were supported by the results of the mixed effects regression model ($R^2_{conditional} = 0.870$), which confirmed the significance of main effects of Block ($F(6,787) = 7.44,\ p < 0.001$) and Dimensional Rule ($F(1,761) = 88.61,\ p < 0.001$), as well as their interactions ($F(6,786) = 45,\ p < 0.001$), on the absolute attention-weight differences. These results hinted at a distinction in early attention bias between groups using different category strategies.

Figure 5b takes a closer look at this issue. Participants were divided into R1 to R5 groups based on their absolute attention-weight differences estimated from the first two learning blocks. Choice proportions were estimated from the last block (the test block), reflecting the category strategies participants employed after the completion of learning phase.

The choice patterns in Figure 5b show a clear effect of early attention-bias degree on subsequent categorization decisions. Participants with a null/marginal initial attention-bias degree (i.e., R1) ended up with a choice pattern consistent

with a 2D-rule. As the initial attention bias increased, there was a reduction in the likelihood of a participant to approach the ambiguous friendly bees (s01 or s10) in the final blocks. Hence, an early bias towards selective attention to a single relevant feature meant that a participant was more likely to eventually fall into a learning trap. The statistical results of the mixed regression model ($R^2_{Conditional} = 0.624$) supported the observed patterns, confirming the significance of main effects of Stimulus Type ($F(4, 2530) = 709.17,\ p < 0.001$) and Early Attention Bias ($F(4, 70) = 5.49,\ p < 0.001$), as well as their interactions ($F(16, 2530) = 21.13,\ p < 0.001$), on choice proportions in the last block.

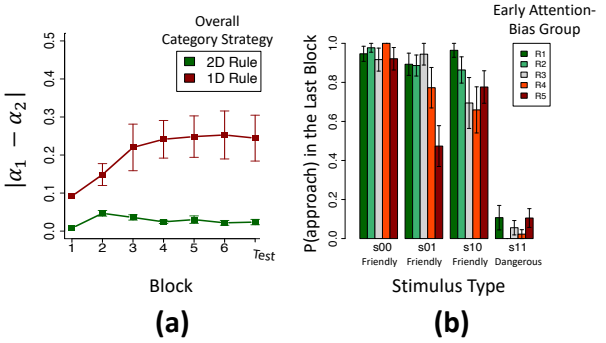

Figure 5: Panel (a): Estimated absolute attention-weight differences across blocks of different category-rule learners. Panel (b): Proportion of approaching bees of different types in the last test block, estimated from participants with different levels of early attention-bias. Bias was quantified by the estimated absolute attention-weight differences in the first three blocks. Error bars denote the standard error of the mean.

Discussion

This study examined the relationship between attention selectivity and the emergence of a learning trap in a multi-dimensional category learning task. After an extended period of learning, we found that a substantial proportion of learners fell into the trap of using an overly simplistic one-dimensional rule to classify stimuli, avoiding stimuli that would have earned them rewards. We utilized a model-based measurement of attention bias, derived from fitted parameters of ALCOVE-RL (Rich & Gureckis, 2018), to probe how individual levels of selective attention related to trap development. A crucial finding was that selective attention in the initial stages of learning was linked to later trap formation. Participants who started with a broad distribution of attention were more willing to explore the full outcome space in the early phase of learning and hence more likely to eventually learn the correct two-dimensional rule. In contrast, participants who showed higher levels of selective attention to a single feature dimension early in learning were less willing to explore the full space of outcomes and more likely to fall into a learning trap.

The current findings are consistent with suggestions that individuals may differ in their tendency to explore novel stimuli or exploit known options (e.g., Gershman & Tzovaras, 2018; Hills & Hertwig, 2010). Importantly, our work makes a novel contribution by highlighting the crucial role of early selective attention in subsequent exploration of novel options, as evidenced by both empirical and modeling results.

Other types of individual differences may also contribute to various attentional bias identified in the current experimental context. It has been shown in range of decision-making environments that the disutility of a loss tends to be perceived greater than the utility of a gain (Kahneman & Tversky, 1979; Tversky & Kahneman, 1991). In our study, we suspect that loss aversion might affect the underlying interplay between attention selectivity and under-exploration in the early phase of learning. Losses from approaching dangerous bees may have a larger hedonic impact than gains from approaching friendly bees for some participants. This may have driven these learners toward a conservative learning strategy, specifically focusing attention on features that could help them predict losses rather than earn rewards. Once this loss-avoidance strategy was in place, no further attention to other stimulus features was required. Li et al. (2021) found that using a contrasting payoff schedule, that is rewarding approaching the conjunctive bee type (s11) while penalizing approaching the disjunctive bee types (s00, s01, and s10), could attenuate the learning trap, even when participants were informed about the frequency of positive versus negative stimuli. These results indicate that the formation of learning trap may relate to the perceived incentives for exploration in addition to loss attention (Lejarraga, Schulte-Mecklenbeck, Pachur, & Hertwig, 2019; Yechiam & Hochman, 2013).

Our results support the Rich and Gureckis(2018) hypothesis that selective attention plays a key role in the development of learning traps. However, as noted, the causal relations between representation of the learning environment, selective attention and exploratory behaviour are complex, and bi-directional in many cases (Turner & Sloutsky, 2024). We therefore need to be cautious about making strong claims about early attentional biases being the primary factor driving later formation of learning traps. Our understanding of these complex issues could be advanced in future work through the use of direct measures of selective attention, such as eye tracking (e.g., Blanco et al., 2023), and direct probes of people's beliefs about the structure of their learning environment at multiple points during learning. These data could be used to further constrain computational modeling, allowing us to track how both attention and learning parameters change over the course of learning and how each is related to trap formation.

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 Experi-*

mental Child Psychology, *226*, 105548.

de Leeuw, J. R. (2015). jspsych: a JavaScript library for creating behavioral experiments in a web browser. *Behav. Res. Methods*, *47*(1), 1–12.

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

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*(1), 193–222.

Eiser, J. R., Fazio, R. H., Stafford, T., & Prescott, T. J. (2003). Connectionist simulation of attitude learning: Asymmetries in the acquisition of positive and negative evaluations. *Personality and Social Psychology Bulletin*, *29*(10), 1221–1235.

Gershman, S. J., & Tzovaras, B. G. (2018). Dopaminergic genes are associated with both directed and random exploration. *Neuropsychologia*, *120*, 97–104.

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.

Hills, T. T., & Hertwig, R. (2010). Information search in decisions from experience: Do our patterns of sampling foreshadow our decisions? *Psychological science*, *21*(12), 1787–1792.

Hoffman, A. B., & Rehder, B. (2010). The costs of supervised classification: The effect of learning task on conceptual flexibility. *Journal of Experimental Psychology: General*, *139*(2), 319.

jamovi. (2021). *The jamovi project.* Retrieved from `https://www.jamovi.org`

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–292.

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

Kruschke, J. K., & Blair, N. J. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & Review*, *7*(4), 636–645.

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* (Vol. 43).

Liquin, E. G., & Gopnik, A. (2022). Children are more exploratory and learn more than adults in an approach-avoid task. *Cognition*, *218*, 104940.

Luce, R. D. (1959). Individual choice behavior.

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., ... 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.

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

R Core Team. (2023). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

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.

Soto, D., Hodsoll, J., Rotshtein, P., & Humphreys, G. W. (2008). Automatic guidance of attention from working memory. *Trends in Cognitive Sciences*, *12*(9), 342–348.

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.

Turner, B. M., & Sloutsky, V. M. (2024). Cognitive inertia: Cyclical interactions between attention and memory shape learning. *Current Directions in Psychological Science*.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The quarterly journal of economics*, *106*(4), 1039–1061.

Yechiam, E., & Hochman, G. (2013). Loss-aversion or loss-attention: The impact of losses on cognitive performance. *Cognitive Psychology*, *66*(2), 212–231.