Trapped by selective attention: The role of attentional processes in the emergence and prevention of learning traps

Watson, P.[1, 2], Lee W. J.,[1] Lee J. E.[1], Liu, Y.,[1] Newell, B. R.[1] & Hayes, B. K.[1]

1 The University of New South Wales, Sydney, Australia

2 The University of Technology, Sydney, Australia

Word Count: 10782

Author Note

The project was supported by an Australian Research Council Discovery Grant DP220101592 to Brett Hayes and Ben Newell. The authors have no known conflicts of interest to disclose. All data, analysis details and stimuli are available at https://osf.io/8fdp9/. We would like to thank Paris Mousamas, Shoha Alam, Anusha Brungi, Jiale Huang, Emily Phan, Mahika Parmar and Teodora Siljanovska for their assistance with data collection.

Poppy Watson: Graduate School of Health, University of Technology Sydney, 100 Broadway, Ultimo, NSW, 2007, AUSTRALIA. Yanjun Liu, Won Jae Lee, Jaimie E. Lee, Ben R. Newell, Brett K. Hayes: School of Psychology, University of New South Wales, Anzac Parade, Kensington, NSW, 2052 AUSTRALIA Sydney.

Corresponding Author: Brett K. Hayes, b.hayes@unsw.edu.au

https://orcid.org/0000-0003-1415-0088

Abstract

Learning traps are cycles of suboptimal decision-making where a false belief about the structure of the environment leads to avoidance of rewarding options. Two experiments (N= 324) examined the role of selective attention in the emergence and prevention of learning traps. Participants learned to approach or avoid members of two categories that were associated with either gains or losses. A rule involving two visual feature dimensions predicted category membership, while a third dimension was irrelevant. Category feedback was provided only when an item was approached. Selective attention during learning was assessed using an eye tracker. Experiment 1 found that many participants fell into the trap of using a single-dimension rule to guide approach/avoid decisions and consequently missed rewards. Eye tracking confirmed that these participants narrowed their visual attention to a single dimension. In Experiment 2, we manipulated visual feature salience across three groups and found evidence for a causal role of selective attention in determining the emergence and prevalence of learning traps.

Keywords: Category learning, Decision-making, Selective Attention, Eye tracking

Public Significance: People often fall into negative cycles of decision-making where they neglect rewarding options. These studies show that encouraging people to spend more time looking at the features of choice options can break this cycle, but only when these features carry useful information.

Learning via experience typically involves a trade-off between the need to *explore*, learning which cues in the environment predict positive or negative outcomes, and *exploiting* this knowledge – taking actions that lead to these outcomes. For example, when you move to a new city you will probably explore several of the local cafés and compare the quality of their brews. Over time, you are likely to select the best of these options for your daily coffee – and avoid those where the brew was subpar. Likewise, after arriving at a social networking event, you might chat to several people. As the evening progresses, you are likely to spend more time with people who created a positive first impression and avoid those where your interaction was less pleasant.

An important finding from research on explore-exploit dilemmas is that false beliefs about the reward structure of the environment that emerge early in learning can lead to *premature termination of exploration* – meaning that the learner may fail to discover additional rewards (Denrell & March, 2001; Rich & Gureckis, 2018, Teodorescu & Erev, 2014a). For example, a poor coffee made by a novice barista may lead you to believe (incorrectly) that all coffees from that vendor are of low quality. Crucially, in environments where feedback is available when one decides to approach a choice option but not when it is avoided, the false belief will not be corrected. This can lead to a persistent cycle of sub-optimal exploration and decision making that has been termed a *learning trap* (Blanco et al., 2023; Lee et al., 2024; Rich & Gureckis, 2018).

Understanding how traps emerge is important because they can have serious negative consequences. Learning traps have been linked to reduced economic rewards (Rich & Gureckis, 2018, Teodorescu & Erev, 2014a), poor management decisions (Elwin, 2013) and distorted first impressions and negative stereotypes of out-groups (Bai et al., 2022; Denrell, 2005). They have also been implicated in psychopathologies like depression (Teodorescu & Erev, 2014b). Once established, learning traps are very difficult to change. Rich and Gureckis (2018), for example, trialed three experimental interventions aimed at preventing or reversing learning traps. None were successful.

The current work therefore had two overarching aims. The first was to better understand the role of selective attention in the formation and maintenance of learning traps. The second was to examine how this knowledge could be used to help prevent traps.

Learning traps in Environments with Predictive Features

Early research on learning traps focused on environments where the only signal of the reward value of a particular option was past experience of rewards or losses (e.g., Denrell & March, 2001; Harris et al., 2020). In such environments, a trap can emerge when the outcomes that result from approaching a particular choice option vary over time; for example, on average outcomes are positive but occasionally a large negative outcome is encountered. Experience of the negative outcome early in learning can lead to a *hot stove* effect, where the learner subsequently avoids that option and never discovers the additional available rewards (Denrell & March, 2001; Denrell & Le Mens, 2020).

In many learning environments, however, we interact with stimuli that have multiple observable features, some of which can be used to predict what will happen if the stimulus is approached (e.g., Blanco et al., 2023; Rich & Gureckis, 2018; Schulz et al., 2018). In such environments *a false belief about which features predict rewards and losses*, can lead to avoidance of potentially rewarding options. Given that learners often prefer simple over complex predictive relationships (Chater & Vitanyi, 2003; Galdo et al., 2022), Rich and Gureckis (2018) suggested that people can fall into the trap of focusing on a subset of the actual features that predict choice outcomes. For example, an early negative experience in a particular café may lead to people to believe that all cafés that belong to the same franchise serve poor quality brews. When subsequently deciding which cafés to approach or avoid, the learner may focus on franchise logos.

Rich and Gureckis (2018) showed how this *simple-rules trap* can emerge in a task that combined elements of category learning and decision-making. Participants learned to approach or avoid members of categories (e.g., friendly vs. dangerous bees), that were respectively associated with gains or losses. The categories could be differentiated by a conjunctive rule involving two feature

dimensions (e.g., approaching bees with spotted bodies and many legs would lead to a loss; approaching stimuli with other feature conjunctions would lead to a gain). When outcome feedback was *choice-contingent* (i.e., only provided when a stimulus was approached), many participants failed to learn the category rule that maximized rewards. Instead, they used a simple one-dimensional rule to guide decisions (e.g., avoid all bees with spotted bodies). This meant that they avoided losses but earned substantially fewer rewards than those who learned the correct rule.

Subsequent work has shown that such traps emerge relatively early in the learning process (Li et al., 2021; Liu et al., 2024) and are more common in adult than child learners (Liquin & Gopnik, 2022). Moreover, such traps can lead the learner to be blind to important changes in the reward structure of the environment (Blanco et al., 2023; Lee et al., 2024).

The Role of Selective Attention in Learning Traps

Simple rules traps involve a focus on a sub-set of predictive features when making decisions about whether to approach or avoid a stimulus. The details of this process, however, have yet to be examined. One possibility is that a simple-rules trap reflects an underweighting of relevant features, but this underweighting occurs only at the decision stage. In this account, multiple features are attended to and encoded by the learner, but only a subset are used when making an approach or avoid decision. In our café example, the learner may attend to many features such as the age and gender of the barista, the type of coffee being ordered and the café logo, but only use a subset of these to make a decision about where to buy their next brew.

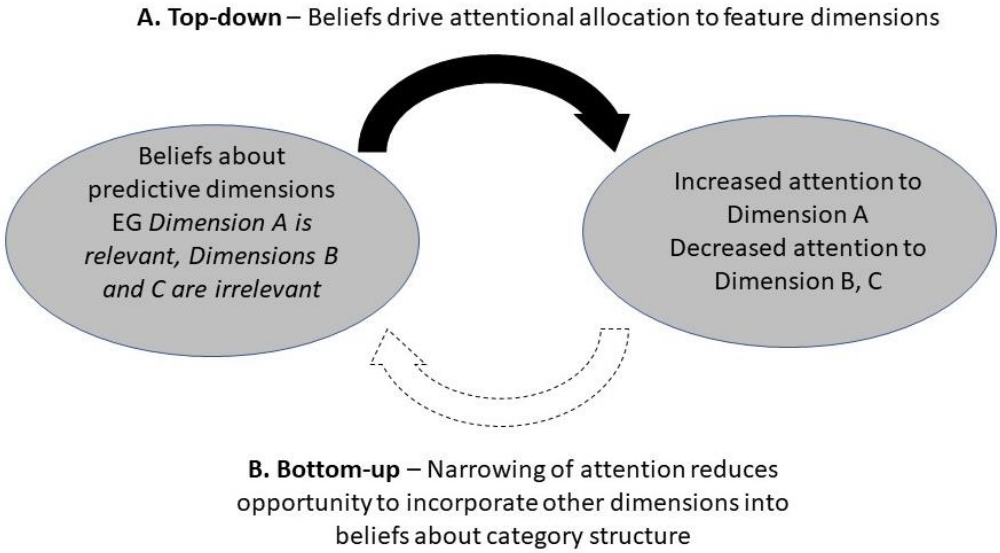
An alternative possibility is that a false belief about which features are relevant for prediction affects both decision-making *and selective attention*. Rich and Gureckis (2018) suggested that a belief that a single feature dimension was relevant for predicting category membership could lead to selective attention to that dimension, with other relevant dimensions ignored. Such selective attention could intensify a nascent trap. Attentional narrowing means that it becomes less likely that the learner will discover other features that may be relevant to learning the true reward structure. For example, if you

focus exclusively on franchise names to decide where to go for coffee, you are likely to miss other, potentially more informative cues.

A schematic version of this account is shown in Figure 1. It envisages a bi-directional relationship between people's beliefs about feature-outcome relations and selective attention to stimulus features (cf. Orquin et al., 2021; Turner et al., 2021). In the initial stages of learning, beliefs about which features are relevant for predicting outcomes guides visual attention – the *top-down* pathway shown in the Figure. As learning progresses, selective attention to a subset of relevant features reduces the likelihood of exploring and learning about stimuli that vary on other relevant features – the *bottom-up* pathway.

Figure 1

Possible pathways in the role of selective attention in the development of learning traps



There is abundant evidence that people selectively attend to predictive features when learning new categories (e.g., Blair et al., 2009a, b; Blanco & Sloutsky, 2019; Galdo et al., 2022; Rehder & Hoffman, 2005a, b). Rehder and Hoffman (2005a), for example, examined visual selective attention as participants learned category structures devised by Shepard et al. (1961) based on one, two, or three deterministic feature dimensions. Early in learning, participants typically fixated on all stimulus

dimensions. As learning progressed, learners shifted their gaze towards the most predictive dimensions. Likewise, over the course of learning, participants ceased to gaze at dimensions that were irrelevant for prediction. These results were robust when analyzed at both group and individual levels.

Subsequent work has shown that these results generalize to more complex cases involving the learning of more than two category alternatives (e.g., Blair et al., 2009a). Moreover, learners modulate their selective attention based on the strength of the predictive relationship between features and categories, attending more to deterministic features than to features that only have a probabilistic relationship with category membership (e.g., Blanco et al., 2023; Rehder & Hoffman, 2005b).

This work shows that people learn to selectively attend to features that actually predict category membership. The open question that we address is whether behavior indicative of the *false* belief that a single feature dimension predicts category outcomes, would be accompanied by a shift in selective attention patterns – namely selective visual attention to a single relevant feature. In two studies, learners were tasked with discriminating between visual categories associated with either gains or losses respectively. As in Rich and Gureckis (2018), on each trial, learners could choose to approach or avoid a stimulus. When feedback about category membership and gain/loss outcomes were contingent on approach, it was expected that many participants would fall into a simple-rules learning trap. Crucially, eye tracking was used to assess visual attention to the various category features as learning progressed. This allowed us to examine the relationship between selective attention to these features and decisions to approach or avoid.

Our first key question, addressed in Experiment 1, was whether those who learned different category rules, as exhibited by their patterns of stimulus approach and avoidance, also differed in their patterns of visual selective attention. In particular, we were interested in whether those who fell into a simple-rules trap showed a narrowing of their visual attention to a single dimension early in learning, such that they subsequently failed to attend to other relevant dimensions.

Our second question was whether *manipulation* of selective attention during learning could

reduce the prevalence of learning traps and promote learning of the optimal decision rule. If it is the case that overly selective attention to a subset of features contributes to trap formation, then taking steps to reduce selectivity (e.g., by increasing the salience of unattended features), may help to prevent traps. This issue was examined in Experiment 2.

Experiment 1

This study tested the hypothesis, suggested by Rich and Gureckis (2018), that the development of a simple-rules learning trap is accompanied by a narrowing of attention to a subset of the features that are relevant for predicting choice outcomes. We used a category learning task where stimuli were composed of three binary feature dimensions. Two were relevant for predicting whether stimuli were friendly (would earn points if approached) or dangerous (would lose points). A third dimension was not predictive of category outcomes. As in previous work on the simple-rules trap (e.g., Lee et al., 2024; Li et al., 2022; Liu et al., 2024; Rich & Gureckis, 2018), participants were only provided with category feedback and corresponding gains or losses when a stimulus was approached. We expected that this contingent feedback would lead a substantial proportion of participants to fall into the trap of relying on a single feature dimension to guide approach/avoid decisions. This would lead to avoidance of losses but a failure to maximize rewards. Further it was expected that that trap would emerge early in learning and persist despite further learning opportunities.[1]

Throughout the task, visual attention to the three stimulus features was monitored using an eye tracker. By examining the proportion of time that learners fixated on each dimension, we were able to track changes in selective attention. Previous studies of selective attention during category learning (e.g., Blair et al., 2009a; Rehder & Hoffman, 2005a) led us to expect that most participants would

[1] A preliminary study carried out with 200 online participants without eye tracking used the same stimuli and general procedure as the current experiments but compared patterns of approach and avoidance when outcome feedback was contingent on approaching a stimulus or when feedback was provided on both avoid and approach trials ("full feedback"). This study confirmed that participants rarely fell into a learning trap when full feedback was provided (only 3% of participants did so), but many (39%) fell into the trap when feedback was decision-contingent. See https://osf.io/8fdp9/ for study details and data.

show a reduction in attention to the irrelevant dimension over the course of learning. The key novel question was whether those showing different patterns of approach and avoid behavior (i.e., patterns consistent with a belief that category membership was predicted by one or two relevant dimensions), would show different patterns of selective attention. In particular, do those who fall into the trap of using a one-dimensional rule reduce their attention to the other relevant dimension? Relatedly, we examined whether an individual's pattern of selective attention to relevant features early in leaning could predict their subsequent decision-making patterns.

Method

Participants

We recruited 109 undergraduates from the University of New South Wales, who received course credit for participation. They could also earn a bonus payment, with every extra point accrued in the study converted to 1 cent, so a total of $1.34 AUD could be earned. Thirty-six participants were excluded because their gaze data did not meet the acceptable threshold for eye tracking analysis (see Results). The demographics for the remaining 73 participants was as follows: M_{age} = 19.95 years, *SEM* = 0.167, Range = 17 to 28 years; 51 females, 22 males. The rationale for the sample size was to replicate or exceed that of Rich and Gureckis (2018, Experiment 2, n≈75 per cell). Ethics approval was obtained from the UNSW Human Research Ethics Panel.

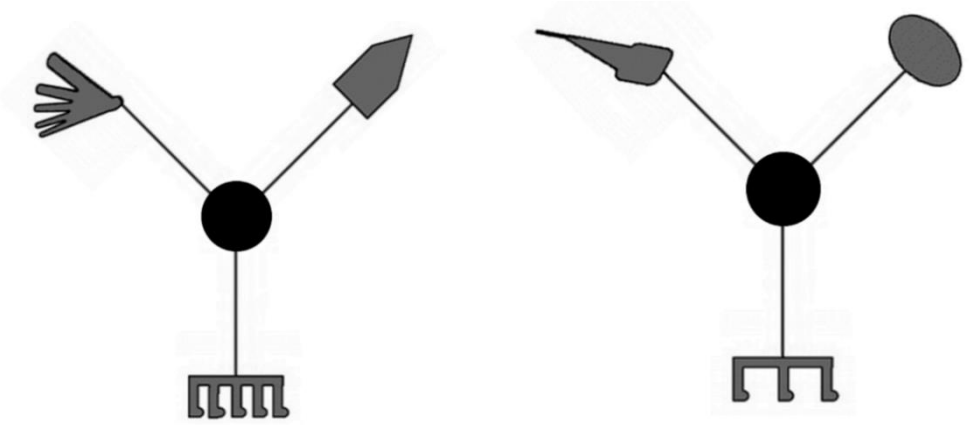
Materials

Our alien stimuli had a similar abstract feature structure to those used in previous learning traps studies (e.g., Lee et at., 2024; Rich & Gureckis, 2018), but incorporated visual features conducive to eye tracking, similar to those used by Rehder and Hoffman (2005a). As shown in Figure 2, exemplars varied along three binary-valued feature dimensions (arm with oval or square head; arm with a claw with three or five fingers; arm with a feathered or pointed tail). Hence, there were eight unique alien stimuli. All were presented on a white background. Each learning trial began with a black fixation cross presented at the center of the screen, 0.5 x 0.5° visual angle. A circular area of interest (AOI)

was defined around the fixation cross with radius of 1.5° visual angle. Once 500ms of gaze had been registered at the fixation location (or after 4000ms), an alien was presented. A rectangular AOI was defined at each of three dimensions measuring 6.35 x 6.35° visual angle.

Figure 2.

Stimulus examples showing features on the three dimensions



Procedure

The experimental paradigm was programmed in MATLAB (MathWorks, 2022) using Psychophysics Toolbox extensions (Brainerd, 1997; Kleiner et al., 2007). Participants were tested individually, with head position stabilized using a chin rest 60 cm from the screen. Gaze was recorded using a Tobii Pro Spectrum eye-tracker (sample rate 600 Hz) mounted on a 23-inch monitor (1920 × 1280 resolution, 120 Hz refresh rate). The eye tracker was calibrated at the start of the task.

Participants were instructed that they were "virtual crystal collectors" tasked with collecting crystals from aliens. Aliens were either "friendly", providing crystals if approached (leading to a 1-point gain), or "dangerous", attacking if approached (leading to a 3-point loss). If an alien was avoided, no category outcome feedback was provided and the points balance was unaffected. Participants started with a balance of 50 points and were paid a monetary bonus depending on their points tally at the end of the experiment. Before commencing learning, participants had to complete an instruction comprehension check consisting of four multiple-choice questions. If a participant made an error on any question, they were informed of the correct answer by the experimenter.

The learning phase consisted of six blocks of 16 trials.[2] Within each block, participants encountered two presentations of the eight unique exemplars, including 2 x 6 friendly and 2 x 2 dangerous items. Transition between blocks was not signaled to participants. Two of the three feature dimensions were relevant to categorizing aliens. A conjunctive two-dimensional rule perfectly predicted category membership (e.g., aliens with three fingers and a feathered tail were dangerous, aliens with other feature combinations on these dimensions were friendly). Features on the remaining dimension were not predictive of category membership. Across trials, features on these dimensions were equally likely to be associated with friendly or dangerous aliens. The relevant dimensions and feature combinations that determined category membership were randomly assigned for each participant by the experimental program.

The trial structure is summarized in the top panel of Figure 3. On each trial, the participant was presented with a single alien stimulus and chose to approach or avoid it by pressing keys (Z or M, counterbalanced across participants). The trial ended and the stimulus was removed after participants made a response. Feedback was presented on a separate screen for 2000ms. If participants approached a 'friendly' alien the feedback read *'You successfully collected the crystal! You earned 1 point.'* If they chose to approach a 'dangerous' alien they saw *'Ouch, you were attacked! You lost 3 points.'* If an alien was avoided, no feedback was provided about the alien identity and the screen simply read *'You avoided this alien. You earned 0 points.'* An on-screen counter tallied the current points earned, and was updated after each trial. The inter-trial-interval was 1000 ms. For the majority of trials, item

[2] Following Rich and Gureckis (2018) we also administered 16 "no feedback" test trials after the final learning block. Each exemplar was shown twice in randomized order with no feedback provided after approach or avoid decisions. However, in Experiment 2 these stimuli lacked the colored ring used to highlight stimulus features (see Experiment 2 Methods). This meant that it was problematic to compare behavioral and eye-tracking test data between conditions where rings were present or absent during learning. Hence, our measures of final category rule use were based on approach/avoid decisions on the last learning block (block 6). Data for test block performance can be found in https://osf.io/8fdp9/. Including test block results in analyses did not alter any of our key findings or conclusions.

presentation order was randomized within blocks.[3]

Participants' beliefs about the structure of the observed categories were assessed after learning was complete. Text descriptions corresponding to the three stimulus feature dimensions ("head", "fingers", "tail") were presented in random order and participants checked boxes to indicate dimensions that they believed were relevant to alien classification. Participants were also asked to estimate the proportion of friendly aliens that they experienced on 0-100 scale.

Data Processing

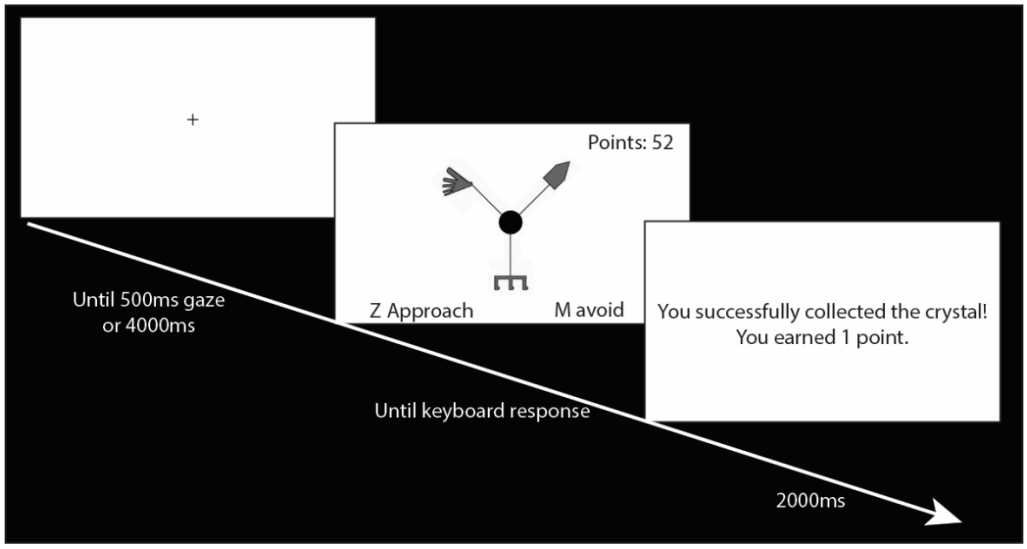
Behavioral data. The key outcome measure was the type of categorization rule that participants used to guide approach/avoid decisions during learning blocks. Within blocks of 16 stimuli, perfect conformity to the optimal two-dimensional rule would lead a learner to approach 12 friendly aliens and avoid the 4 dangerous aliens (resulting in a net gain of 12 points). Perfect conformity to a one-dimensional rule would lead a learner to approach 8 friendly aliens, avoid 4 dangerous *and* 4 friendly aliens (net gain of 8 points). In a given block, participants were classified as "2D rule users" or "1D rule users" if their choice patterns were consistent with one of these rules on at least 15 trials. The stimulus structure meant that there were always two possible one-dimensional rules for a given stimulus set. Participants using either one-dimensional rule were classified together. Those whose choices did not satisfy either criterion were said to be using an "unclassified rule". Rule use in the final learning block (2D, 1D or unclassified) was the key outcome variable in behavioral analyses.

[3] In this study, we were also interested in the effects of different sequences of item presentation on learning trap formation. Hence, for trials 1-24, two sub-groups saw the same stimuli presented in different semi-randomized sequences. One sub-group saw multiple instances of dangerous aliens, before they saw all of the friendly aliens. For the other subgroup, this sequence was reversed. This manipulation had no measurable effect on the way people learned category rules, so in all analyses we collapsed over this factor.

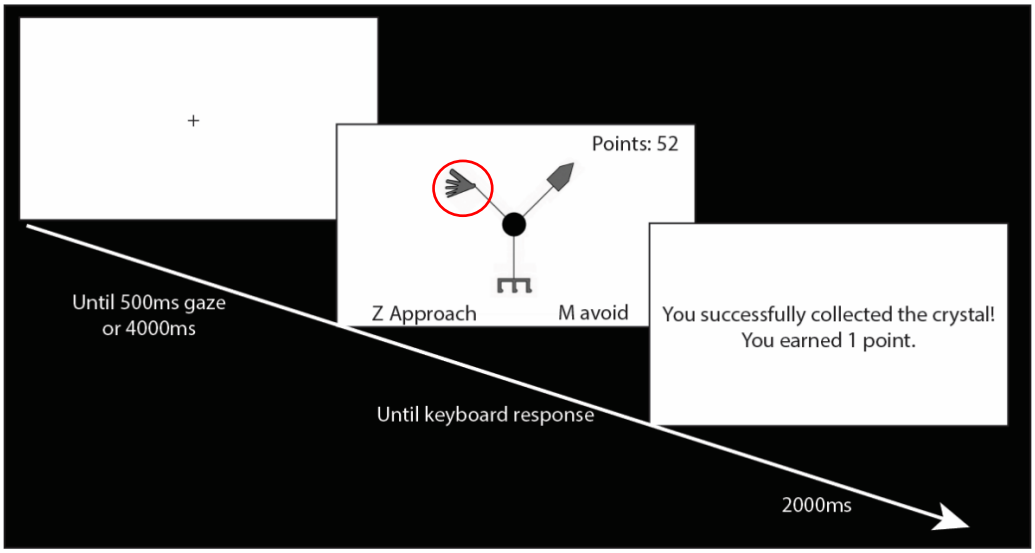
Figure 3

Example trials in the learning traps task with eye tracking.

A. Without visual highlighting (Experiments 1, 2)



B. With visual highlighting (Experiment 2)



Note. During learning, participants received feedback about the correct category and points outcome when an approach response was made. No category feedback was provided if an avoid response was made (feedback screen said that no points were earned). In the Experiment 2 random rings group, participants saw a highlighted ring around one of the features on each trial (randomized order). In the Experiment 2 informative rings group, on two-thirds of the trials, one of the two relevant features was shown with a highlighted ring (randomized order).

Eye tracking data. We were primarily interested in the relative amount of time that participants spent looking at the three alien features, on each trial. We identified fixations (and saccades) using a velocity threshold identification algorithm (cf. Salvucci & Goldberg, 2000). Following linear interpolation across gaps in the raw gaze data of <75 ms, the data were smoothed with a five-point moving average filter. Fixations were defined as periods when the gaze location was stable with a velocity criterion of less than 40° visual angle per second. The mean *x* and *y* coordinates across the entire fixation were then calculated. The three alien features were defined as circular AOIs with diameter 5.86°. If the mean fixation coordinates fell within one of the specified AOIs, that fixation was coded as being at a particular feature. We excluded all fixation data unless the fixation was greater than 100 ms at one of the three feature locations. Fixations to other screen areas (e.g., the points tally or the text for the approach/avoid question) were not included in our analyses.

Trials with no fixation data on any of the three features (23% of all trials) were excluded, as were trials where the eye tracker captured insufficient samples (i.e., below the threshold of 0.25 valid samples; 0.22% of remaining trials). Thirty-six participants were excluded from the analyses for having more than 30% trial exclusions (13 1D learners, 17 2D learners and 6 unclassified). The mean proportion of valid eye tracking samples per trial was .93 (*SEM* = .007).

For each participant, we calculated the mean number of fixations (of at least 100 ms) on each of the three features and the duration of those fixations. Mean total dwell time to the features varied substantially between participants (e.g., mean dwell time to all AOIs ranged from 268 ms to 1923 ms, *M* = 741.3 ms, *SD* = 352.24). Because of this large individual variation, and because we were primarily interested in the extent to which participants' visual attention was distributed across dimensions, we focused on the *proportion* of time that participants spent fixating on each of the three features.

As a summary measure of a participant's distribution of attention across features, we calculated mean *attentional entropy* per block H_t (see Blanco et al., 2023 for a similar measure). The calculation is given in Equation 1, where t_d is the mean proportion of time spent gazing at each dimension *d*,

averaged across trials in that block. Entropy scores were normalized so that 0 represents selectively looking at one feature on each trial of an entire block and 1 represents equal distribution of attention across the three features. The normative value of H_t in this context was 0.693, which would reflect an equal distribution to the two relevant features but no attention to the irrelevant feature.

$$H_t = -\sum_{d}^{D} \ln(t_d)\ t_d \qquad (1)$$

Transparency and Openness

We report the rationale for determining our sample size, all data exclusions, all manipulations and measures. All data, analysis details and stimuli are available at https://osf.io/8fdp9/. Data were analyzed using JASP (JASP Team, 2019) and SPSS (IBM, 2024). Neither study was pre-registered.

Results

Behavioral Measures: Category Rule Use and Category Beliefs

Figure 4 shows the proportion of participants using differ category rules in each learning block. There was clear evidence of learning with the proportion using a 1D or 2D rule increasing over blocks, Cochran's Q (5, N=73) =120.939, p<.001. By the final learning block, a majority learned the optimal 2D rule (49.3%), but a substantial minority had fallen into a learning trap, using a suboptimal 1D rule (28.8%). This trap typically emerged early in learning. Most of those using a 1D rule on the final block (75%) first showed evidence of 1D responding within the first three learning blocks. A one-way analysis of variance (ANOVA) confirmed that those using a 2D rule on the final block earned more bonus points (M = 52.444, SEM = 1.760) than those using a 1D rule (M = 34.0, SEM = 2.304) or unclassified rule (M = 25.0, SEM = 2.64), F(2, 72) = 44.122, p<.001, η^2_p = 0.558.

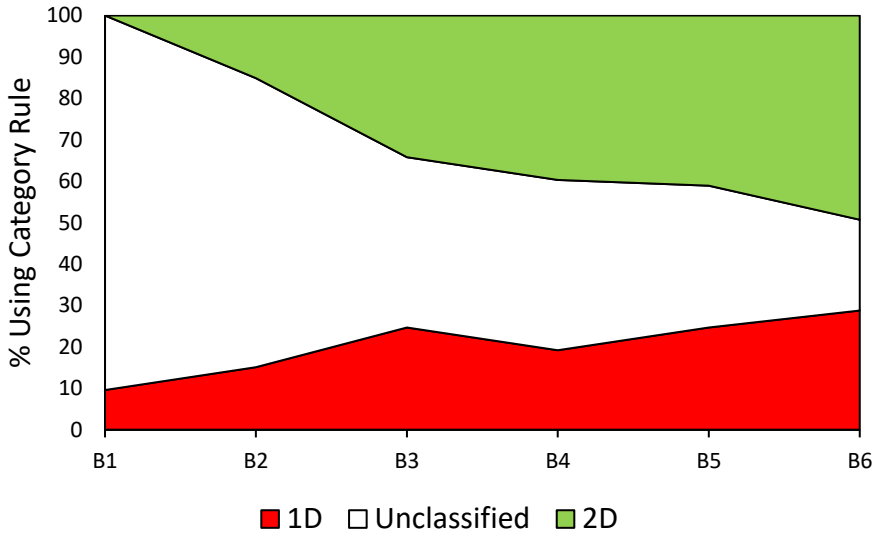
Post-test queries probed participants' beliefs about the structure of the learning environment.[4] We first compared the number of dimensions identified as relevant for categorization by participants

[4] A programming error meant that post-test responses were not available for some participants.

using different category rules at the end of learning. The number of identified features differed across each rule category, $\chi^2(4, N=71) = 23.24$, $p < .001$. Beliefs about relevant features were generally consistent with these rules. Most of those using a 1D rule identified just one dimension as relevant (89.5%), with a minority identifying two features (10.5%). This pattern was reversed for those using a 2D rule, who most often identified two dimensions as relevant (63.9%), with a minority identifying one relevant feature (36.1%). Those using an unclassified rule most often identified just one dimension as relevant (62.5%), with minorities identifying two (25%) or all three dimensions (12.5%). Post-test estimates of the percentage of friendly aliens also reflected the rules used at the end of learning. Those using a 2D rule gave estimates (M = 76.69) close to the correct figure (75%). Lower estimates were given by those using a 1D rule (M = 50.71) or an unclassified rule (M = 65.69), $F(2, 58) = 29.12$, $p<.001$, $\eta^2_p = 0.510$.

Figure 4

Experiment 1. Category rule use in each learning block (N=73)



Note: The Figure shows the proportion of participants whose pattern of approach/avoidance decisions in a given block was consistent with the optimal two-dimensional categorization rule, a suboptimal one-dimensional rule or an unclassified rule. See the online article for the color version.

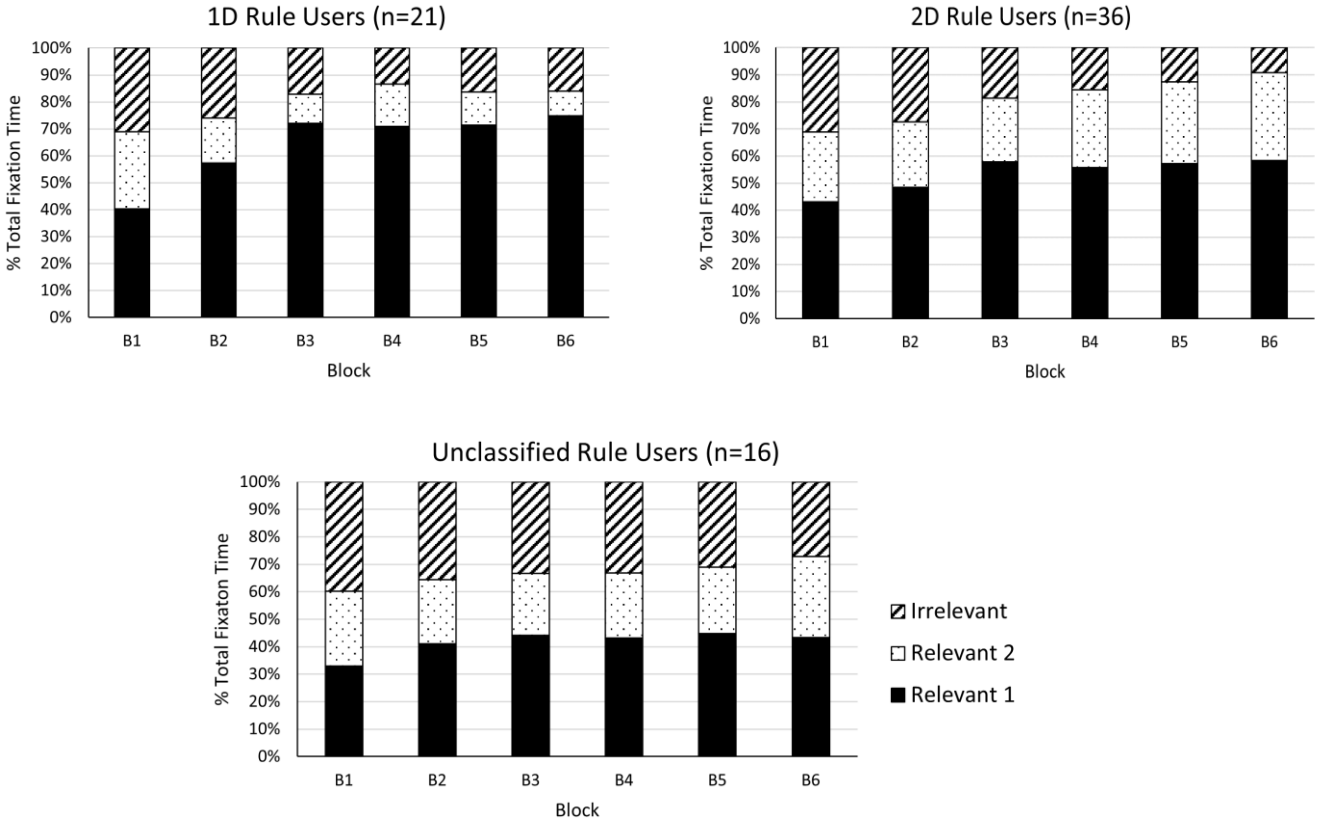
Eye tracking

Our primary interest was in the relative amount of time that participants spent examining the areas of interest associated with features on each stimulus dimension. For each participant, we calculated the proportion of time spent fixating to each of the three features on each trial. This data was aggregated to give the mean proportion of fixation time per feature for each learning block. To allow for interpretable comparisons between groups, for a given participant, the feature that was relevant to the category rule that had the highest mean fixation within a block was designated as "relevant feature 1" for that block. The other feature that was relevant to the rule was designated as "relevant feature 2". Figure 5 shows patterns of visual attention to these features, as well as to the irrelevant feature, for sub-groups using different category rules at the end of learning. Those who eventually learned a 1D rule showed selective attention to a single relevant feature, with increasing relative fixation to that feature (i.e., relevant feature 1) over the course of learning and decreasing fixation time to the other features. Those who eventually learned a 2D rule showed increasing levels of fixation to both relevant features, but decreasing attention to the irrelevant feature. Those who used an unclassified rule showed similar levels of attention to all three features, and this pattern remained relatively stable across learning.

Attentional entropy scores represent a summary measure of the distribution of attention across feature dimensions (see Figure 6). The entropy scores for sub-groups using different category rules at the end of learning data were entered into a 3 (final rule-use) x 6 (block) ANOVA with repeated measures on the second factor. The sub-groups using different rules at the end of the learning showed different entropy patterns, $F(2, 70) = 18.870$, $p<.001$, $\eta^2_p = 0.350$. Those using a 1D rule showed lower attentional entropy ($M = 0.535$, $SEM = 0.038$) than those who learned the optimal 2D rule ($M = 0.773$, $SEM = 0.035$) or an unclassified rule ($M = 0.882$, $SEM = 0.040$). Entropy declined across blocks, $F(5, 70) = 29.112$, $p<.001$, $\eta^2_p = 0.294$, but this effect interacted with rule use, $F(10, 350) = 7.985$, $p<.001$, $\eta^2_p = 0.186$. Figure 6 shows that 1D rule users showed the steepest decline in entropy.

Figure 5

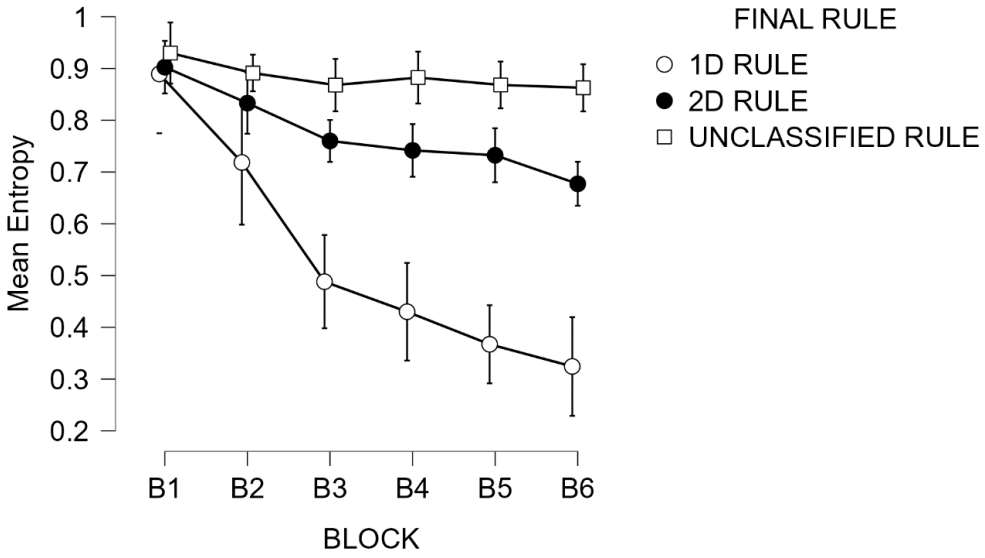
Experiment 1. Percentage of total fixation time attending to each feature



Note. For each participant, the feature that was relevant to the optimal two-dimensional category rule that had the highest mean fixation within a block was designated as “relevant feature 1” for a given participant. The other feature that was relevant to the rule was designated as “relevant feature 2”.

Figure 6

Mean entropy scores in each learning block for each rule-use group



Note. An entropy score of 0.0 means that, across trials, all fixation time was devoted to a single feature. An entropy score of 1.0 means that fixation time was evenly distributed between the three features. The normative entropy score, reflecting equal distribution of attention to the two relevant features but ignoring the irrelevant feature, was 0.693. Error bars represent 95% confidence intervals.

Predicting category rule use from early visual attention

An important question was whether individuals' pattern of selective attention early in learning could be used to predict the type of categorization rule that they ultimately used at the end of learning. To examine this, we entered individual entropy scores from early learning blocks (blocks 1, 2 or 3) as covariates in a series of multinomial logistic regressions, with final rule use group (i.e., the rule used in block 6) as a categorical dependent measure. Adding block 3 entropy scores to the regression equation led to a better fit to the rule use data than an intercept-only model, $\chi^2(2, N=73) = 21.357$, $p < .001$, but no reliable evidence of improvement in fit was found for entropy scores on earlier blocks, p's >0.07. The details of the relationship between block 3 entropy and final rule use are revealed by examining the regression equation parameters associated with the entropy predictor. Lower levels of

entropy at block 3 significantly increased the odds that an individual would use a suboptimal 1D rule as compared to the 2D rule, β = 1.276, *SE* = 0.417, Odds Ratio (OR) = 3.583, z = 9.349, p =.002, and increased the odds that an individual would use a 1D rule as compared to an unclassified rule, β = 2.259, *SE* = 0.612, OR = 9.569, z = 13.601, p <.001. Block 3 entropy, however, did not reliably predict whether someone was more likely to use a 2D as compared with an unclassified rule, β = 0.982, *SE* = 0.525, OR = 2.671, z = 3.497, p=.061. In short, overly selective attention to a single feature early in learning meant that individuals were more likely to find themselves in a learning trap. However, the same measure of selective attention was not able to discriminate between those who eventually learned a 2D rule and those who used an unclassified rule.

Discussion

This experiment examined the relationship between visual selective attention to exemplar features and the formation of a simple-rules learning trap. Our behavioral data showed that a majority learned the optimal two-dimensional rule that predicted category membership and associated rewards. However, a substantial minority fell into a one-dimensional learning trap, using an overly simplistic category rule to predict which exemplars they should approach or avoid. This led them to earn fewer rewards than those who learned the two-dimensional rule. A minority of participants failed to learn either rule. These behavioral patterns are similar to those found in previous studies of simple-rules learning traps (e.g., Lee et al., 2024; Li et al., 2021; Rich & Gureckis, 2018).

The experiment revealed a number of interesting relationships between patterns of selective attention, as measured by relative fixation time to the various features, and the learning of category rules. Those who acquired a 1D or 2D rule showed learned inattention to the irrelevant feature, spending minimal time fixating on this feature by the last learning block. This is consistent with previous work showing that, when learning to discriminate between category members composed of multiple features, people generally shift their attention away from non-predictive features (Blair et al., 2009a; Blanco et al., 2023; Rehder & Hoffman, 2005a, b).

Those learning 1D or 2D rules, however, showed markedly different patterns of attention to the two relevant features. Those who eventually learned a 1D rule narrowed their attention to a single relevant feature over the course of learning. Those who learned a 2D rule continued to divide their attention between both relevant features. Notably, patterns of visual attention at a relatively early stage of learning predicted whether an individual was more likely to eventually learn the correct 2D rule or fall into a 1D learning trap.

These results show that learning traps do not simply involve an underweighting of relevant features when making decisions about whether to approach a stimulus. Rather, as suggested by Rich and Gureckis (2018), such traps involve systematic changes in selective attention – those falling into a simple-rules trap attended to just one of the two feature dimensions that were relevant for predicting category membership and associated rewards. Figure 6 suggests that such narrowing of attention began relatively early (by block 3) and increased over subsequent learning blocks.

The minority of participants who did not use an identifiable rule showed less selective attention than those learning a 1D rule over the course of learning. At the end of learning, this sub-group still showed considerable attention to the feature dimension that was irrelevant to predicting category outcomes. It is also worth noting that patterns of selective attention early in learning did not discriminate between those who eventually learned the unclassified rule users and those who learned the 2D rule. This shows that attending to both relevant features early in learning was a necessary but not sufficient condition for learning the correct categorization rule. Although the unclassified group did attend to these features, they failed to learn the contingencies between features and the outcomes of approach and avoid decisions.

Experiment 2

Experiment 1 confirmed a close connection between approach/avoid decisions and patterns of selective attention as learning traps emerge. These results are consistent with the bi-directional relationship between beliefs about relevant category features and selective attention shown in Figure

1. Experiment 2 examines the novel question of whether *increasing attention to multiple features* can have a positive effect on learning – reducing the prevalence of learning traps and increasing learning of the optimal two-dimensional decision rule. In other words, our goal was to examine whether intervening in the bottom-up pathway shown in Figure 1 would benefit learning.

Our general strategy was to introduce a manipulation to enhance the salience of each feature, increasing the likelihood that learners would distribute their visual attention more broadly across feature dimensions. Feature salience can be affected by a range of perceptual factors, statistical learning and previous reward history (see Awh et al., 2012; Chun et al., 2011 for reviews). In the current work, we used a method for increasing salience similar to that used in several previous experimental (e.g., Müller et al., 2009) and applied studies (e.g., Milosavljevic et al., 2012), namely visual highlighting of features using a colored ring.

As in many learning environments outside the laboratory, our category stimuli were composed of a mixture of feature dimensions that were relevant for predicting category membership and gain/loss outcomes, and a feature dimension that was irrelevant for prediction. This raises a further interesting question about the impact of the salience manipulation. It is possible that increasing attention to all feature dimensions, regardless of their predictive value, could benefit learning. According to this view, enhancing the salience of all feature dimensions will prevent the narrowing of attention to a single dimension characteristic of learning traps. This in turn will lead to more exploration of stimuli with variations on other feature dimensions, and increase the likelihood of discovering the optimal two-dimensional rule.

This argument bears some similarity to suggestions that changes in visual attention can have a causal impact on preferential choice (e.g., Fisher, 2021; Krajbich & Rangel, 2011). According to this view, increased visual attention to a particular choice option can increase the likelihood that it will be chosen or preferred, even when its objective reward value is the same as other choice options (Bhatnagar & Orquin, 2022; Shimojo et al., 2003; Smith &Krajbich, 2019, but see Mormann & Russo,

2021 for an alternative view). Such general effects of visual attention on decision-making have been reported in studies with two-alternative forced choices (e.g., Sui et al., 2020) and tasks involving choice between options with multiple features (Krajbich & Rangel, 2011; Yang & Krajbich, 2023).

Alternately, although boosting the salience of all features may reduce selective attention to a single feature, it could *interfere* with learning of the optimal rule. The selective attention data in Experiment 1 showed that those learning an optimal two-dimensional rule distributed their visual attention across the two relevant features but showed little attention to the irrelevant feature. Highlighting all features could interfere with this filtering process. By this account, learning outcomes would only be improved if the enhancement of feature salience was restricted to prediction-relevant features.

To test these alternatives, we manipulated feature salience in two different ways. In the *random rings* group, a single ring appeared on each trial, presented with equal frequency around a feature of each of the three stimulus dimensions (two relevant, one irrelevant). The intention was for the highlighting ring to draw attention to a range of stimulus features, without adding information about which was relevant for prediction. In the *informative rings* group, the highlighting ring also carried information that would assist with learning about category structure. In this case, on a given trial, the ring was only presented around features from one of two dimensions relevant to predicting outcomes. Features from the irrelevant dimension were never highlighted. Visual attention and learning of category rules in these two groups were compared to a no rings baseline, which followed a procedure similar to the previous study.

In both ring conditions, we expected that feature highlighting would lead to a broader distribution of visual attention compared to the baseline. The crucial question was whether such changes in selective attention were accompanied by a reduction in trap prevalence and increased learning of the correct two-dimensional rule in both rings conditions or only when highlighting also provided information about feature relevance.

Method

Participants

We recruited 295 undergraduates from the University of New South Wales, who received course credit and could earn a bonus payment as detailed blow. Participants were randomly allocated to a no rings baseline, random rings group or informative rings group. Using the same criteria for acceptable gaze thresholds across trials as the previous study, 44 participants (16 no rings, 13 random rings, 15 informative rings) were excluded. The final sample details were as follows: N = 251 (n= 77 no rings; n= 83 random rings, 91 informative rings); M_{age} = 19.61 years, *SEM* = 0.258, Range = 17 to 53 years; 190 females, 61 males).[5]

Procedure

The learning phase and eye tracking procedure for the no rings condition was the same as that used in Experiment 1, except that a) the order of presentation of category exemplars within every block of 16 learning trials was fully randomized (cf. Footnote 3), and b) there was a change in the monetary bonus, as detailed below.

In the random rings condition, after passing the comprehension check, participants were told that "on some trials a red ring will appear around one of the arms of the alien. These rings may or may not help you learn about the aliens". On each learning trial in this condition, a red circle appeared around one of the three alien feature dimensions, overlaid on the alien image (see Figure 3, panel B). The ring remained visible throughout the trial. The specific feature highlighted by a ring on a given trial was selected randomly, with the constraint that each of the binary features on the three dimensions was highlighted four times over every sequence of 24 trials. Hence, over the course of learning, each feature on each dimension was highlighted on 16 occasions.

The procedure for the informative rings condition was the same as for random-rings, except that

[5] Data for the majority of the no rings and random rings conditions was collected five months before the data for informative rings. There were no differences in the demographic composition of participants tested in these two periods. A comparison between no rings participants tested in the first or the second period found no differences in attention or learning patterns.

the red highlighting ring only ever appeared around features on the relevant dimensions. One of the features on one of these dimensions was randomly selected for highlighting on 16 trials out every each 24-trial sequence (the same frequency of highlighting of these features as in the random condition). On the remaining 8 trials in the sequence no ring appeared. In all other respects, the procedure for the random and informative rings conditions was identical to the no rings condition.

To increase compliance with the procedure and reduce participant attrition due to insufficient gaze time, larger bonus payments were available in this study, with every extra point accrued in the study converted to 5 cents, so up to $6.70 AUD could be earned. Notably, fewer participants were excluded due to insufficient gaze (15%) than in Experiment 1.

Results and Discussion

Behavioral Measures: Category Rule Use and Category Beliefs

Figure 7 shows the proportion in each group using various category rules in each learning block. Each group showed clear evidence of rule learning with the proportion using a 1D or 2D rule increasing over blocks (no rings: Cochran's Q (5, N=77) = 158.443, p<.001; random rings: Cochran's Q (5, N=83) = 106.772, p<.001; informative rings: Cochran's Q (5, N=91) = 128.875, p<.001).

The pattern of rule use in the no rings condition was generally similar to that observed in Experiment 1, except that 1D rule use was slightly higher (36.4%) and 2D rule use was slightly lower (41.6%), with a similar level of use of an unclassified rule: 22%. In the random rings condition, fewer participants used either a 1D rule (20.5%) or a 2D rule (34.9%) by the final learning block, compared to the no rings baseline. A majority (44.6%) used an unclassified rule. In the informative rings condition, relatively few participants fell into the 1D trap (13.2%) with close to half using the optimal rule (46.2%), and 40.6% using an unclassified rule.

We compared rule use at the end of learning between rings groups using a multinomial logistic regression. This regression used group membership to predict the log odds of being assigned to the 1D, 2D or unclassified rule subgroups. Adding group membership as a predictor led to a significantly

better fit to the rule use data than an intercept-only model, $\chi^2(2, N=251) = 18.079$, $p < .001$. To understand the details of group effects we examined regression parameter estimates (see https://osf.io/8fdp9/ for full details). The odds of using a 2D rule as compared to a 1D rule did not differ between the random rings and no rings baseline, $\beta = 0.401$, $SE = 0.400$, OR = 1.493, $z = 1.001$, $p = .317$. Notably, however, learners were three times more likely to use a 2D compared to a 1D rule in the informative rings than the no rings group, $\beta = 1.119$, $SE = 0.417$, OR = 3.062, $z = 7.195$, $p = .007$. The odds of using an unclassified rule compared to a 1D rule, increased in both the random rings, $\beta = 1.277$, $SE = 0.425$, OR = 3.585, $z = 9.036$, $p = .003$, and informative rings groups, $\beta = 1.625$, $SE = 0.453$, OR = 5.078, $z = 12.888$, $p < .001$, relative to the no rings group. Those in the random rings group were also more likely to use an unclassified rule than a 2D rule, compared to the baseline, $\beta = 0.876$, $SE = 0.389$, OR = 2.402, $z = 5.064$, $p = .024$.

In sum, although the addition of the random rings led to some reduction in the use a one-dimension rule, it also impeded learning of the optimal rule; most participants in the random rings group learned no systematic rule. In contrast, adding informative rings reduced the prevalence of one-dimensional rule-use relative to the baseline without reducing learning of the 2D rule.

As in the previous study, those learning the optimal rule earned more bonus points ($M = 50.662$, $SEM = 1.291$) than those using a 1D ($M = 32.279$, $SEM = 1.820$) or unclassified rule ($M = 24.863$, $SEM = 1.450$), $F(2, 242) = 94.017$, $p < .001$, $\eta^2_p = 0.402$. Those in the random rings group earned fewer points ($M = 33.120$, $SEM = 1.722$) than those in the no rings baseline ($M = 40.662$, $SEM = 1.788$), $F(1, 158) = 9.233$, $p = .003$, $\eta^2_p = 0.055$. There was no difference in points earnings between the informative rings group ($M = 38.440$, $SEM = 1.877$) and the baseline, $F(1, 166) = 0.643$, $p = .424$.

Post-test queries about participants' beliefs about category structure again reflected decision rules on the final learning block, $\chi^2(4, N=251) = 68.642$, $p < .001$. Those using a 1D rule usually identified just one dimension as relevant (87.7%), with a minority correctly identifying two dimensions (12.3%), whereas those using a 2D rule showed the reverse pattern (1 relevant dimension: 29.1%; 2

relevant dimensions: 68.9%; 3 dimensions: 1.9%). Those using an unclassified rule most often identified just one dimension as relevant (58.2%), with minorities identifying two (30.8%) or three (11.0%). Those using a 2D rule gave more accurate estimates of the proportion of friendly aliens (M = 70.56, SEM = 2.956) than those using other rules (1D rule: M = 54.346, SEM = 4.079; unclassified rule: M = 63.113, SEM = 3.289), $F(2, 231) = 8.971$, $p<.001$, $\eta^2_p = 0.073$.

Notably, the pattern of explicit rules differed between rings conditions, $\chi^2(4, N=251) = 13.954$, $p=.01$. Correct identification of the two relevant dimensions was more common in the informative rings group (1 relevant: 44.0%; 2 relevant: 54.9%; all dimensions: 1.1%) than the random rings (1 relevant: 62.7%; 2 relevant: 30.1%; all dimensions: 7.2%) or no rings baseline (1 relevant: 53.2%; 2 relevant: 40.3%; all dimensions: 6.5%).

Eye tracking

Mean total fixation to stimulus AOIs per trial was longer in the rings (M = 1126 ms; SEM = 60.613) than the no rings condition (M = 987 ms; SEM = 61.665), $F(1, 159) = 5.118$, $p=.025$, $\eta^2_p = 0.031$, but as in the previous study, there was substantial individual variation. Hence, we again focus on relative proportion of fixation to each feature. Figure 8 shows these proportions for each rings condition and sub-group using different category rules. We computed entropy scores for each participant on each block, shown in Figure 9. These data were entered into a 3 (rings group) x 3 (rule group) x 6 (block) mixed model ANOVA. Overall, entropy decreased over blocks, $F(5, 1210) = 87.542$, $p<.001$, $\eta^2_p = 0.266$. However, as shown in the Figure, entropy decreased less over the course of learning in the random and informative rings groups than in the no rings group, $F(10, 1210) = 2.946$, $p=.01$, $\eta^2_p = 0.024$. Those using a 1D rule showed a steeper decrease in entropy across blocks than those using other rules, $F(10, 1210) = 11.256$, $p<.001$, $\eta^2_p = 0.085$. Across blocks, entropy was significantly lower in the no rings group (M = 0.766, SEM = 0.016) compared to the random (M = 0.975, SEM = 0.017) and informative rings conditions (M = 0.812, SEM = 0.016), $F(2, 242) = 11.133$, $p<.001$, $\eta^2_p = 0.084$. Entropy was also lower for those using a 1D rule (M = 0.737, SEM = 0.017) than

for those using a 2D rule (M = 0.816, SEM = 0.015) or an unclassified rule (M = 0.900, SEM = 0.016), $F(2, 242) = 21.241$, $p<.001$, $\eta^2_p = 0.084$. Notably, as shown in Figure 9, the effect of rule use on entropy differed between rings conditions, $F(4, 242) = 3.116$, $p=.016$, $\eta^2_p = 0.049$. Entropy patterns in the no rings condition were similar to those observed in Experiment 1, with lower entropy for 1D users than for those using a 2D rule or an unclassified rule. Entropy differences between these rule-use groups were smaller in the random rings and informative rings conditions. Follow-up tests confirmed that, by the final learning block, there was no reliable entropy difference between those using 1D or 2D rules in either the random (1D: M = 0.693, SEM = 0.074; 2D: M = 0.798, SEM = 0.034), $F(1, 44) = 2.171$, $p=.148$, or the informative rings groups (1D: M = 0.628, SEM = 0.093; 2D: M = 0.717, SEM = 0.028, $F(1, 52) = 1.554$, $p=.218$.

In the no rings group, adding individual block 3 entropy scores to the regression equation significantly increased fit to the rule use data over an intercept-only model, $\chi^2(2, N=77) = 21.269$, $p <.001$. Lower levels of early entropy significantly increased the odds of using a 1D rule compared to a 2D rule, β = 2.940, SE = 1.133, OR = 18.908, z = 6.727, $p=.009$. No such predictive relationship, however, was found in the random ($p=.226$) or informative rings conditions ($p=.211$).

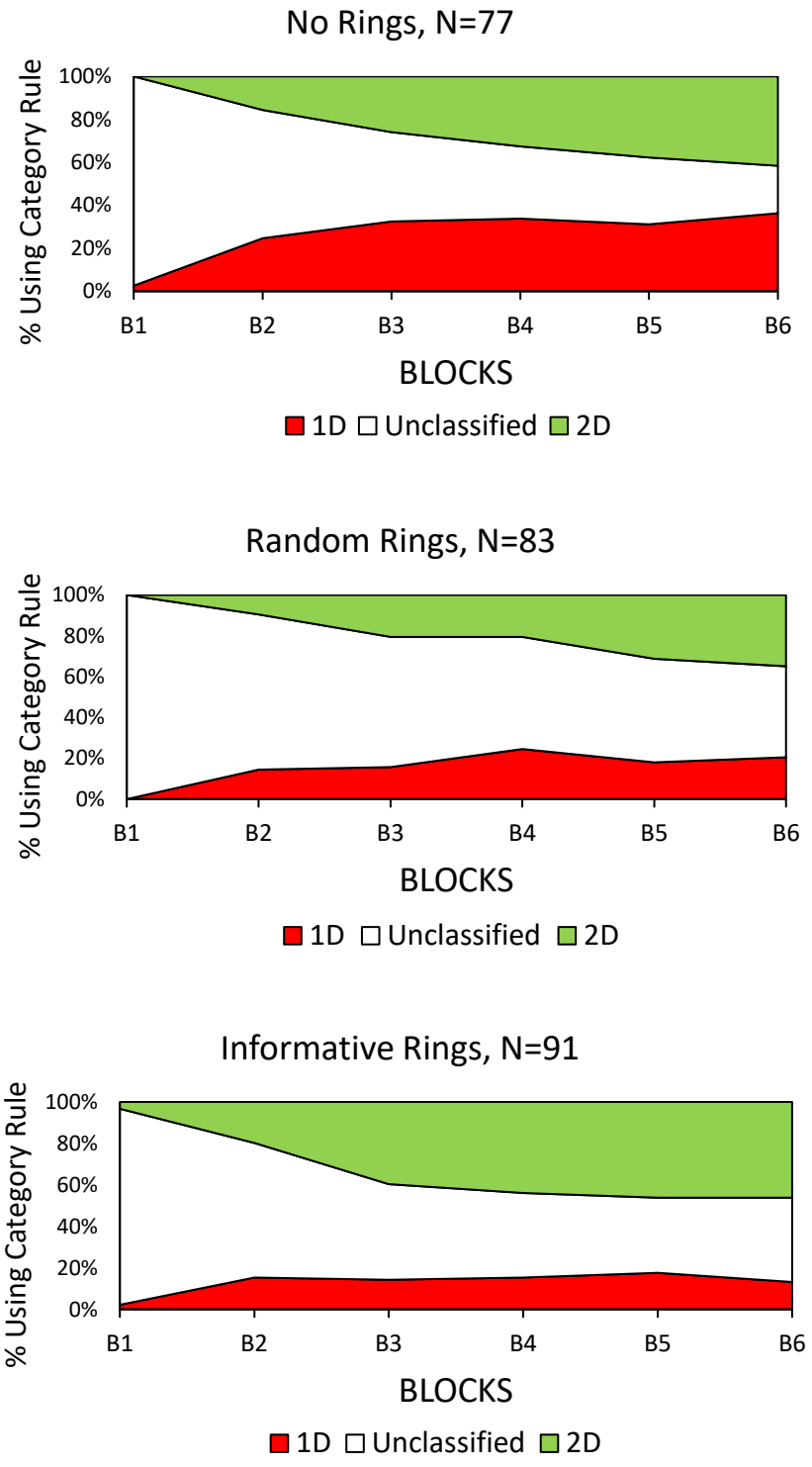
Summary

The addition of highlighting rings successfully altered patterns of selective visual attention. Adding rings to the features reduced the decline in entropy previously observed over the course of learning and reduced differences in entropy between the three sub-groups using different categorization rules. These effects were evident in both the random and informative rings conditions.

The alteration in patterns of visual attention in the random rings condition was associated with some reduction in use of a one-dimensional rule by the end of learning. However, in that condition, we also observed a reduction in the use of the optimal rule and a corresponding increase in the use of an unclassified rule. In the informative rings group, highlighting only relevant features reduced the proportion of learners who used a one-dimensional rule without impeding learning of the optimal rule.

Figure 7

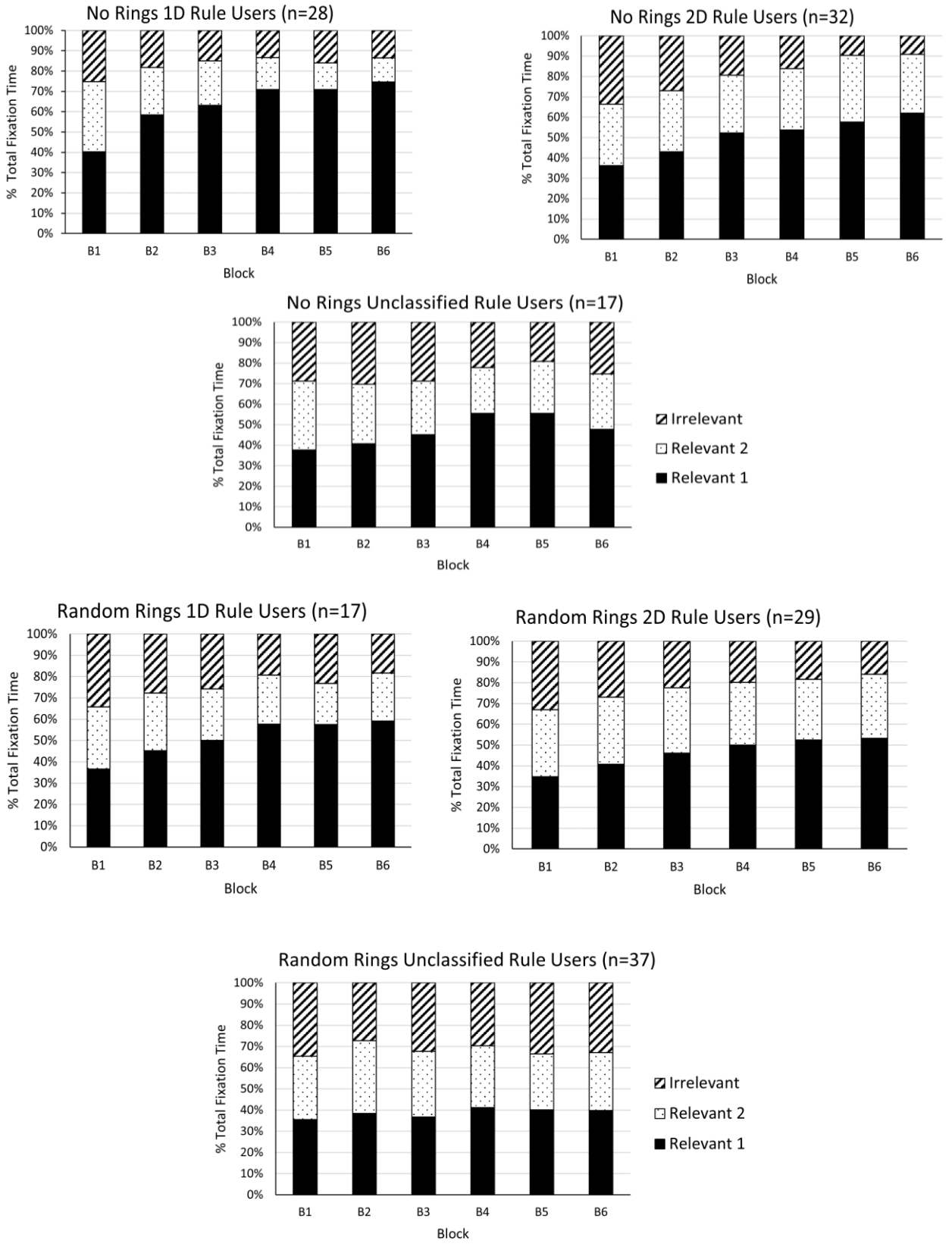
Experiment 2. Category rule use in each learning block

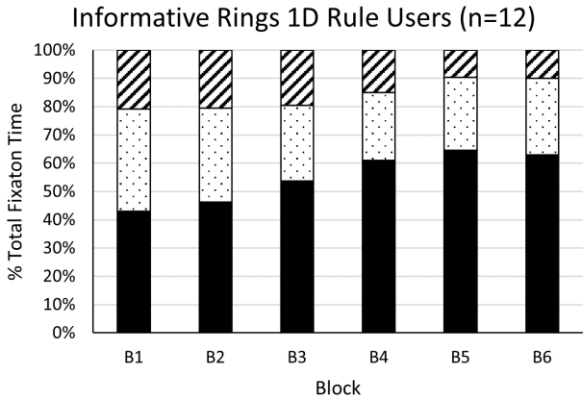


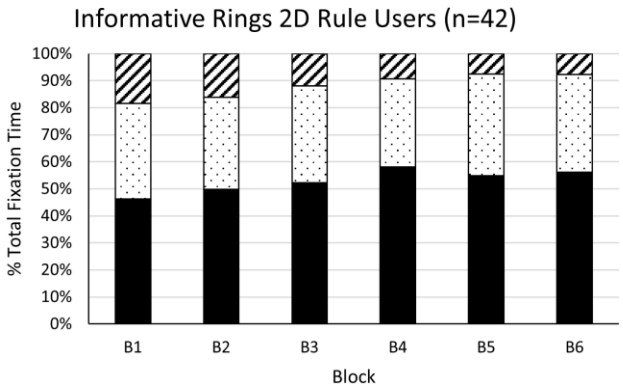
Note: The Figure shows the proportion of participants whose approach/avoidance decisions in a given block was consistent with the optimal two-dimensional categorization rule, a suboptimal one-dimensional rule or an unclassified rule. See the online article for the color version.

Figure 8

Experiment 2. Percentage of total fixation time attending to each feature per learning block




Informative Rings 1D Rule Users (n=12)
% Total Fixaton Time
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
B1
B2
B3
B4
B5
B6
Block


Informative Rings 2D Rule Users (n=42)
% Total Fixation Time
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
B1
B2
B3
B4
B5
B6
Block

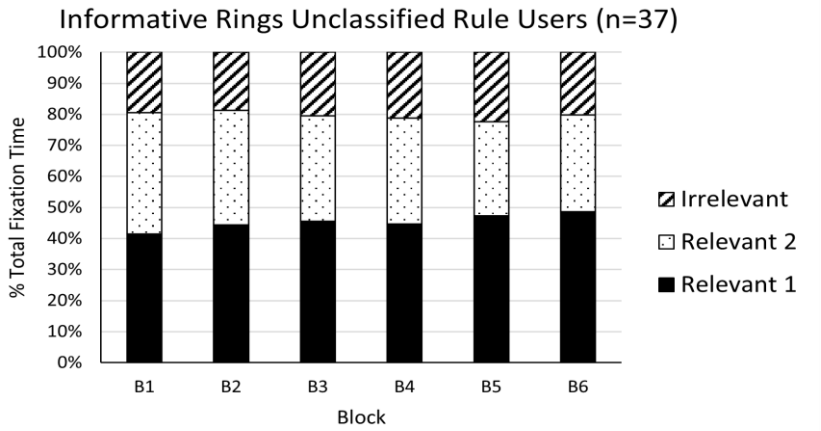
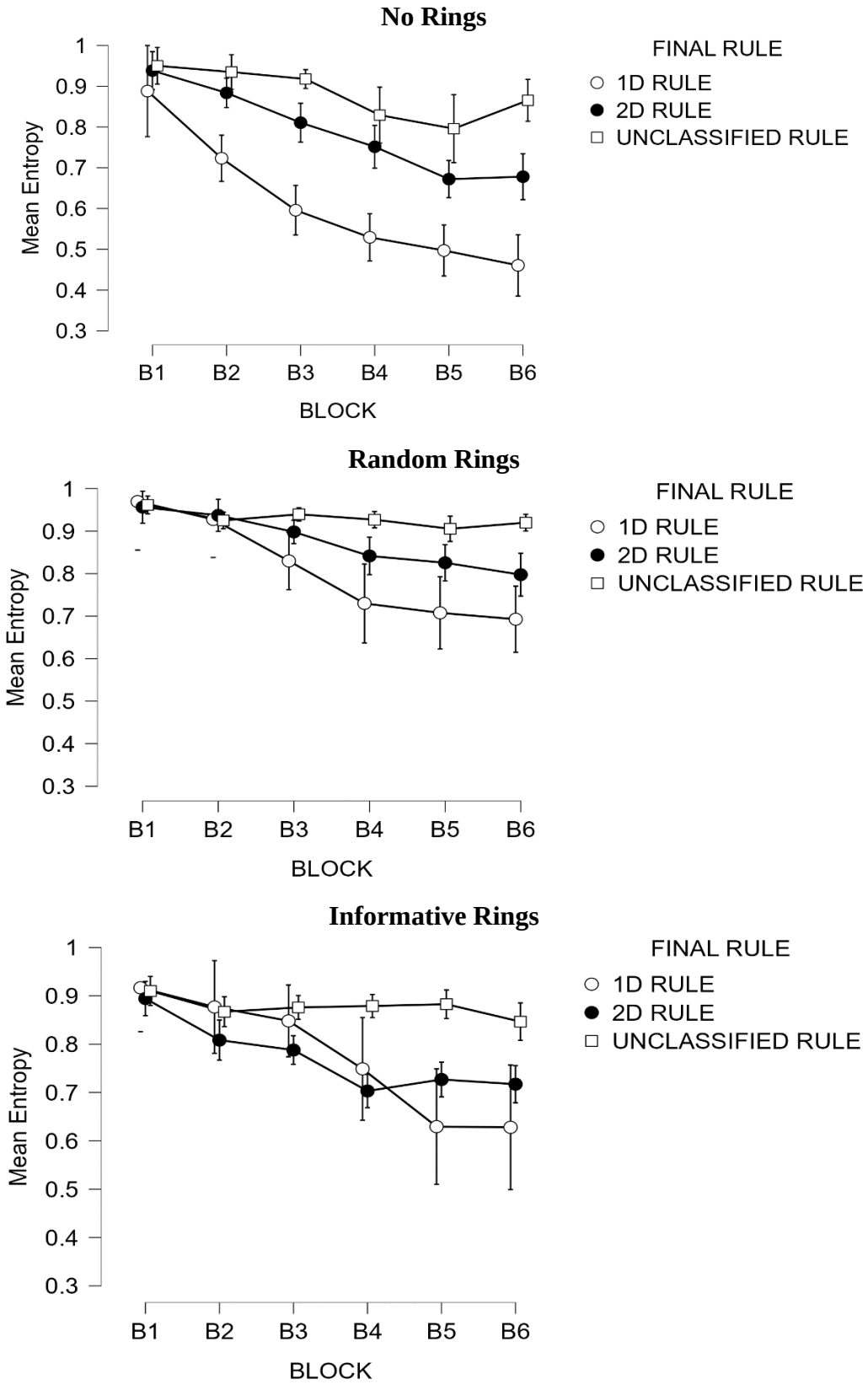

Informative Rings Unclassified Rule Users (n=37)
% Total Fixation Time
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
B1
B2
B3
B4
B5
B6
Block
Irrelevant
Relevant 2
Relevant 1

Figure 9

Experiment 2. Mean entropy across learning blocks for each rings condition



Note. Error bars represent 95% confidence intervals

General Discussion

Two studies examined the role of selective attention in learning traps and how this knowledge could be used to reduce trap prevalence. In each experiment, learners could decide to approach or avoid category exemplars in order to learn which types of stimuli led to gains or losses. Eye tracking allowed us to document changes in selective attention to exemplar features over the course of learning. Two feature dimensions were relevant for predicting category membership and associated gains and losses, while one was irrelevant. As in previous studies of selective attention during categorization (e.g., Blair et al., 2009a, b; Galdo et al., 2022; Rehder & Hoffman, 2005a, b), we found that most participants learned to reduce attention to the irrelevant feature.

In both studies, however, we also found a close correspondence between learners' pattern of approach/avoid decisions and their distribution of attention across feature dimensions. A key finding from Experiment 1 and the no rings condition in Experiment 2, was that those who fell into the trap of using a single dimension to guide their decisions, narrowed their visual attention to that dimension. By the end of learning, these participants spent little time gazing at the other dimension that was part of the optimal decision rule. Attentional narrowing commenced relatively early in learning. Individuals' pattern of selective attention in the first few blocks predicted the types of decision rules that they eventually acquired.

Experiment 2 examined whether the narrowing of attention had a direct causal role in the development of a simple-rules trap. In these experiments, we added a visual highlight designed to draw attention to multiple features. Our entropy measure showed that this successfully reduced attentional narrowing. However, the effect of highlighting on learning of category rules depended on whether the highlighting cue was presented for all feature dimensions or for predictive features only. In the former case, there was some reduction in the proportion of participants using a one-dimensional rule, but there was also a reduction in learning of the optimal two-dimensional rule. Those exposed to the random rings feature actually earned fewer rewards overall than those in a no rings baseline. In contrast,

highlighting only relevant features reduced trap prevalence without interfering with learning of the optimal rule. In other words, the highlighting cue was only effective in tackling traps when it also carried some information about which features were relevant for predicting outcomes.

Beliefs about Category Structure, Selective Attention and Learning Traps

Our results confirm the close coupling between selective attention and formation of learning traps hypothesized by Rich and Gureckis (2018) and illustrated in Figure 1. Previous work has used eye tracking to show that people selectively attend to feature dimensions that predict category outcomes (e.g., Blair et al., 2009a; Rehder & Hoffman, 2005a). Blanco et al. (2023) used eye tracking to show that variations in selective attention can predict whether or not learners detect a change in the predictive status of exemplar features. Our work, however, suggests that people often selectively narrow their attention to features based on a false or incomplete belief about the reward structure of the environment and that this attentional narrowing persists throughout subsequent learning. This relationship is reflected in the "top down" pathway linking beliefs about predictive features and selective attention shown in Figure 1.

Experiment 2 examined how manipulating the "bottom up" attention pathway in Figure 1 could impact trap formation. We found that boosting the salience of all features led to a broader distribution of feature attention, but this did not necessarily improve learning overall. Increasing the visual salience of all features made it harder for learners to discover the optimal rule. Although overly selective attention can lead to one to fall into a learning trap, learning of an optimal decision rule often requires *some* level of selective attention, allowing learners to filter out prediction-irrelevant features.

In many respects, our conclusions accord with recent attempts to model the relationships between beliefs or representation of the environment, selective attention and decision-making (Liu et al., 2024; Turner & Sloutsky, 2024; Turner et al., 2021). Turner et al. (2021) outlined a theoretical framework in which beliefs, attention and decisions are represented as three separate nodes in the learning process. A suite of possible models of the relations between these nodes was generated by

varying the causal pathways between them. For example, in the *Fully Constrained by Representation* model, beliefs about the predictive structure of the environment have a primary causal role, guiding both selective attention and decisions, with a subsidiary "bottom-up" route from attention to beliefs. These models were used to simulate human accuracy, eye gaze and response time data from a category learning study where a single visual feature always predicted category membership while other features either had a weaker probabilistic relationship or were not predictive (similar to Blanco et al., 2023). During the course of learning, there was an unsignaled change in the category structure such that the previous deterministic feature no longer predicted categorization and a previously irrelevant feature became a perfect predictor. Many participants fell into the trap of attending only to the deterministic feature early in learning, making it difficult for them to learn the new structure. In simulations of this data, the Fully Constrained model outperformed rival models including a *Representation Unconstrained by Attention* model, in which all features are encoded during learning. Follow-up modeling evaluated the relative contribution of the various casual pathways in the Fully Constrained model to fitting the learning and response time data. The most important pathway reflected the "top-down" effect of beliefs on the deployment of attention. The bottom-up route from attention to belief made a smaller but nonetheless significant contribution to data fit. This pattern of strong top-down effects and a weaker 'bottom up' signal converges with our findings in Experiment 2.

Although our primary focus was on those learning either an optimal two-dimensional rule or falling into the one-dimensional trap, it is also interesting to consider why a considerable proportion of participants in each study failed to discover either rule. We suspect that this is a heterogenous sub-group and there is no single answer to this question. However, our data provide some interesting clues. Our eye tracking data show that, unlike those who learned a one-dimensional rule, the unclassified subgroup attended to all three feature dimensions over the course of learning. This means they had more opportunity to discover the optimal rule than the 1D rule learners. The fact they did not acquire this rule suggests these participants may have had difficulty in learning the contingencies between

observed features and outcomes. Some support for this interpretation comes from recent modeling of participant learning data in a traps task that was similar to that used in the current studies. Liu et al. (2024) fitted a version of the Attention Learning Covering Map (ALCOVE, Kruschke, 1992) model of category learning from response-contingent feedback (Rich & Gureckis, 2018) to individual approach and avoid decisions. This model contains separate parameters that reflect a) attention to feature dimensions, and b) the learning of feature-outcome associations. Parameter comparison between sub-groups indicated that those using an unclassified rule distributed their attention more broadly than those using a one-dimensional rule, but showed poorer mapping of the relations between exemplar features and category membership.

The random rings condition in Experiment 2 showed that increasing attention to all features can exacerbate this learning problem. Drawing attention to irrelevant as well as relevant features may have led some people to search for predictive relations that did not exist, resulting in a failure to learn any identifiable rule.

Broad to Narrow Change in Attention during Category Learning

Our results are also helpful in adjudicating an ongoing debate about how selective attention shifts as people learn categories composed of multi-dimensional stimuli (see Wills et al., 2015 for a review). The *broad to narrow* view suggests that people commence learning by distributing their attention across a range of features, and gradually learn to selectively attend to those features that are the best predictors of category membership. This view has been proposed in many studies of visual attention during category learning (e.g., Blanco et al., 2023; Rehder & Hoffman, 2005a) and is implicit in categorization models such as the Generalized Context Model (GCM, Nosofsky, 1984) and ALCOVE. In contrast, the *narrow to broad* view, suggests that people begin learning by searching for simple category rules based on a single feature or small set of features. If simple rules lead to categorization errors, then additional dimensions are added. This is the pattern suggested by models such as the rules-plus-exceptions RULEX model (Navarro, 2015; Nosofsky et al., 1994). Within this

framework, learning traps would arise when there is an insufficient expansion of attention across stimulus dimensions.

Both our attentional and learning data are more consistent with the broad-to-narrow view. Most participants began by attending to all three features and narrowed their attention over the course of learning. To quantify this trend, for each participant in the two experiments, we calculated the difference in entropy scores between the first and last learning block. Overall, a large majority (82%) showed a reduction in entropy over the course of learning, reflecting a progressive narrowing of attention. Likewise, inspection of individual rule use across learning blocks showed that it was relatively rare for learners to start with a narrow one-dimensional rule and then later shift to a two-dimensional rule. Across studies, only 14% of participants showed this pattern.

Using Attentional Highlighting to Overcome Learning Traps

Learning traps can have a pernicious and persistent effect on decision-making in many domains outside the laboratory. Our results provide some useful guidance for what sorts of strategies might be used to prevent traps from forming. Repeatedly drawing attention to *all* features of a decision option may reduce the tendency to selectively attend to a subset of relevant features. However, our results show that this might have the negative side-effect of increased processing of features that have no predictive value, obscuring the path to the optimal decision strategy. Drawing attention to relevant features only, however, reduced trap formation without this side-effect.

To our knowledge, the approach exemplified in our informative rings group represents the first successful attempt to prevent the emergence of a learning trap. Rich and Gureckis (2018) attempted to reduce trap prevalence by 1) adding individuating features to each training item, 2) presenting just a single prediction-relevant or -irrelevant feature on some trials, or 3) adding noise to the categorization rule, so that the outcomes predicted by feature conjunctions was switched on ten percent of trials. The first two strategies produced little change in trap formation compared to a contingent feedback baseline. The third strategy led to some reduction in use of a one-dimensional rule, but also suppressed

learning of the optimal two-dimensional rule – similar to the results of our random rings group.

Although our approach of highlighting attention to informative features seems promising, there are challenges in translating this into a more general strategy for countering learning traps. The approach requires that someone has control over the design of the decision architecture and that the most relevant features for predicting choice outcomes are known. This may be possible in some environments. For example, in defense, intelligence and finance, decision-makers are often confronted with multiple sources of information that may or may not be relevant to making a consequential decision. Where it is known that some of these cues are especially helpful in making decisions, an interface can be designed that highlights these cues; the goal being to reduce neglect of less obvious but relevant cues or cue combinations (e.g., Mandel et al., 2023). Of course, such control over the decision environment may not always be possible. Hence, it is worthwhile to continue to search for other strategies that might help overcome learning traps. This might include providing additional feedback about choice outcomes (J. Lee et al., 2023) or collaboration between learners in deciding which stimuli to explore (Budiono et al., 2024).

Constraints on Generality

Learning traps are thought to arise through the operation of fundamental psychological processes such as selective attention and generalization from early learning. Hence, although the current studies were carried out with a mixed gender sample of university undergraduates, we expect that the key characteristics of learning traps, such as selective attention to a subset of predictive features, will generalize to many other populations. Evidence for this comes from studies demonstrating similar levels of learning trap prevalence in participants recruited from online crowdsourcing platforms, who were more diverse in age and educational background than the current samples (Li et al., 2021; Lee et al., 2024, also see Footnote 1). Our novel finding that traps can be prevented by promoting attention to a range of informative features requires replication with these more diverse populations. Again, however, because of the fundamental nature of the underlying

processes, we predict that these results will show considerable generality. Notably, visual search based on the learner's expectations (i.e., top-down attention) shows relatively little age-change during adulthood (Madden, 2007). Hence, our remediation of traps via manipulation of selective attention may also prove effective with older adults.

There are, however, some populations that are likely to show a very different pattern of responses. Young children, aged 4-5 years, have been shown to show less proclivity to fall into learning traps than older children and adults (Blanco & Sloutsky, 2019; Liquin & Gopnik, 2022). Such children also often show less intentional control over their attentional deployment (Hanania & Smith, 2010). Hence, we expect that they will show lower levels of trap prevalence than those we observed and that attempts to promote a broader distribution of attention via manipulation of feature salience would be less effective. Likewise, our attentional intervention is less likely to benefit those with clinical disorders that involve impaired attentional control (e.g., Attention Deficit Hyperactivity Disorder).

Conclusions

Many consequential decisions involve observing the features associated with choice outcomes and learning which feature combinations predict positive or negative outcomes. The current studies showed how selective attention to a subset of features contributes to suboptimal decision rules that persist despite further learning. We found that drawing attention to neglected features can help to overcome this simple-rules learning trap, but only when these features are relevant for outcome prediction. This work extends our understanding of the cognitive mechanisms that give rise to learning traps and suggests how we can use this knowledge to overcome such traps. The simple-rules trap that was the focus of the current work, however, is just one way that under-exploration of the environment early in the learning process can lead us to decision-making strategies that are sub-optimal in the long term (see Bai et al., 2022; Denrell & March, 2001; Harris et al., 2020; Pilditch & Custers, 2018 for other examples). An important goal for future work is to examine the role of selective attention in these other types of learning traps.

References

Awh, E., Belopolsky, A. V., & Theeuwes, J. (2012). Top-down versus bottom-up attentional control: A failed theoretical dichotomy. *Trends in Cognitive Sciences*, *16*(8), 437–443.

Bai, X., Fiske, S. T., & Griffiths, T. L. (2022). Globally inaccurate stereotypes can result from locally adaptive exploration. *Psychological Science*, *33*(5), 671-684.

Bhatnagar, R., & Orquin, J. L. (2022). A meta-analysis on the effect of visual attention on choice. *Journal of Experimental Psychology: General*, *151*(10), 2265-2283. https://doi.org/10.1037/xge0001204

Blair, M. R., Watson, M. R., Walshe, R. C., & Maj, F. (2009a). Extremely selective attention: eye-tracking studies of the dynamic allocation of attention to stimulus features in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1196-1206. https://psycnet.apa.org/doi/10.1037/a0016272

Blair, M. R., Watson, M. R., & Meier, K. M. (2009b). Errors, efficiency, and the interplay between attention and category learning. *Cognition*, *112*(2), 330-336.

Blanco, N. J., & Sloutsky, V. M. (2019). Adaptive flexibility in category learning? Young children exhibit smaller costs of selective attention than adults. *Developmental psychology*, *55*(10), 2060-2076.

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. https://doi.org/10.1016/j.jecp.2022.105548

Brainard D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10, 433–436.

Budiono, R., Hartley, C. A., & Gureckis, T. M. (2024). How does social learning affect stable false beliefs? 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*. pp.

Chater, N., & Vitányi, P. (2003). Simplicity: a unifying principle in cognitive science? *Trends in*

Cognitive Sciences, *7*(1), 19-22. https://doi.org/https://doi.org/10.1016/S1364-6613(02)00005-0

Chun, M. M., Golomb, J. D., & Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual Review of Psychology*, *62*(1), 73-101.

Denrell, J. (2005). Why most people disapprove of me: Experience sampling in impression formation. *Psychological Review, 112*(4), 951– 978. https://doi.org/10.1037/0033-295X.112.4.951

Denrell, J., & March, J. G. (2001). Adaptation as information restriction: The hot stove effect. *Organization Science*, *12*(5), 523–538. http://dx.doi.org/ 10.1287/ orsc.12.5.523.10092

Elwin, E. (2013). Living and learning: Reproducing beliefs in selective experience. *Journal of Behavioral Decision Making*, *26*(4), 327-337.

Fisher, G. (2021). Intertemporal choices are causally influenced by fluctuations in visual attention. *Management Science*, *67*(8), 4961-4981.

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. https://doi.org/10.1016/j.cogpsych.2022.101508

Hanania, R., & Smith, L. B. (2010). Selective attention and attention switching: Towards a unified developmental approach. *Developmental Science*, *13*(4), 622-635.

Harris, C., Fiedler, K., Marien, H., & Custers, R. (2020). Biased preferences through exploitation: How initial biases are consolidated in reward-rich environments. *Journal of Experimental Psychology: General*, *149*(10), 1855-1877.

IBM (2024). *IBM SPSS Statistics for Windows, (Version 29.0.1)* [Computer software]. Armonk, NY: IBM Corp

JASP Team (2024). *JASP (Version 0.19.0)* [Computer software].

Kleiner M., Brainard D. H., & Pelli D. G. (2007). *What's new in Psychtoolbox-3*? Paper presented at

the European Conference on Visual Perception, Arezzo, Italy

Krajbich, I., & Rangel, A. (2011). Multi-alternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, *108*(33), 13852-13857.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychology Review*, *99*, 22-44.

Lee, J. E., Li, A. X., & Hayes, B. K. (2023). Overcoming learning traps with summative feedback. Unpublished preprint. *PsyArXiv,* doi: 10.31234/osf.io/kztxj

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*. Accepted 30/06/2024

Li, A. X., Gureckis, T. M., & Hayes, B. (2021). Can losses help attenuate learning traps? In T. Fitch et al. (Eds.) *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*, pp. 1201-1207.

Liquin, E. G., & Gopnik, A. (2022). Children are more exploratory and learn more than adults in an approach-avoid task. *Cognition*, *218*, 104940. https://doi.org/https://doi.org/10.1016/j.cognition.2021.104940

Liu, Y., Newell, B. R., Lee, J. E., & Hayes, B. K. (2024). Examining the Relationship Between Selective Attention and the Formation of Learning Traps. 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*. pp. 4632-4638.

Madden, D. J. (2007). Aging and visual attention. *Current Directions in Psychological Science*, *16*(2), 70-74.

Mandel, D. R., Irwin, D., Dhami, M. K., & Budescu, D. V. (2023). Meta-informational cue inconsistency and judgment of information accuracy: Spotlight on intelligence

analysis. *Journal of Behavioral Decision Making*, *36*(3), e2307.

MathWorks Inc. (2022). MATLAB version: 9.13.0 (R2022b), Natick, Massachusetts: The MathWorks Inc. https://www.mathworks.com

Milosavljevic, M., Navalpakkam, V., Koch, C., & Rangel, A. (2012). Relative visual saliency differences induce sizable bias in consumer choice. *Journal of Consumer Psychology, 22*, 67-74.

Mormann, M., & Russo, J. E. (2021). Does attention increase the value of choice alternatives? *Trends in Cognitive Sciences, 25*(4), 305-315.

Müller, H. J., Geyer, T., Zehetleitner, M., & Krummenacher, J. (2009). Attentional capture by salient color singleton distractors is modulated by top–down dimensional set. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(1), 1−16

Navarro, D. J. (2005). Analyzing the RULEX model of category learning. *Journal of Mathematical Psychology*, *49*(4), 259-275.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 104–114.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101*, 53–79.

Orquin, J. L., Lahm, E. S., & Stojić, H. (2021). The visual environment and attention in decision making. *Psychological Bulletin*, *147*(6), 597-617.

Pilditch, T. D., & Custers, R. (2018). Communicated beliefs about action-outcomes: The role of initial confirmation in the adoption and maintenance of unsupported beliefs. *Acta Psychologica, 184*, 46-63.

Rehder, B., & Hoffman, A. B. (2005a). Eye tracking and selective attention in category learning. *Cognitive Psychology*, *51*(1), 1-41.

Rehder, B., & Hoffman, A. B. (2005b). Thirty-something categorization results explained: selective attention, eyetracking, and models of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 811-829.

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), 1553–1570. https://doi.org/10.1037/xge0000466

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye tracking Research & Applications* (pp. 71-78). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/355017.355028

Schulz, E., Konstantinidis, E., & Speekenbrink, M. (2018). Putting bandits into context: How function learning supports decision making. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *44*(6), 927–943. https://doi.org/10.1037/xlm0000463

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*, 1– 42. doi:10.1037/h0093825

Shimojo, S., Simion, C., Shimojo, E., & Scheier, C. (2003). Gaze bias both reflects and influences preference. *Nature Neuroscience*, *6*(12), 1317-1322.

Smith, S. M., & Krajbich, I. (2019). Gaze amplifies value in decision making. *Psychological Science*, *30*(1), 116-128.

Sui, X. Y., Liu, H. Z., & Rao, L. L. (2020). The timing of gaze-contingent decision prompts influences risky choice. *Cognition*, *195*, 104077.

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

Teodorescu, K., and Erev, I. (2014b). Learned helplessness and learned prevalence: exploring the causal relations among perceived controllability, reward prevalence, and exploration. *Psychological Science, 25*, 1861–1869.

Turner, B. M. & Sloutsky, V. M. (2024). Cognitive inertia: cyclical interactions between attention and memory shape learning. *Current Directions in Psychological Science, 1-8*, https://doi.org/10.1177/09637214231217989

Turner, B. M., Kvam, P. D., Unger, L., Sloutsky, V., Ralston, R., & Blanco, N. J. (2021). Cognitive inertia: How loops among attention, representation, and decision-making distort reality. *PsyArXiv*

Wills, A. J., Inkster, A. B., & Milton, F. (2015). Combination or differentiation? Two theories of processing order in classification. *Cognitive Psychology*, *80*, 1-33.

Wilson, R. C., Bonawitz, E., Costa, V. D., & Ebitz, R. B. (2021). Balancing exploration and exploitation with information and randomization. *Current Opinion in Behavioral Sciences, 38*, 49-56.

Yang, X., & Krajbich, I. (2023). A dynamic computational model of gaze and choice in multi-attribute decisions. *Psychological Review*, *130*(1), 52-70.