

1 An inductive bias for slowly changing features in human 2 reinforcement learning

3 Noa L. Hedrich^{1,2,3,4*}, Eric Schulz^{5,6}, Sam Hall-McMaster^{1,7,¶}, Nicolas W. Schuck^{1,2,4,¶}

- 4 1. Max Planck Research Group NeuroCode, Max Planck Institute for Human Development, Berlin,
5 Germany.
- 6 2. Max Planck UCL Centre for Computational Psychiatry and Ageing Research, Berlin, Germany.
- 7 3. Einstein Center for Neurosciences Berlin, Charité Universitätsmedizin Berlin, Berlin, Germany.
- 8 4. Institute of Psychology, Universität Hamburg, Hamburg, Germany.
- 9 5. Max Planck Research Group Computational Principles of Intelligence, Max Planck Institute for
10 Biological Cybernetics, Tübingen, Germany.
- 11 6. Helmholtz Institute for Human-Centered AI, Helmholtz Center Munich, Neuherberg, Germany.
- 12 7. Department of Psychology, Harvard University, United States of America.

13 * Corresponding Author

14 E-mail: noa.hedrich [at] uni-hamburg.de (NLH)

15 ¶ SHM and NWS are Joint Senior Authors.

16

Abstract

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

Identifying goal-relevant features in novel environments is a central challenge for efficient behaviour. We asked whether humans address this challenge by relying on prior knowledge about common properties of reward-predicting features. One such property is the rate of change of features, given that behaviourally relevant processes tend to change on a slower timescale than noise. Hence, we asked whether humans are biased to learn more when task-relevant features are slow rather than fast. To test this idea, 100 human participants were asked to learn the rewards of two-dimensional bandits when either a slowly or quickly changing feature of the bandit predicted reward. Participants accrued more reward and achieved better generalisation to unseen feature values when a bandit's relevant feature changed slowly, and its irrelevant feature quickly, as compared to the opposite. Participants were also more likely to incorrectly base their choices on the irrelevant feature when it changed slowly versus quickly. These effects were stronger when participants experienced the feature speed before learning about rewards. Modelling this behaviour with a set of four function approximation Kalman filter models that embodied alternative hypotheses about how feature speed could affect learning revealed that participants had a higher learning rate for the slow feature, and adjusted their learning to both the relevance and the speed of feature changes. The larger the improvement in participants' performance for slow compared to fast bandits, the more strongly they adjusted their learning rates. These results provide evidence that human reinforcement learning favours slower features, suggesting a bias in how humans approach reward learning.

36

Author Summary

37

38

39

40

41

42

43

44

45

46

47

48

49

Learning experiments in the laboratory are often assumed to exist in a vacuum, where participants solve a given task independently of how they learn in more natural circumstances. But humans and other animals are in fact well known to “meta learn”, i.e. to leverage generalisable assumptions about *how to learn* from other experiences. Taking inspiration from a well-known machine learning technique known as slow feature analysis, we investigated one specific instance of such an assumption in learning: the possibility that humans tend to focus on slowly rather than quickly changing features when learning about rewards. To test this, we developed a task where participants had to learn the value of stimuli composed of two features. Participants indeed learned better from a slowly rather than quickly changing feature that predicted reward and were more distracted by the reward-irrelevant feature when it changed slowly. Computational modelling of participant behaviour indicated that participants had a higher learning rate for slowly changing features from the outset. Hence, our results support the idea that human reinforcement learning reflects a priori assumptions about the reward structure in natural environments.

50 Introduction

51 A remarkable amount of information is reaching our senses at any given time, yet often only a small
52 subset of it is relevant to our current goal. Determining which aspects of our environment are relevant
53 is therefore a crucial challenge for learning goal-directed behaviour. But addressing this challenge is
54 hard. The space of possibilities is often too large to be explored fully within the time limits we need
55 to consider, and yet limiting attention to only a subset of features risks ignoring relevant information
56 [1, 2]. One approach to this problem is to not learn every problem anew, but instead use knowledge of
57 properties that have been relevant in the past as a starting point, in the form of so-called priors, also
58 known as inductive biases [3–7]. Here, we study the role of one such prior in human learning, namely a
59 bias to focus learning on slowly changing features in our environments, and their potential association
60 to rewards.

61 Analogous to the concept of a ‘prior’ in Bayesian statistics, priors are pre-existing beliefs about
62 the underlying structure of an environment, based on generalised past experiences or evolutionary
63 transmission [3, 8]. Previous research has shown that priors can expedite the learning process by
64 focusing information processing on what is common across many environments [4, 9, 10]. The resulting
65 decision-making biases are numerous [10–13] and can for instance be observed in the form of adaptive
66 heuristics that reflect constraints on time or resources [14], or in the form of visual illusions that reflect
67 the simplifying assumptions of our visual system, such as that light tends to come from above [15].
68 Studying useful priors for representation learning is also an active field of development in artificial
69 intelligence [8, 16–18], in particular for reinforcement learning (RL), where knowledge about which
70 actions maximise reward and minimise punishment is acquired through a trial-and-error process [19].
71 While the RL framework has been very successful in furthering our understanding of learning and
72 decision-making, [20–23], it becomes notoriously inefficient in high dimensional environments [19].
73 This problem can be alleviated through a process known as representation learning, where learning is
74 limited to a subset of features that help predict future rewards, known as task states [19, 24–28]. The
75 difficulties of learning the state space for each new problem *de novo* have been widely recognized [29],
76 underscoring the potential benefit of leveraging prior knowledge.

77 A useful prior for reinforcement learning should therefore help quickly build appropriate task states
78 from rich perceptual observations in novel environments [8, 30]. A characteristic shared across many
79 environments is that the causal process generating observations develops on a slower timescale than
80 the sensory signals we observe [31–33]. For example, the appearance of a ball flying toward you in a
81 park might fluctuate rapidly as it passes through patches of sun and shade, but its true colour will
82 remain unchanged. Similarly, other relevant properties such as its speed and trajectory will change
83 in a slower, continuous manner compared to low-level perceptual features. In short, signal tends to
84 vary more slowly than noise [34]. It follows that a way to extract features relevant to building task
85 states, while remaining impartial to the exact nature of those features or the causal process underlying
86 the perceptual observations, is to focus on slowly changing features. Indeed, some research has shown
87 that humans have a bias toward perceiving slower speeds in the spatial domain [34–36]. This idea
88 has gained more traction in machine learning, where a slowness prior has been shown to enable the
89 discovery of task states from raw observations [8, 28, 37, 38].

90 A well-known implementation of this prior is Slow Feature Analysis (SFA), an unsupervised learn-
91 ing algorithm that reduces the dimensionality of its input by identifying slowly changing dimensions
92 in the data [31, 39, 40]. SFA first isolates independent components in the data and then extracts
93 those components that change slowly, under the premise that slower features are more meaningful
94 representations of the data [31]. This insight has been shown to be relevant for RL, for instance in the
95 context of a spatial learning task where SFA can provide an effective representation learning mechanism
96 [41]. The same paper showed that the SFA agent produced similar learning trajectories to rats solving
97 a comparable task, underscoring the relevance of a slowness prior for animals. Theoretical research
98 also demonstrated that extracting slow features can explain the activity of complex cells in the visual
99 cortex, the formation of allocentric spatial maps in the hippocampus and can be implemented in a bi-
100 logically plausible network [42–46]. Hence, a slowness prior promises a domain-general and biologically
101 plausible way to extract task states from environmental input.

102 Despite its potential for representation learning and the abundance of research in the machine
103 learning domain, studies on the slowness prior in human reinforcement learning are largely lacking.
104 Here we explored the idea that humans rely on a slowness prior during reinforcement learning. We
105 developed a novel decision-making task, in which participants had to repeatedly learn which of two

106 stimulus features predicted reward. We manipulated the speed of change of the features and asked
107 whether participants were faster to learn when the relevant feature changed slowly versus when it
108 changed quickly. Across two studies and extensive model comparison, our results indicate that they
109 do. This finding enriches our understanding of human inductive biases in RL and can prompt further
110 studies about other such biases in human learning, as well as inform artificial intelligence about how
111 to best build human-like agents.

112 Results

113 We investigated whether humans have a prior to preferentially process slowly changing features of the
114 environment that impacts reinforcement learning. We hypothesised that given such a prior, partici-
115 pants should be better at learning the task if reward-predictive features changed slowly, rather than
116 quickly. To test this, we developed a task that required participants to learn the rewards associated
117 with a set of visual stimuli characterized by two features, a colour and a shape (Fig 1a). During each
118 trial of learning, participants saw a stimulus composed of both features and decided between rejecting
119 or accepting the stimulus. While rejecting always led to a fixed reward of 50 coins, accepting led to
120 reward between 0 to 100 coins that was higher than 50 for half of all stimuli. Across trials, the two
121 features changed independently and with different speeds: one feature changed slowly (e.g., partici-
122 pants saw relatively similar shapes from trial to trial), while the other feature changed quickly (e.g.,
123 participants saw relatively distinct colours from trial to trial, Fig 1a). Our core manipulation was that
124 in each block either the slowly-changing or the fast-changing feature was reward-predictive, while the
125 other had no relation to reward (relevant and irrelevant feature, respectively). The relevant feature
126 had a fixed relation to reward in each block, with the maximum reward of 100 assigned to one position
127 and decreasing rewards assigned to other positions based on their distance to the maximum. This split
128 the circular feature space into two semicircles: high- and low-reward (Fig 1b). Hence participants had
129 to learn which feature was reward-predictive in general, and which specific feature positions should be
130 accepted vs rejected.

131 We conducted a pilot experiment and a main experiment, each with 50 participants. The key differ-
132 ence between the pilot and main experiments was that the main experiment included a demonstration
133 of stimulus changes before each block. Hence, in the pilot experiment participants directly started
134 reward learning, and could observe which feature changed fast vs. slow while they also had to observe
135 the reward outcomes. In the main experiment, we ensured participants knew how fast each feature
136 would change *before* each block by displaying a sequence of 30 trials without reward that participants
137 observed passively before learning (*Observation phase*, see Fig 1d). Participants were not informed
138 about which feature was relevant in either experiment but had to learn it in each block through trial
139 and error from the *Learning phase*, as described above (pilot experiment: 45 trials, main experiment:
140 60 trials, Fig 1e). Due to the continuous reward structure, it was beneficial to generalise observed out-
141 comes to nearby feature positions. We probed generalisation of learned values at the end of each block
142 in a *Test phase* in which participants were asked to choose the more valuable stimulus among pairs of
143 stimuli not seen during learning, without feedback (pilot experiment: 15 trials, main experiment: 36
144 trials, Fig 1f).

145 Participants performed eight blocks in total. In half of the blocks the slow feature was reward-
146 predictive (slow blocks), in the other half the fast feature was reward-predictive (fast blocks, Fig 1c).
147 Within each of these conditions, colour and shape were assigned as the relevant feature an equal
148 number of times.

149 Participants learned feature rewards and generalised their knowledge

150 We first analysed participant choices to confirm learning of the feature-reward mapping. In the main
151 experiment, participants' choice accuracy on the learning task increased from an average of 51% in the
152 first ten trials of a block to 74% in the last ten trials ($t(49) = 13.699$ $p < .001$, Fig 2a). This increase
153 in accuracy was accompanied by a gradual decrease in 'accept' choices throughout the learning phase,
154 reducing from 86% in the first ten trials to 61% in the last ten trials ($t(49) = -12.755$ $p < .001$,
155 Fig 2b). Note that 'accept' choices allowed participants to gather information on stimulus values and
156 therefore were necessary for exploration early in a block. Accordingly, participants learned with time
157 to selectively reject low-value stimuli, while they continued to accept high-value stimuli (Fig 2c). We
158 confirmed participants did not engage in simplified strategies by fitting two control models, one which
159 captures possible biases for accept choices (Random Choice model), and one which can capture a bias
160 for one of the response keys (Random Key model). These models did not explain participant choices
161 well, compared to the learning models discussed below (Fig 2a, details see below and Methods). These
162 results show that participants learned the feature-reward mapping and are consistent with data from
163 the pilot experiment, see S1 Fig

164 We also found that participants could correctly identify the higher value stimulus in the test phase,
165 in which previously unseen feature positions were presented, for which participants never witnessed

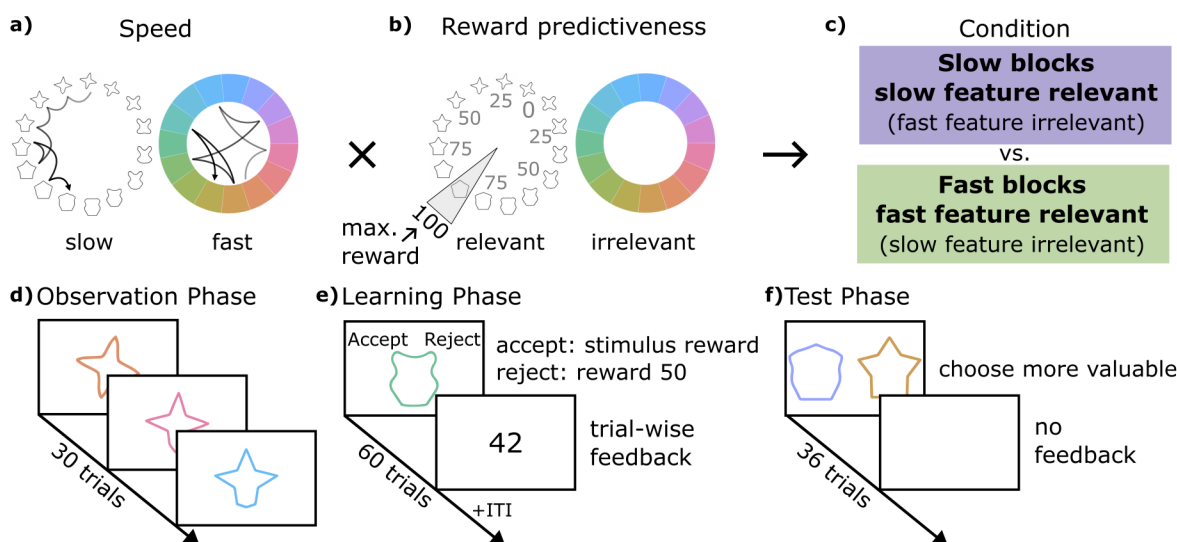


Figure 1: Continuous reward features learning task. **a)** The two stimulus features and their possible speeds. Each jump of the arrows indicates the change in the feature on a trial. The slow feature (here: shape) changes gradually, while the fast feature (here: colour) changes randomly. The feature-speed mapping is only for illustration, in each block, either shape or colour could change slowly. **b)** The mapping of reward onto the relevant feature space. The relevant feature (here: shape) determines the stimulus reward. The closer the stimulus shape is to the maximum reward location, the higher the reward. The irrelevant feature (here: colour) was uncorrelated with reward. The feature-reward mapping is only for illustration, in each block, either shape or colour could be relevant and the maximum reward location changed. **c)** How feature speed and reward predictiveness were combined to form slow and fast blocks. Note that which feature was slow/relevant was counterbalanced across blocks. **d-f)** Schematic of the three phases in each task block in the main experiment. In the pilot experiment, the observation phase **d)** was omitted.

166 reward feedback (mean accuracy 75% significantly higher than the chance level of 50% $t(49) = 17.378$,
 167 $p < .001$). Further, participant choice probabilities reflected true stimulus values (Fig 2d). Performance
 168 during the test phase did not differ statistically from end-of-learning performance in the learning phase
 169 ($t(49) = -1.48$, $p = .143$). Hence, our data suggests that participants generalised values successfully
 170 across task and stimulus differences between the two phases. These results were a replication of what
 171 we observed in the pilot experiment, see S1 Fig

172 Performance improved when the relevant feature changed slowly

173 Having established that participants learned and generalised well in our task, we turned to our
 174 main question, namely, whether reward learning and generalisation differed for slowly versus fast-
 175 changing features. The hypothesis and main analyses were preregistered prior to data collection
 176 (<https://osf.io/6dy8f>). Note that some changes were made to the design and follow-up analyses after
 177 the preregistration (e.g. ANOVAs were replaced with linear mixed effect models). None of these
 178 changes were material to the main conclusions of our paper. For specific changes in the rationale
 179 behind them, see S1 Text. All mixed effect models used the maximal random effects structure that
 180 converged. We first included all main effects and interactions between predictors in the fixed effects
 181 and sequentially removed all terms that did not significantly improve the model. Predictors were z-
 182 scored and no response trials were excluded, see Methods for details. Full model descriptions including
 183 effect sizes and confidence intervals can be found in S2 - S7 Tables.

184 **Improved learning** We measured performance in the learning phase by subtracting the cumulative
 185 reward expected by chance (50 per trial) from the cumulative reward obtained by participants.
 186 In line with our hypothesis, the cumulative reward gain was higher in slow compared to fast blocks
 187 ($M_S = 248.62 \pm 21.54$, $M_F = 217.57 \pm 22.43$, $t(49) = 2.17$, $p_{1-sided} = .017$, $d = 0.31$, Fig 2e). To

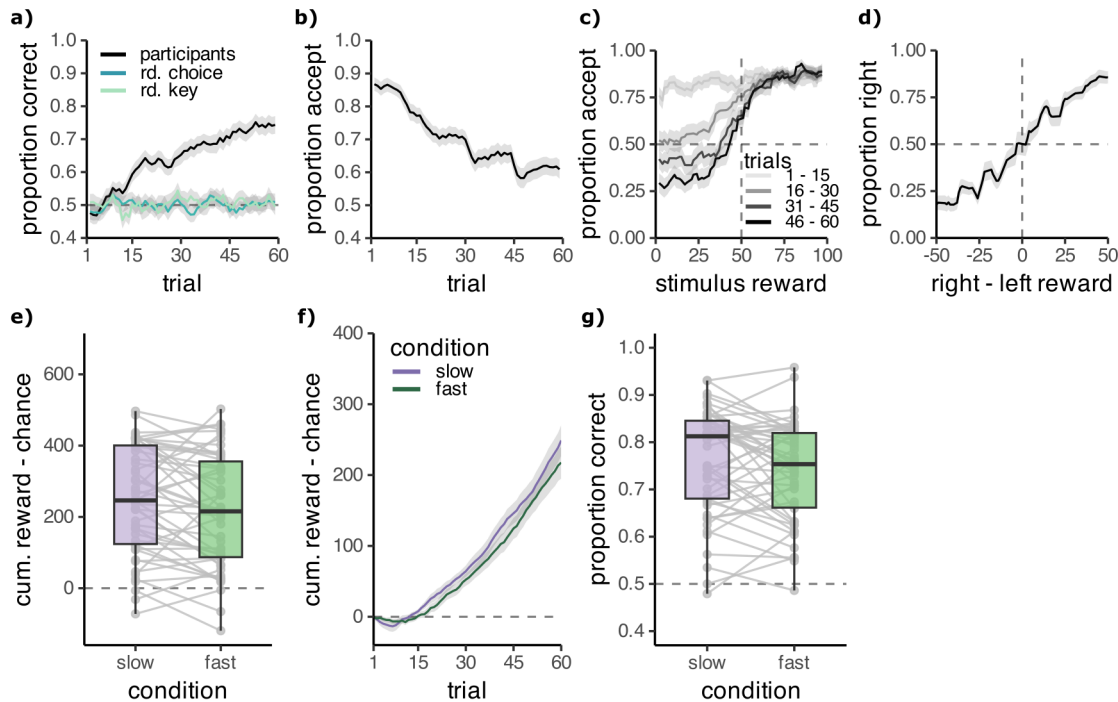


Figure 2: Participants performed better in slow blocks. **a)** Proportion correct choices across trials in the learning phase. The behaviour of two control models which capture aspects of random behaviour are shown in blue/green colours. **b)** The proportion of accept choices in the learning phase reduces across trials. **c)** The proportion of accept choices depending on the true stimulus reward, for every 15 trials from the start to the end of the block. Participants learn to selectively reject low-value stimuli. **a-c)** Curves were averaged across 3 adjacent values. **d)** Proportion of choosing the right stimulus in the test trials, depending on the difference in value between the right and left stimulus, shows sensitivity to the true reward value. Curves were averaged across 5 adjacent values. **e)** Cumulative reward obtained in a block of the learning phase above a chance baseline of 50 per trial is higher in slow than in fast blocks. **f)** Cumulative reward obtained relative to a chance baseline of 50 on each trial increases more rapidly in slow blocks. **g)** Mean accuracy in the test phase is higher in slow than in fast blocks. **e-g)** separately for blocks where the slow feature (purple) and fast feature (green) were relevant. Individual participants in grey. Grey ribbons show the standard error of the mean.

188 test more specifically whether the rate at which participants accumulated reward was greater in slow
 189 blocks, we modelled the trial-wise cumulative reward with a linear mixed effects model with trial
 190 number, condition (slow/fast), and trial \times condition interaction as predictors. We found a significant
 191 trial \times condition interaction, indicating that the rate of reward accumulation was greater in slow com-
 192 pared to fast blocks ($\beta = 39.07$, 95% CI = [2.44 to 75.70], likelihood ratio test comparing to model
 193 without interaction: $X^2(1) = 4.19$, $p = .041$, Fig 2g).

194 The learning benefit was also evident in an analysis of the average percent of correct choices in slow
 195 vs fast blocks ($M_S = 65.26\% \pm 1.24$, $M_F = 63.54\% \pm 1.35$, $t(49) = 1.98$ $p_{1-sided} = .028$, $d = 0.28$).
 196 A logistic mixed effects model of choice accuracy with fixed effects for condition (slow/fast), trial
 197 number, stimulus value difference to 50, and trial \times value difference showed that including the effect of
 198 condition marginally improved the model predictions ($X^2(1) = 3.33$, $p = .068$), reflecting that correct
 199 choices were more likely in slow blocks, albeit marginally ($\beta = 0.08$, 95% CI = [0.00 to 0.16]). In sum,
 200 participants made more correct choices in slow relative to fast blocks and hence accumulated more
 201 rewards at a faster pace. This lends support to the idea that participants benefited when the relevant
 202 feature was changing slowly.

203 Given that the slowness prior proposes that slow-changing features will be more likely to be con-
 204 sidered relevant, we hypothesised that the lower reward and accuracy on fast blocks could result from
 205 incorrectly basing choices on the slow feature, even when it was irrelevant. To test this, we used the

206 feature positions for both the relevant and irrelevant feature, trial number, and their interactions to
207 predict participant choices separately for slow and fast blocks, using a logistic mixed effects model.
208 We found that on fast blocks, there was a significant impact of the irrelevant slow feature on choice,
209 while on slow blocks the effect of the irrelevant fast feature was marginal (Type II Wald X^2 tests
210 irrelevant slow feature: $X^2(1) = 7.07$, $p = .008$, irrelevant fast feature: $X^2(1) = 2.75$, $p = .097$).
211 Hence, participants tended to base their choices on the slowly changing feature, even when it was not
212 predictive of reward.

213 **Improved generalisation** We next asked whether a difference between slow and fast blocks was
214 also evident in the test phase. Indeed, participants' accuracy was again greater in slow versus fast
215 blocks ($M_S = 76\% \pm 1.6$, $M_F = 74\% \pm 1.5$, $t(49) = 1.85$, $p_{1-sided} = .035$, $d = 0.26$, Fig 2f). A logistic
216 mixed effects model of choice accuracy with fixed effects for condition (slow/fast) and the absolute value
217 difference between the shown stimuli supported this finding, as evidenced by a significant fixed effect for
218 condition ($\beta = 0.14$, 95% CI = [0.01 to 0.28], model comparison to a model without a condition effect:
219 $X^2(1) = 3.99$, $p = .046$). The same picture emerged when modelling participant left/right choices
220 rather than choice accuracy in a logistic mixed effects model, with the condition, value difference and
221 the condition \times value difference interaction as predictors. In slow blocks the true difference in value
222 between the shown stimuli had a greater influence on choice than in fast blocks ($\beta = 0.13$, 95% CI
223 = [0.04 to 0.21]). Hierarchical model comparison showed that a model including the condition \times value
224 difference interaction explained choices better than a model without ($X^2(2) = 7.93$, $p = .005$). Hence,
225 participants were better able to infer and generalise the feature values in the test phase when the
226 relevant feature had changed slowly during the learning phase.

227 **Control analyses** One possible concern regarding the interpretation of these effects is that the auto-
228 correlation of reward outcomes could facilitate learning for slow but not for fast blocks. Our results
229 speak against this interpretation. First, we tested a control model that ignored the stimulus features
230 and simply learned a value estimate from successive reward outcomes (henceforth: Bandit Model).
231 This model performed badly on the task and could not predict participant choices well (see Fig 4a
232 and h below, and Methods), suggesting that auto-correlation alone could not explain the differences in
233 performance between slow and fast blocks. Second, we tested a control model that used a win-stay-lose-
234 shift strategy (henceforth: WSL Model) [47, 48]. This strategy can be helpful in slow blocks, where
235 consecutive trials are likely to require the same choice, but not in fast blocks, where the correct choice
236 is likely to change often. Indeed, this model performed well in slow blocks and badly in fast blocks
237 (see S4 Fig), but could not explain participant choices well (see Fig 4h below, and Methods). Third,
238 we observed better performance on slow blocks in the test trials, where no feedback was provided and
239 rewards on successive trials were not auto-correlated, and participants could not rely on the preceding
240 trials in this phase to guide choices. As both the Bandit and WSL model ignored feature values, they
241 could not account for generalisation in the test phase.

242 **Pilot experiment** The pilot experiment, in which the observation phase was omitted, yielded con-
243 sistent but overall weaker results. Briefly, the difference in cumulative reward during the learning
244 phase pointed in the same direction, but was marginal ($M_S = 128.11 \pm 14.03$, $M_F = 108.88 \pm 14.97$,
245 $t(49) = 1.57$, $p_{1-sided} = .061$, $d = 0.22$, S1 Fig), and the analysis of reward accumulation rate also
246 only numerically pointed toward faster learning in slow blocks ($\beta = 25.16$, 95% CI = [-6.52 to 56.83],
247 $X^2(1) = 2.37$, $p = .124$, S1 Fig). We did not find evidence for a difference in accuracy between
248 conditions in the learning phase, neither in the group means ($M_S = 60\% \pm 1.11$, $M_F = 59\% \pm 1.13$,
249 $t(49) = 1.14$, $p_{1-sided} = .130$, $d = 0.16$), nor in the mixed effects analysis ($\beta = -0.02$, 95% CI = [-0.44
250 to 0.40], $X^2(1) = 1.26$, $p = .263$). However, we did find that the irrelevant feature interfered with
251 choices more on fast blocks than on slow blocks. Specifically, in fast blocks, the effect of the irrelevant
252 feature increased across trials (Type II Wald X^2 tests irrelevant feature \times trial: $X^2(1) = 4.40$, $p = .036$),
253 while in slow blocks it did not ($X^2(1) = 2.71$, $p = .100$). No evidence for condition differences in the
254 test phase was found (all $p > .05$, S1 Fig). The differences between the pilot and main experiment
255 indicate that the observation phase, which explicitly provided information on the speed of the features,
256 critically strengthened the behavioural effect, although other explanations cannot be ruled out (e.g.
257 the pilot had shorter blocks compared to the main experiment).

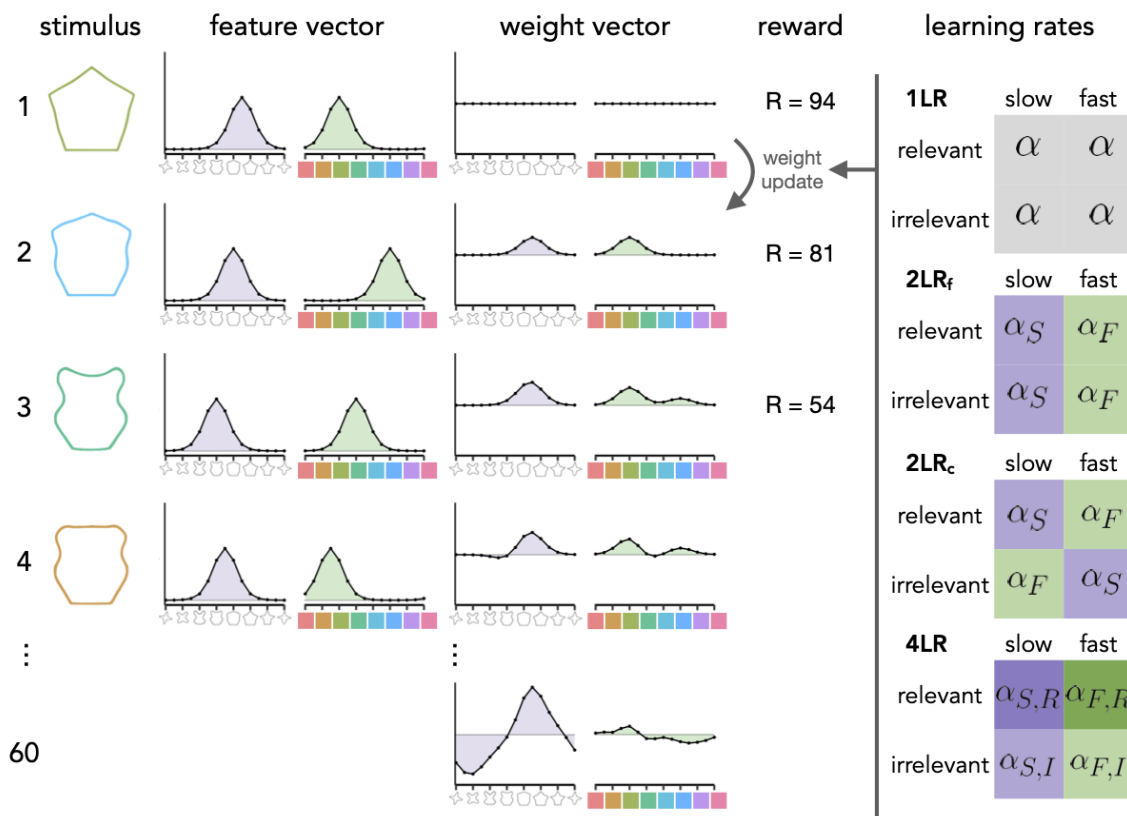


Figure 3: Schematic of the RL models. From left to right: A stimulus is converted to a feature vector, which is a distribution across neighbouring feature values. The feature vector is combined with the weight vector, which stores the value estimates. The resulting value for the stimulus is compared against the reward outcome. This reward prediction error is used to update the weight vector on each trial (shown as rows in the figure). By the end of the block (bottom row), the model learns a mapping between the relevant feature (in this case shape) and reward. The right column shows how the learning rates map onto the stimulus features and experimental condition.

258 Computational Models

259 To examine which mechanisms might underlie the difference in learning between the conditions, we
 260 fitted four reinforcement learning (RL) models to participant choices during the learning phase. Based
 261 on our behavioural findings above, all considered models sought to (a) reflect participants' learning
 262 from outcomes, (b) account for learning about which stimulus feature is relevant and which is not, (c)
 263 incorporate generalisation between stimuli of similar appearance, and (d) reflect participant's tendency
 264 to explore by accepting many stimuli when uncertainty is high. Our major aim was to test whether
 265 the learning process differed depending on whether participants learned about slow or fast-changing
 266 features, i.e. in slow vs fast blocks. To this end, we formulated a set of four models that embodied
 267 alternative hypotheses about how feature speed could affect learning, as described below.

268 All models used linear function approximation and a Kalman filter to account for participants'
 269 generalisation and exploration behaviour, respectively (see Fig 3 and Methods). Briefly, each stimulus
 270 was converted into a 30-dimensional feature vector \mathbf{x}_t that indicated which colour and shape stimulus on
 271 trial t had (one entry for each of the 15 possible shapes and 15 colours). To reflect feature similarity
 272 across the circular stimulus space, a von-Mises distribution was centered around the true stimulus
 273 features, such that activation of node i was determined by its distance from the node assigned to the
 274 true feature t

$$x_{t,i} = \frac{e^{\cos(d_{t,i})\kappa}}{\sum_{i=1}^{360} e^{\cos(d_{t,i})\kappa}} \quad (1)$$

275 where $d_{t,i}$ is the distance between node i and t in radians and κ determines the width (a.k.a. concen-
276 tration) of the von-Mises distribution. We then modelled the expected value V_t of a stimulus as the
277 inner product of the feature vector \mathbf{x}_t and the weight vector \mathbf{w}_t :

$$V_t = \mathbf{x}_t^T \mathbf{w}_t \quad (2)$$

278 and updated \mathbf{w}_t after each accept choice to reflect the outcome R_t of trial t with a learning rate α , as
279 follows:

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha_t \mathbf{x}_t (R_t - V_t) \quad (3)$$

280 To account for exploration behaviour, we modelled participants' uncertainty, U_t , about the value
281 of a stimulus on trial t using a Kalman Filter. Akin to an upper confidence bound mechanism [49],
282 the uncertainty was added to stimulus value in model choices, serving as an exploration bonus (see
283 Methods for details):

$$V_{a,t} = V_t + cU_t \quad (4)$$

284 where c mediates how strongly the exploration bonus is weighted at choice. The uncertainty U_t
285 also determined the learning rate on the current trial, α_t . As the environment was stationary the
286 uncertainty and learning rate reduced across trials. Finally, the model's choice was guided by the
287 probability of the value of accepting, $V_{a,t}$, being larger than a normal random variable centred on 50
288 (the value of rejecting), with standard deviation σ :

$$\begin{aligned} p(\text{accept}) &= P[X \leq V_{a,t}] \\ X &\sim N(50, \sigma^2) \end{aligned} \quad (5)$$

289 While all of the four models reported here used the above-described mechanisms, they differed
290 in whether they could adapt their learning rates to the slowness of the features, the relevance of
291 the features to predict reward, or both (see Fig 3 right column). A baseline model used the same
292 learning rate α for all conditions and features (one learning rate model, short 1LR). Hence, this
293 model was indifferent to slowness and could not account for a difference in performance between the
294 slow and fast blocks. A second model used separate learning rates for the slow vs. fast-changing
295 feature (α_S/α_F), irrespective of whether the feature was relevant in a given block (feature learning
296 rates model, 2LR_f). This model could account for the difference in performance between slow and
297 fast blocks, but since it disregarded the relevance of the features for predicting reward it is an unlikely
298 candidate to explain participant behaviour *a priori*. In a third model (condition learning rates model,
299 2LR_c), separate learning rates were used depending on whether the relevant feature was changing
300 slowly (α_S) or quickly (α_F), but used the same learning rate for both features within the same block,
301 regardless of their relevance. Finally, the fourth model had four separate learning rates for the slow
302 and fast-changing features, when they were relevant and irrelevant (4LR model, learning rates $\alpha_{S,R}$,
303 $\alpha_{F,R}$ vs $\alpha_{S,I}$, $\alpha_{F,I}$, respectively). This model could accommodate both differences in learning due to
304 the slowness of the features and the reward structure of the task, for which reason we expected this
305 model to predict participant choices best.

306 **All models can learn the task** To ensure that all models represent useful accounts of behaviour,
307 we first fitted model parameters to maximise reward obtained by the model. This showed that given
308 optimal parameters all learning models achieved a near-ceiling cumulative reward gain of around 600
309 coins per block, significantly above the cumulative reward expected by chance (all $p < .001$, theoretical
310 maximum of clairvoyant agent: ca. 735 coins). In contrast, above mentioned Random Choice, Random
311 Key, Bandit, and WSLs control models, were all significantly worse at the task (all $p < .001$, Fig 4a).
312 In the test phase, the differences were even starker – only the learning models learned a mapping of
313 stimulus features to reward, so only these models could generalise to unseen feature values (Fig 4b).
314 Hence all learning models were capable of performing our task.

315 We next evaluated which models could in principle reproduce the above-reported condition differ-
316 ence by simulating the models with a higher learning rate for the slow compared to the fast feature
317 (0.6 vs 0.3, respectively; for the 1LR model, we used $\alpha = 0.3$). As expected, all models with 2 or 4
318 learning rates (2LR_f, 2LR_c and 4LR) could, given appropriate parameters, account for a difference
319 between the slow and fast conditions (Fig 4c), while the 1LR model could not reproduce this effect.

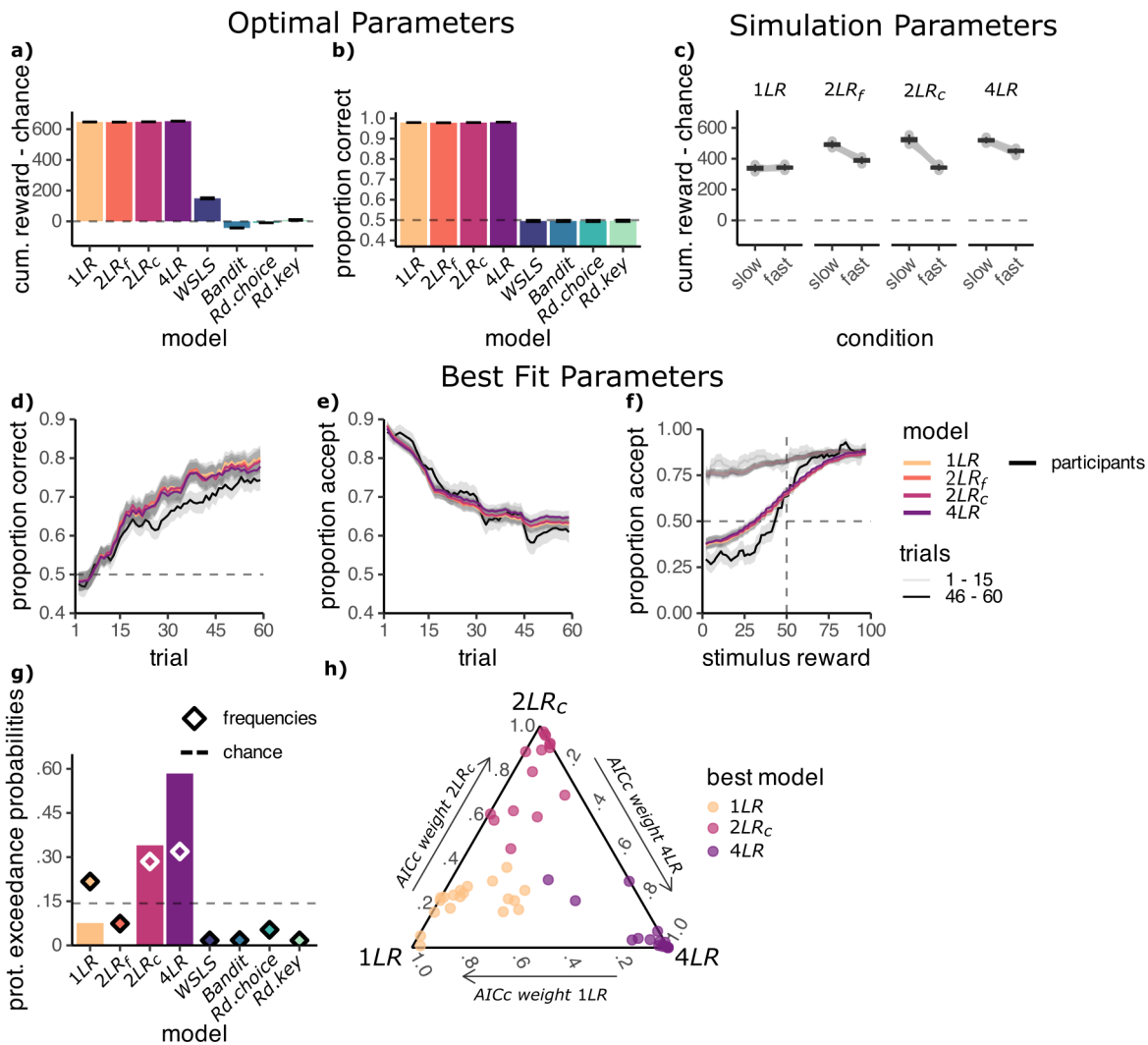


Figure 4: Models including slowness effect explain participant behaviour best. **a)** Mean reward in the learning phase for the models using optimal parameters. Learning models: one learning rate model (1LR), separate learning rates per feature ($2LR_f$), separate learning rates per condition ($2LR_c$) and the four learning rates model (4LR). Control models: win-stay-lose-shift (WSLS), learning model ignoring features (Bandit), random responding with a bias for accept choices (Rd. Choice) or response key (Rd. Key). **b)** Mean accuracy in the test phase for the models using optimal parameters. **c)** Mean reward for slow and fast blocks in the learning phase for the models simulated using hand-picked learning rates, $\alpha/\alpha_F = 0.3$ $\alpha_S = 0.6$. For the 4LR model both relevant learning rates, $\alpha_{S,R}$, $\alpha_{F,R}$, were increased by 0.1. **d)** Proportion correct choices across trials in the learning phase. **e)** Proportion of accept choices across trials in the learning phase. **f)** Proportion of accept choices depending on the true stimulus reward, for the first and last 15 trials of the learning phase. **d-f)** Using best fit model parameters. Lines smoothed with width of 3. Models are shown in coloured lines and participants in black. Control models are not shown. **h)** Protected exceedance probabilities (bars) and estimated frequencies (diamonds) of the models. **i)** Simplex of AICc weights (larger values indicate better fit), calculated considering only the three best-fitting models: 4LR, $2LR_c$ and 1LR. Each point is one participant, coloured by their best fit model.

320 **Learning is affected by slowness** Having established that all models in principle represent plausible
 321 accounts of behaviour, we next asked which model fits participant choices best, using maximum
 322 likelihood fitting and compared models using protected exceedance probabilities. Protected exceedance
 323 probabilities were calculated with the `bmsR` package in R, with model evidence approximated with AICc

weights, relative to the 1LR model [50], for details see Methods. Following maximum likelihood fitting, we first simulated the models with the best-fit parameters (see Table 1). This showed that all models were able to qualitatively match participant learning curves, increasing from 50% to just under 80% correct choices across the 60 trials in a learning block (Fig 4d, see S3 Fig for individual participant fits). Models also captured the decrease in accept choices from around 85% to approximately 63% by the end of learning (Fig 4e), as well the increase in sensitivity to expected reward in both the learning and test phase (Fig 4f and g).

Notably, comparing protected exceedance probabilities [51] and corrected AIC (AICc) scores [52] indicated that the model with four different learning rates (4LR model) fitted behaviour best (XP = .584, AICc = 471.2, see Fig 4h), followed by the model with separate learning rates per condition (2LR_c model, XP = .340, AICc = 471.3) and the 1LR and 2LR_f models (1LR: XP = .076, AICc = 473.5; 2LR_f: XP < .001, AICc = 473.0). The 4LR model was estimated as the most frequent model out of those tested (32%), closely followed by the 2LR_c model (29%, Fig 4h). Together these two models best explained the behaviour of most participants (N=28), however some participants were best fit by the 1LR model (N=15, estimated frequency 22%).

To ask how clear the evidence in favor of the winning model was within each participant, we inspected the distribution of AICc weights for the three best-performing models on a simplex (4LR, 2LR_c and 1LR, Fig 4i). The AICc distribution indicated that participants best fit by the 4LR model were unambiguously best fit by this model, i.e., participants best fit by this model had relatively low weights for the other models. A similar picture emerged for the 2LR_c model. In the case of the one learning rate model (1LR) the difference in fit between the best and alternative models was less pronounced. In sum, the evidence that the best-performing models, 4LR and 2LR_c, adapted their learning rates to the feature speed suggests that participants' learning was affected by feature slowness.

	c	σ	κ	$\alpha/\alpha_S/\alpha_{S,R}$	$\alpha_F/\alpha_{F,R}$	$\alpha_{S,I}$	$\alpha_{F,I}$
1LR	6.08 ± 2.84	41.96 ± 20.37	6.70 ± 8.29	.59 ± .33			
2LR _f	6.57 ± 3.08	44.52 ± 7.43	5.88 ± 7.43	.69 ± .34	.55 ± .34		
2LR _c	6.19 ± 2.65	43.71 ± 8.76	6.82 ± 8.76	.61 ± .33	.57 ± .32		
4LR	6.33 ± 2.96	47.48 ± 8.01	6.80 ± 8.01	.78 ± .33	.70 ± .37	.39 ± .36	.40 ± .32

Table 1: Mean and standard deviation of the best estimates for the exploration parameter (c), decision noise (σ), von Mises concentration (κ), and learning rates on the first trial (α) for the slow (S) or fast (F), and relevant (R) or irrelevant (I) feature, obtained through maximum likelihood fitting.

The 4LR model captures participant behaviour Given that the 4LR model emerged as the winning model, we asked how this model related to the behavioural differences between slow and fast blocks. We compared 4LR model fits to the 1LR model to examine the improvement in fit conferred by the adaptation of learning rates to feature speed, while accounting for the remaining learning mechanisms and ability to solve the task, which were the same across all models (see Fig 4a). Simulating 4LR model choices using the best-fit parameters showed a similar condition difference in accumulated reward as seen in participants (Fig 5a). We found that larger differences in participants' cumulative reward in slow compared to fast blocks in the learning phase were related to a better fit of the 4LR relative to the 1LR model ($r = .28$, $p = .045$, Fig 5b top). We also found that stronger behavioural effects in the test phase were related to a better relative fit of the 4LR model ($r = .30$, $p = .032$, Fig 5b bottom). No such relationships were found for the 2LR_c model ($p > .05$, all p values uncorrected).

We also found that the fitted learning rates related to participant behaviour. Note that due to the Kalman filter aspect of our model, the learning rates decreased across trials (see S5 Fig). Therefore, we examined the mean learning rate across all trials in a block, instead of using the fit value, which was the learning rate on the first trial. When it was relevant, the slow feature benefited from higher mean learning rates than the fast feature ($M_S = .68 \pm .36$, $M_F = .57 \pm .37$, $t(49) = 2.09$, $p = 0.042$, $d = 0.30$). For the irrelevant learning rates, we found no such difference ($M_S = .28 \pm .29$, $M_F = .27 \pm .23$, $t(49) = 0.16$, $p = 0.875$, $d = 0.02$, Fig 5c, all p values uncorrected). Larger mean learning rates for the relevant slow feature were correlated with more reward being accrued on slow than on fast blocks in the learning phase ($r = .41$, $p = .012$ Fig 5d). No other learning rate showed a significant relationship to the behavioural effect (all $p > .05$). These results indicate that the effect of

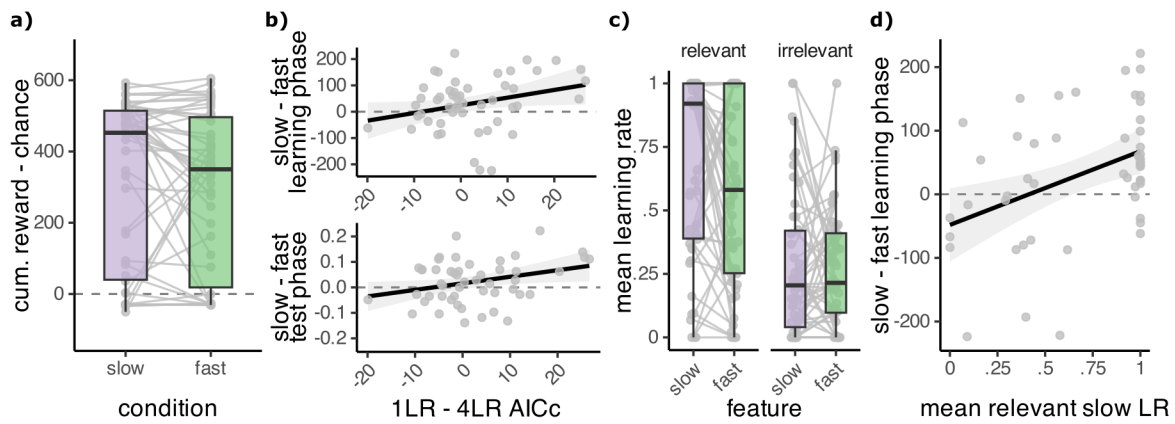


Figure 5: The four learning rates model captures participant behaviour. **a)** Simulating the 4LR model with the best-fit learning rates leads to higher collected reward in slow compared to fast blocks. **b)** A better fit of the 4LR model (x) is related to greater collected reward in slow than in fast blocks in the learning phase (top) and (bottom) greater accuracy in slow than in fast blocks in the test phase (bottom). **c)** Distribution of learning rates for the 4LR model, obtained from maximum likelihood fitting. Mean across all trials in a block. **d)** Higher mean learning rates for the slow feature (x) are correlated with greater collected reward in slow than in fast blocks in the learning phase (y). Points are individual participants. Grey ribbons show standard error of the mean.

369 feature speed on learning was mainly modulated by improved learning from the slow feature. Hence,
370 individual differences in model parameters and fit captured differences in how strongly the slowness
371 prior influenced participants' choices.

372 Discussion

373 Causal processes tend to evolve on a slower timescale than noise [31]. To investigate whether humans
374 employ a slowness prior to identify potentially relevant features during reinforcement learning, we
375 tested participants in a decision-making task with stimuli composed of one reward-predictive and one
376 reward-irrelevant feature. Participants learned the value of stimuli faster when the reward-predictive
377 feature changed slowly and the irrelevant feature changed quickly, compared to when the opposite was
378 the case. Participants were also more distracted by the irrelevant feature when it changed slowly than
379 when it changed quickly. By comparing models with different structures for the learning rates, we
380 showed that participants adjusted their learning to the speed of the features. Specifically the learning
381 rate for the slow feature when it was relevant mediated the behavioural effect, suggesting that the
382 observed behavioural differences between conditions were being driven by increased learning from the
383 slow feature. Our study extends research on the slowness prior to humans and suggests that it aids
384 learning task states, in a reinforcement learning domain.

385 Our work relates to a broader discussion of how the human brain solves representation learning
386 problems [27, 30]. Previous work has shown how representation learning can be implemented in
387 parallel to reinforcement learning by using feedback signals to guide selective attention [53, 54], or
388 through replay mechanisms during offline periods [55, 56]. Although these approaches represent flexible
389 mechanisms that allow on-the-fly adaptation to the current environment, it is unlikely to be feasible in
390 environments with hundreds of possible signals to attend to [3, 6, 29]. Our results suggest that for this
391 reason representation learning mechanisms during RL are supplemented with inductive biases. Our
392 findings are in line with previous research showing that priors have a pervasive influence on behaviour,
393 shaping perception [15, 35], remaining stable in the face of exposure to contradictory training [57], and
394 hindering learning of structures which do not align with them [58, 59]. More indirectly, our work raises
395 the question about the origins of such priors, and whether they are learned themselves. One possibility
396 in this regard is that meta-learning, or learning to learn, is the core mechanism that humans use in
397 order to extract regularities of their environment and develop priors that aid perception and learning
398 [60].

399 While our results align with several theoretical studies on the slowness prior [34, 37, 41], it is
400 important to consider other ways in which slowness can benefit learning. For instance, the temporal
401 auto-correlation of features and rewards inherent to a slowly changing environment could enable the
402 use of heuristic strategies, such as a win-stay-lose-shift rule [47, 48]. We addressed this concern
403 through model comparison and found that these strategies were unable to explain the behaviour of
404 participants. Another possibility is that presenting stimuli in an ordered fashion yields benefits, as
405 suggested in function learning studies [61]. In our task, slow blocks were more likely to be ordered
406 than fast blocks, but due to the periodic nature of our feature-reward mapping, ordering might not be
407 immediately apparent in either condition. Still, future research should aim to disentangle the effects
408 of ordering and slowness on learning. Importantly, assuming relevant processes change slowly only is
409 a useful assumption given the physical laws that govern our world, i.e., Newton's first law of motion,
410 inertia [37]. Under these conditions, slow acceleration and changes in acceleration are likely to also
411 provide useful priors, as has been shown in motion perception studies in humans [36]. Human learning
412 likely incorporates a host of priors, reflecting other properties determined by our (intuitive) physical
413 understanding of the world [16].

414 Our findings also relate to previous work on curriculum learning, which has shown that humans
415 benefit from blocked, rather than interleaved, training on a context-dependent categorisation task [62].
416 In the blocked curriculum the relevant features for categorisation were the same across trials, whereas
417 in the interleaved curriculum the relevant features could switch from trial to trial, even though the
418 stimuli characteristics changed in both curricula. This raises the possibility that slowness, not only in
419 feature dynamics but also in task rules, may aid learning. However, it is worth noting that interleaved
420 training might promote the formation of more generalisable representations [63], suggesting that the
421 optimal learning curriculum may differ depending on the task at hand. In sum, multiple lines of
422 research point toward a beneficial effect of slowness on learning. Here, we propose that part of this
423 effect is due to the existence of a slowness prior.

424 Our task and models make some simplifying assumptions. In our task, participants need to reduce
425 a two-dimensional stimulus to a one-dimensional representation. Despite its simplicity, the task itself
426 posed a considerable challenge to participants, as indicated by their end-of-learning performance,
427 which still left room for improvement. Consequently, the task contained the necessary elements to

428 test our hypothesis and provides a controlled test bed for looking at dimensionality reduction. Our
429 winning model, the four learning rate model, assigned learning rates to the features based on their
430 speed and relevance from the first learning trial of the block. While it is reasonable to assume that
431 participants in the main experiment knew the speed of the features based on the preceding observation
432 phase, they could not yet have known which feature was relevant. However, it is important to note
433 that due to our models being Kalman Filters, we merely fit the learning rates on the first trial, and
434 the development of learning rates throughout the blocks was determined by the experience with the
435 environment. Additionally, participants' accuracy increased within the first learning trials in a block,
436 leading us to believe that they quickly developed a sense for the relevance of the features. We chose
437 this approach for its computational simplicity, but it remains a potential avenue for future research.
438 It is for instance possible that the dynamics of learning rates are influenced by a number of additional
439 factors, such as volatility or the size of prediction errors [64–66]. In addition, participants might learn
440 a belief about which feature is relevant to determine learning rates [67].

441 Overall, the results of our experiments suggest that participants were able to infer, learn and
442 generalise the values of stimuli better when the relevant feature changed slowly. By providing empirical
443 evidence for the role of a slowness prior in human learning and connecting to a large number of machine
444 learning findings [31, 37, 39], our study sheds light onto how humans might rapidly learn representations
445 in complex environments.

446 References

- 447 1. Schuck NW, Gaschler R, Wenke D, et al. Medial Prefrontal Cortex Predicts Internally Driven
448 Strategy Shifts. *Neuron* 2015;86:331–40.
- 449 2. Löwe AT, Touzo L, Muhle-Karbe PS, Saxe AM, Summerfield C, and Schuck NW. Regularised
450 neural networks mimic human insight. arXiv:2302.11351 [cs, q-bio]. 2023. URL: <http://arxiv.org/abs/2302.11351> (visited on 10/31/2023).
451
- 452 3. Kemp C and Tenenbaum JB. Structured statistical models of inductive reasoning. *Psychological*
453 *Review* 2009;116:20–58.
- 454 4. Gershman SJ and Niv Y. Novelty and Inductive Generalization in Human Reinforcement Learn-
455 ing. *Topics in Cognitive Science* 2015;7:391–415.
- 456 5. Griffiths TL, Chater N, Kemp C, Perfors A, and Tenenbaum JB. Probabilistic models of cognition:
457 exploring representations and inductive biases. *Trends in Cognitive Sciences* 2010;14:357–64.
- 458 6. Gershman SJ, Cohen JD, and Niv Y. Learning to Selectively Attend. *Proceedings of the Annual*
459 *Meeting of the Cognitive Science Society* 2010.
- 460 7. Battaglia PW, Hamrick JB, Bapst V, et al. Relational inductive biases, deep learning, and graph
461 networks. arXiv:1806.01261 [cs, stat]. 2018. URL: <http://arxiv.org/abs/1806.01261> (visited
462 on 10/16/2023).
- 463 8. Bengio Y, Courville A, and Vincent P. Representation Learning: A Review and New Perspectives.
464 *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2013;35:1798–828.
- 465 9. Wilke A and Mata R. Cognitive Bias. In: *Encyclopedia of Human Behavior*. Elsevier, 2012:531–5.
466 DOI: [10.1016/B978-0-12-375000-6.00094-X](https://doi.org/10.1016/B978-0-12-375000-6.00094-X). URL: [https://linkinghub.elsevier.com/](https://linkinghub.elsevier.com/retrieve/pii/B978012375000600094X)
467 [retrieve/pii/B978012375000600094X](https://linkinghub.elsevier.com/retrieve/pii/B978012375000600094X) (visited on 12/12/2023).
- 468 10. Hermsdorff GB, Pereira T, and Niv Y. Quantifying Humans’ Priors Over Graphical Representa-
469 tions of Tasks. In: *Unifying Themes in Complex Systems IX*. Ed. by Morales AJ, Gershenson C,
470 Braha D, Minai AA, and Bar-Yam Y. Series Title: Springer Proceedings in Complexity. Cham:
471 Springer International Publishing, 2018:281–90. DOI: [10.1007/978-3-319-96661-8_30](https://doi.org/10.1007/978-3-319-96661-8_30). URL:
472 http://link.springer.com/10.1007/978-3-319-96661-8_30 (visited on 10/06/2023).
- 473 11. Schulz E, Tenenbaum JB, Duvenaud D, Speekenbrink M, and Gershman SJ. Compositional in-
474 ductive biases in function learning. *Cognitive Psychology* 2017;99:44–79.
- 475 12. Gershman SJ and Niv Y. Perceptual estimation obeys Occam’s razor. *Frontiers in Psychology*
476 2013;4.
- 477 13. Quiroga F, Schulz E, Speekenbrink M, and Harvey N. Structured priors in human forecasting.
478 Pages: 285668 Section: New Results. 2018. DOI: [10.1101/285668](https://doi.org/10.1101/285668). URL: [https://www.biorxiv.org/](https://www.biorxiv.org/content/10.1101/285668v1)
479 [content/10.1101/285668v1](https://www.biorxiv.org/content/10.1101/285668v1) (visited on 10/31/2023).
- 480 14. Gigerenzer G and Gaissmaier W. Heuristic Decision Making. *Annual Review of Psychology*
481 2011;62:451–82.
- 482 15. Coren S and Girgus J. Seeing is Deceiving: The Psychology of Visual Illusions. Google-Books-ID:
483 uyX5DwAAQBAJ. Routledge, 2020.
- 484 16. Lake BM, Ullman TD, Tenenbaum JB, and Gershman SJ. Building machines that learn and think
485 like people. *Behavioral and Brain Sciences* 2017;40:e253.
- 486 17. Dubey R, Agrawal P, Pathak D, Griffiths TL, and Efros AA. Investigating Human Priors for
487 Playing Video Games. arXiv:1802.10217 [cs]. 2018. URL: <http://arxiv.org/abs/1802.10217>
488 (visited on 10/06/2023).
- 489 18. Saanum T, Éltető N, Dayan P, Binz M, and Schulz E. Reinforcement Learning with Simple
490 Sequence Priors. arXiv:2305.17109 [cs]. 2023. URL: <http://arxiv.org/abs/2305.17109> (visited
491 on 10/31/2023).
- 492 19. Sutton RS and Barto AG. Reinforcement learning: an introduction. Adaptive computation and
493 machine learning. Cambridge, Mass: MIT Press, 1998.
- 494 20. Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning.
495 *Nature* 2015;518:529–33.

- 496 21. Niv Y. Reinforcement learning in the brain. *Journal of Mathematical Psychology* 2009;53:139–54.
- 497 22. Rescorla RA and Wagner AR. A theory of Pavlovian conditioning : Variations in the effective-
498 ness of reinforcement and non-reinforcement. *Classical conditioning, Current research and theory*
499 1972;2. Publisher: Appleton-Century-Crofts:64–9.
- 500 23. Schultz W, Dayan P, and Montague PR. A Neural Substrate of Prediction and Reward. *Science*
501 1997;275:1593–9.
- 502 24. Kaplan R, Schuck NW, and Doeller CF. The Role of Mental Maps in Decision-Making. *Trends*
503 *in Neurosciences* 2017;40:256–9.
- 504 25. Schuck NW, Cai MB, Wilson RC, and Niv Y. Human Orbitofrontal Cortex Represents a Cognitive
505 Map of State Space. *Neuron* 2016;91:1402–12.
- 506 26. Niv Y. Learning task-state representations. *Nature Neuroscience* 2019;22:1544–53.
- 507 27. Radulescu A, Shin YS, and Niv Y. Human Representation Learning. *Annual Review of Neuro-*
508 *science* 2021;44:253–73.
- 509 28. Lesort T, Díaz-Rodríguez N, Goudou JF, and Filliat D. State representation learning for control:
510 An overview. *Neural Networks* 2018;108:379–92.
- 511 29. Bellman R and Kalaba R. On adaptive control processes. *IRE Transactions on Automatic Control*
512 1959;4. Conference Name: IRE Transactions on Automatic Control:1–9.
- 513 30. Schuck NW, Wilson R, and Niv Y. Chapter 12 - A State Representation for Reinforcement
514 Learning and Decision-Making in the Orbitofrontal Cortex. In: *Goal-Directed Decision Making*.
515 Ed. by Morris R, Bornstein A, and Shenhav A. Academic Press, 2018:259–78. DOI: [10.1016/B978-](https://doi.org/10.1016/B978-0-12-812098-9.00012-7)
516 [0-12-812098-9.00012-7](https://doi.org/10.1016/B978-0-12-812098-9.00012-7). URL: [https://www.sciencedirect.com/science/article/pii/](https://www.sciencedirect.com/science/article/pii/B9780128120989000127)
517 [B9780128120989000127](https://www.sciencedirect.com/science/article/pii/B9780128120989000127) (visited on 12/06/2023).
- 518 31. Wiskott L and Sejnowski TJ. Slow Feature Analysis: Unsupervised Learning of Invariances. *Neural*
519 *Computation* 2002;14:715–70.
- 520 32. Körding KP, Kayser C, Einhäuser W, and König P. How Are Complex Cell Properties Adapted
521 to the Statistics of Natural Stimuli? *Journal of Neurophysiology* 2004;91:206–12.
- 522 33. Roth S and Black MJ. On the Spatial Statistics of Optical Flow. *International Journal of Com-*
523 *puter Vision* 2007;74:33–50.
- 524 34. Weiss Y, Simoncelli EP, and Adelson EH. Motion illusions as optimal percepts. *Nature Neuro-*
525 *science* 2002;5:598–604.
- 526 35. Stocker AA and Simoncelli EP. Noise characteristics and prior expectations in human visual
527 speed perception. *Nature Neuroscience* 2006;9:578–85.
- 528 36. Lu H, Lin T, Lee A, Vese L, and Yuille AL. Functional form of motion priors in human motion
529 perception. *Advances in neural information processing systems* 2010;23.
- 530 37. Jonschkowski R and Brock O. Learning state representations with robotic priors. *Autonomous*
531 *Robots* 2015;39:407–28.
- 532 38. Anand A, Racah E, Ozair S, Bengio Y, Cote MA, and Hjelm RD. Unsupervised State Represen-
533 tation Learning in Atari. *Advances in neural information processing systems* 2019;32.
- 534 39. Becker S and Hinton GE. Self-organizing neural network that discovers surfaces in random-dot
535 stereograms. *Nature* 1992;355:161–3.
- 536 40. Song P and Zhao C. Slow Down to Go Better: A Survey on Slow Feature Analysis. *IEEE Trans-*
537 *actions on Neural Networks and Learning Systems* 2022;1–21.
- 538 41. Legenstein R, Wilbert N, and Wiskott L. Reinforcement Learning on Slow Features of High-
539 Dimensional Input Streams. *PLoS Computational Biology* 2010;6. Ed. by Morrison A:e1000894.
- 540 42. Berkes P and Wiskott L. Slow feature analysis yields a rich repertoire of complex cell properties.
541 *Journal of Vision* 2005;5:9–9.
- 542 43. Rolls ET. Learning Invariant Object and Spatial View Representations in the Brain Using Slow
543 Unsupervised Learning. *Frontiers in Computational Neuroscience* 2021;15:686239.
- 544 44. Földiák P. Learning Invariance from Transformation Sequences. *Neural Computation* 1991;3:194–
545 200.

- 546 45. Franzius M, Sprekeler H, and Wiskott L. Slowness and Sparseness Lead to Place, Head-Direction,
547 and Spatial-View Cells. *PLoS Computational Biology* 2007;3. Ed. by Friston KJ:e166.
- 548 46. Lipshutz D, Windolf C, Golkar S, and Chklovskii D. A Biologically Plausible Neural Network for
549 Slow Feature Analysis. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran
550 Associates, Inc., 2020:14986–96. URL: [https://proceedings.neurips.cc/paper/2020/hash/
551 ab73f542b6d60c4de151800b8abc0a6c-Abstract.html](https://proceedings.neurips.cc/paper/2020/hash/ab73f542b6d60c4de151800b8abc0a6c-Abstract.html) (visited on 10/31/2023).
- 552 47. Posch M. Win–Stay, Lose–Shift Strategies for Repeated Games—Memory Length, Aspiration
553 Levels and Noise. *Journal of Theoretical Biology* 1999;198:183–95.
- 554 48. Thorndike E. *Animal Intelligence: Experimental Studies*. New York: Routledge, 2017. DOI: [10.
555 4324/9781351321044](https://doi.org/10.4324/9781351321044).
- 556 49. Auer P. Using Confidence Bounds for Exploitation-Exploration Trade-offs. *Journal of Machine
557 Learning Research* 2002;3:397–422.
- 558 50. Wagenmakers EJ and Farrell S. AIC model selection using Akaike weights. *Psychonomic Bulletin
559 & Review* 2004;11:192–6.
- 560 51. Stephan KE, Penny WD, Daunizeau J, Moran RJ, and Friston KJ. Bayesian model selection for
561 group studies. *NeuroImage* 2009;46:1004–17.
- 562 52. Sugiura N. Further analysis of the data by Akaike’s information criterion and the finite corrections:
563 Further analysis of the data by akaike’ s. *Communications in Statistics - Theory and Methods*
564 1978;7:13–26.
- 565 53. Niv Y, Daniel R, Geana A, et al. Reinforcement Learning in Multidimensional Environments
566 Relies on Attention Mechanisms. *The Journal of Neuroscience* 2015;35:8145–57.
- 567 54. Jones M and Canas F. Integrating Reinforcement Learning with Models of Representation Learn-
568 ing. *Proceedings of the Annual Meeting of the Cognitive Science Society* 2010;32.
- 569 55. Russek EM, Momennejad I, Botvinick MM, Gershman SJ, and Daw ND. Predictive representa-
570 tions can link model-based reinforcement learning to model-free mechanisms. *PLOS Computa-
571 tional Biology* 2017;13. Ed. by Daunizeau J:e1005768.
- 572 56. Wittkuhn L, Chien S, Hall-McMaster S, and Schuck NW. Replay in minds and machines. *Neu-
573 roscience & Biobehavioral Reviews* 2021;129:367–88.
- 574 57. Roark CL and Holt LL. Long-term priors constrain category learning in the context of short-term
575 statistical regularities. *Psychonomic Bulletin & Review* 2022;29:1925–37.
- 576 58. Best CT, McRoberts GW, and Goodell E. Discrimination of non-native consonant contrasts
577 varying in perceptual assimilation to the listener’s native phonological system. *The Journal of
578 the Acoustical Society of America* 2001;109:775–94.
- 579 59. Kuhl PK, Conboy BT, Coffey-Corina S, Padden D, Rivera-Gaxiola M, and Nelson T. Phonetic
580 learning as a pathway to language: new data and native language magnet theory expanded (NLM-
581 e). *Philosophical Transactions of the Royal Society B: Biological Sciences* 2008;363:979–1000.
- 582 60. Braun DA, Mehring C, and Wolpert DM. Structure learning in action. *Behavioural Brain Research*
583 2010;206:157–65.
- 584 61. Byun E. Interaction between prior knowledge and type of nonlinear relationship on function
585 learning. PhD thesis. Purdue University, 1995.
- 586 62. Flesch T, Balaguer J, Dekker R, Nili H, and Summerfield C. Comparing continual task learning
587 in minds and machines. *Proceedings of the National Academy of Sciences* 2018;115.
- 588 63. Zhou Z, Singh D, Tandoc MC, and Schapiro AC. Building Integrated Representations Through
589 Interleaved Learning. *Journal of Experimental Psychology* 2023;152:2666–84.
- 590 64. Nassar MR, Wilson RC, Heasly B, and Gold JI. An Approximately Bayesian Delta-Rule Model
591 Explains the Dynamics of Belief Updating in a Changing Environment. *The Journal of Neuro-
592 science* 2010;30:12366–78.
- 593 65. Yu AJ and Dayan P. Uncertainty, Neuromodulation, and Attention. *Neuron* 2005;46:681–92.
- 594 66. Koch C, Zika O, Bruckner R, and Schuck NW. Influence of surprise on reinforcement learning
595 in younger and older adults. preprint. *PsyArXiv*, 2022. DOI: [10.31234/osf.io/unx5y](https://doi.org/10.31234/osf.io/unx5y). URL:
596 <https://osf.io/unx5y> (visited on 01/04/2024).

- 597 67. Palminteri S, Khamassi M, Joffily M, and Coricelli G. Contextual modulation of value signals in
598 reward and punishment learning. *Nature Communications* 2015;6:8096.
- 599 68. Li AY, Liang JC, Lee ACH, and Barense MD. The validated circular shape space: Quantifying
600 the visual similarity of shape. *Journal of Experimental Psychology: General* 2020;149:949–66.
- 601 69. Leeuw JRd, Gilbert RA, and Luchterhandt B. jsPsych: Enabling an Open-Source Collaborative
602 Ecosystem of Behavioral Experiments. *Journal of Open Source Software* 2023;8:5351.
- 603 70. R Core Team T. R: A language and environment for statistical computing. R Foundation for
604 Statistical Computing 2017.
- 605 71. Team R. RStudio: Integrated Development Environment for R. PBC 2020.
- 606 72. Bates D, Mächler M, Bolker B, and Walker S. Fitting Linear Mixed-Effects Models Using **lme4**.
607 *Journal of Statistical Software* 2015;67.
- 608 73. Barr DJ, Levy R, Scheepers C, and Tily HJ. Random effects structure for confirmatory hypothesis
609 testing: Keep it maximal. *Journal of Memory and Language* 2013;68:255–78.
- 610 74. Burnham KP and Anderson DR. Model selection and multimodel inference: a practical information-
611 theoretic approach. 2nd ed. OCLC: ocm48557578. New York: Springer, 2002.
- 612 75. Dixon P. Models of accuracy in repeated-measures designs. *Journal of Memory and Language*
613 2008;59:447–56.
- 614 76. Quené H and Van Den Bergh H. On multi-level modeling of data from repeated measures designs:
615 a tutorial. *Speech Communication* 2004;43:103–21.

616 Methods

617 Participants

618 For each of the two experiments 50 participants (*pilot experiment*: female = 19, age = 18-38 years, M
619 = 24.4 years, SD = 5.3 years, *main experiment*: female = 15, age: 18-39 years, M = 24.6 years, SD
620 = 5.4 years) were recruited through Prolific (www.prolific.co) and completed the experiment online.
621 None reported being colour blind and none were currently receiving treatment or taking medication
622 for mental illness. Participants were compensated £3.75, plus a performance-dependent bonus of up
623 to £1.50. The sample size was based on a power analysis set to achieve a power of .8, using the results
624 from a preparatory study ($d = .36$, paired one-tailed t-test with alpha of .05). The study was approved
625 by the Ethics Committee of the Max Planck Institute for Human Development.

626 Materials

627 Stimuli were coloured shapes, with shapes originating from the Validated Circular Shape space [68]
628 and colours defined as a slice in CIELAB colour space, with luminance 70, chroma 51 and origin [0,0].
629 Shapes and colours were parameterized on a circular space, so each position (0-359°) corresponded to
630 one colour or one shape (Fig 1a), and colour/shape similarity varied continuously but had no hard
631 boundaries. The feature spaces were perceptually uniform, so that the angular distance between feature
632 values corresponded to the perceived difference between them. Small angular distances correspond to
633 similar shapes (or colours), whereas large angular distances correspond to distinct shapes (or colours).

634 In the learning phase of each block, a subset of 15 positions was shown, spaced uniformly around
635 the circle in steps of 24°. Each block used a distinct set of positions, offset from the positions used in
636 other blocks in multiples of 3° and assigned to blocks in a random order. In the test phase, stimuli were
637 constructed from 15 feature positions offset by 12° from the positions used in the preceding learning
638 phase. This offset ensured that shapes and colours seen at test were maximally different from those
639 seen during learning, providing a strong and semi-independent test of participants' knowledge about
640 the feature-reward mapping.

641 The task was programmed as an online experiment using the jsPsych library version 6.1.0 [69].

642 Design

643 Participants completed a task that required them to learn the rewards associated with a set of visual
644 stimuli characterized by two features (colour and shape) (Fig 1). Unbeknownst to participants, stimulus
645 rewards were related to only one of the two features in each block. We refer to the feature that predicted
646 reward as the relevant feature and the feature that did not predict reward as the irrelevant feature
647 (Fig 1b). For each block one position in the relevant feature space was chosen as the maximum reward
648 position. Maximum reward positions were at 10°, 100°, 190°, or 280° in the feature space. Each of
649 these reward positions was used once for colour-relevant and once for shape-relevant blocks, in random
650 order. The closer the relevant stimulus feature was to the maximum reward position, the higher the
651 stimulus reward. The stimulus reward was calculated as the absolute distance between the relevant
652 feature position and the maximum reward position, subtracted from the maximum possible distance of
653 180°. The resulting value was re-scaled from the angular distance range (0-180°) to the reward range
654 (0-100 coins).

655 We manipulated feature speed, by controlling the trial-to-trial variability of the two features.
656 Within each block, one feature had low variability across trials (e.g. participants see relatively similar
657 shapes from trial to trial), while the other feature had high variability (e.g. participants see relatively
658 distinct colours from trial to trial). We refer to these as the slow and fast feature, respectively (Fig
659 1a). The slow feature was sampled using a Gaussian random walk centred on 0°, with a standard
660 deviation of 30°. The fast feature was sampled randomly, while preventing the smallest step-size (24°)
661 from occurring. Within each block, the 15 feature positions (see Materials) repeated three times in the
662 pilot experiment and four times in the main experiment, with each position being shown once before
663 repeating. In this way, we ensured comparable exposure to the slow and fast feature spaces, despite
664 their differing variability.

665 We counterbalanced the relevant feature dimension (shape relevant/colour irrelevant or vice versa)
666 and the feature speed (shape slow/colour fast or vice versa). Each combination of relevant feature

667 dimension and relevant feature speed was repeated twice, resulting in eight task blocks. In half of the
668 eight task blocks, the slow feature was relevant (slow blocks), in the other half the fast feature was
669 relevant (fast blocks, Fig 1c). The block order was pseudo-randomised, so that each combination was
670 experienced once before repeating.

671 Procedure

672 Each task block consisted of three phases, observation, learning, and test (Fig 1d-f).

673 The observation phase served to demonstrate the variability of the features to participants. Thirty
674 individual stimuli were shown in rapid succession (500ms each) and without intervening screens. The
675 speed of the features in the observation phase matched that in the subsequent learning phase. Both
676 phases used the same set of 15 feature positions, however, sequences for observation and learning were
677 sampled independently and started at randomly selected positions in feature space. In the learning
678 phase, participants played an accept-reject task and were asked to maximise coins earned by collecting
679 valuable gems. Each trial began with a gem (a coloured shape) being displayed centrally on the screen.
680 Using the ‘F’ or ‘J’ key, participants could either accept the stimulus, and receive the reward associated
681 with it (between 0 to 100 coins), or reject the stimulus and receive an average reward (50 coins). The
682 reject/accept key mapping was counterbalanced across trials. If participants failed to respond after
683 four seconds they received zero coins. Immediately after a key press, the number of coins earned was
684 displayed on the screen for one second, followed by a blank screen for a variable inter-trial interval
685 (0.5 to 1.5s). A correct response was defined as accepting a stimulus with a value above 50 coins or
686 rejecting a stimulus with a value below 50 coins.

687 Following the learning phase, participants completed a two alternative forced choice task to test
688 their understanding of the stimulus values. In this test phase, participants were presented with pairs
689 of stimuli and asked to choose the more valuable stimulus in the pair, based on the preceding learning
690 phase. On each trial, participants could choose the left or right stimulus with the ‘F’ or ‘J’ keys,
691 respectively, with no time limit. After their response a blank screen was shown for a variable inter-trial
692 interval (0.5 to 1.5s). There was no trial-wise feedback during the test phase. A correct response was
693 defined as choosing the stimulus with the higher value. Here, feature speed was no longer manipulated.
694 Instead, the difference in value between the two stimuli in a pair was systematically varied. By
695 controlling the relevant feature positions of the two stimuli, it was possible to probe choices from
696 easier comparisons, where stimuli had more distinct values (the maximum included difference was 54
697 coins), to increasingly difficult comparisons, where the values of the two stimuli were more similar (the
698 minimum difference was 2 coins in the main experiment and 13 coins in the pilot experiment). Overall
699 block accuracy (including both learning and test phase) was reported to participants at the end of the
700 block and used to determine the performance bonus.

701 We ran two versions of the experiment. In the *pilot experiment* the observation phase of the
702 experiment was omitted. Nonetheless, the speed of the features was still manipulated during the
703 learning phase, so slowness information was available, but less evident and presented concurrently
704 with the reward learning task. The *main experiment* included an observation phase prior to the
705 learning phase, as described above, which explicitly demonstrated the speed of the features prior to
706 learning their values. Additionally, there were differences in the length of each task. In the *pilot*
707 *experiment* participants completed 45 learning trials and 15 test trials per block, while in the *main*
708 *experiment* participants completed 30 observation trials, 60 learning trials, and 36 test trials per block.
709 In all other aspects, the experiments were identical.

710 Data Analysis

711 **Mixed Effects Models** We ran mixed effect models in R (R version 4.3.1, RStudio version 2023.09.1
712 + 494), using `lmer` (linear) and `glmer` (logistic) from the `lme4` package (version 1.1-32) [70–72]. To
713 obtain parameter values we ran the Bound Optimisation by Quadratic Approximation (BOBYQA)
714 algorithm for 100.000 evaluations. We initially included all relevant fixed effects and their interactions
715 in the models and subsequently used the `drop1` function in R to test which terms contributed to the
716 fit. All terms that did not significantly improve the fit were removed. We used a maximal random
717 effects structure whenever possible [73]. That is, all variables and interactions initially included as
718 fixed effects were included in the random effects, even if they were later dropped from the fixed effects.
719 Random effects were only simplified if the maximal structure led to fitting issues. All continuous

720 predicting variables were scaled, trial number was normalised to range between zero and one. Trials
721 with no response were excluded from all analyses.

722 We first analysed performance in the learning phase by using a linear mixed effects model to look
723 at the cumulative reward obtained by participants relative to a chance level reward of 50 per trial.
724 The best model was:

$$CR_t = \beta_0 + \beta_1 \text{Condition}_t + \beta_2 t + \beta_3 \text{Condition}_t \times t + (1 + \text{Condition}_t + t + \text{Condition}_t \times t | \text{Subject})$$

725 where CR_t is the cumulative reward relative to chance on trial t , and the predictors are the Condition
726 (slow/fast block), the trial number t , and their interaction.

727 We then examined correct vs. incorrect choices in the learning phase using a logistic mixed effects
728 model. After backwards model comparison the best model was:

$$ACC_t = \beta_0 + \beta_1 \text{Condition}_t + \beta_2 t + \beta_3 |R_t - 50| + \beta_4 t \times |R_t - 50| + (1 + \text{Condition}_t | \text{Subject})$$

729 where ACC_t denotes whether a choice on trial t was correct and $|R_t - 50|$ is the absolute difference
730 between the stimulus reward on trial t and the choice boundary of 50 coins.

To examine the effect of the relevant and irrelevant feature on choice we used a logistic mixed effects model to predict choices based on the stimulus colour and shape positions on each trial. As the features were angles in the shape and colour circles, each feature was included as a $\cos()$ and $\sin()$ predictor in the model. As this analysis was run separately for slow and fast blocks, no model comparison was done.

$$\begin{aligned} C_t = & \beta_0 + \beta_1 t + \beta_2 \cos(\theta_R) + \beta_3 \sin(\theta_R) + \beta_4 \cos(\theta_I) + \beta_5 \sin(\theta_I) \\ & + \beta_6 t \times \cos(\theta_R) + \beta_7 t \times \sin(\theta_R) + \beta_8 t \times \cos(\theta_I) + \beta_9 t \times \sin(\theta_I) \\ & + (1 + \cos(\theta_R) + \cos(\theta_I) | \text{Subject}) \end{aligned}$$

731 where θ_R is the position of the relevant feature and θ_I is the position of the irrelevant feature.

732 To look at performance in the test phase, we examined correct versus incorrect choices using a
733 logistic mixed effects model and found the following model predicted accuracy best:

$$ACC_t = \beta_0 + \beta_1 \text{Condition}_t + \beta_2 |R_{\text{diff},t}| + (1 + \text{Condition}_t | \text{Subject}) \quad (6)$$

734 where $|R_{\text{diff},t}|$ is the absolute difference in value between the left and right stimulus on trial t .

735 The probability of choosing the right stimulus on a test trial was best explained by the following
736 logistic mixed effects model:

$$C_t = \beta_0 + \beta_1 \text{Condition}_t + \beta_2 R_{\text{diff},t} + \beta_3 \text{Condition}_t \times R_{\text{diff},t} + (1 + \text{Condition}_t | \text{Subject}) \quad (7)$$

737 where $R_{\text{diff},t}$ is the difference in value between the left and right stimulus on trial t .

738 Computational Models

739 To analyse trial-by-trial learning, we fit eight computational models to the choices of participants in
740 the learning task. Four learning models embodied alternative hypotheses about how the prior could
741 affect learning and differed in their ability to adapt their learning rates to the slowness of the features.
742 The other four models served as control models and tested for competing hypotheses or tested whether
743 participants engaged with the task.

744 **Learning models** The reinforcement learning (RL) models used the outcome of each trial to update
745 their estimate of the value of the features and predict the next choices of participants. To account for
746 the fact that continuous feature dimensions in the task allowed participants to generalise their learning
747 within each feature (i.e., learning about the value of red was also informative of the value of orange),
748 stimuli were represented as a distribution in feature space, instead of being represented as only their
749 specific colour and shape angles (Fig 3). A stimulus on trial t was represented as a feature vector \mathbf{x}_t .
750 Note that, as each stimulus was made up of two feature dimensions, it was represented by two feature
751 vectors: one for the slow, $\mathbf{x}_{t,S}$, and one for fast-changing feature, $\mathbf{x}_{t,F}$ (corresponding to colour/shape
752 as determined by the current block condition). Therefore, the feature vector for a stimulus \mathbf{x}_i was

753 the concatenation of the slow and fast feature vectors: $\mathbf{x}_t = [\mathbf{x}_{t,S}, \mathbf{x}_{t,F}]$. The feature vectors for the
 754 slow and fast feature angles of a stimulus were obtained from a von Mises like distribution, which
 755 approximates a normal distribution in circular space, as follows:

$$x_{t,i} = \frac{e^{\cos(d_{t,i})\kappa}}{\sum_{i=1}^{360} e^{\cos(d_{t,i})\kappa}} \quad (8)$$

756 where:

$$d_{t,i} = \frac{\theta_t - \theta_i}{360} 2\pi \quad (9)$$

757 where $x_{t,i}$ is the i th entry of feature vector \mathbf{x}_t , and $d_{t,i}$ is the distance from the stimulus' feature
 758 angle on trial t to feature angle i . The parameter κ determines the concentration of the function. With
 759 large κ , the distribution becomes concentrated around the stimulus feature angle, and less surrounding
 760 angles are included. With κ approaching 0, the distribution becomes uniform. Representing stimuli
 761 in this way allowed the model to learn about the value of unobserved angles, based on perceptual
 762 similarity.

763 For each of the two feature dimensions, the models learned a feature weight vector, $\mathbf{w}_{t,S}$ and
 764 $\mathbf{w}_{t,F}$, which were concatenated in the weight vector $\mathbf{w}_t = [\mathbf{w}_{t,S}, \mathbf{w}_{t,F}]$. This vector corresponds to the
 765 estimated value for each feature position on trial t . The expected value V_t of a stimulus on trial t was
 766 calculated as the inner product of the feature vector \mathbf{x}_t with the weight vector \mathbf{w}_t :

$$V_t = \mathbf{x}_t^T \mathbf{w}_t \quad (10)$$

767 This value estimate flowed into the prediction of the choice on the next trial and could guide choices
 768 to maximise reward. However, before being fully guided by value estimates, it is necessary to gather
 769 information and become certain that the estimates are meaningful (as participants do, see Fig 2b).
 770 To mediate between the pressures of exploring and exploiting, we supplemented the value estimate
 771 for each stimulus with an exploration bonus U_t , which reflects how uncertain the model is in its value
 772 estimate. The value of accepting stimulus on trial t , $V_{a,t}$, was then calculated as follows:

$$V_{a,t} = V_t + c \cdot U_t \quad (11)$$

773 where c mediates how strongly the exploration bonus is weighted at choice.

774 Due to the continuous nature of the features and the flexible recombination of features across
 775 stimuli, a simple count-based uncertainty estimate (as in the Upper Confidence Bound method [49])
 776 would be ineffective. Instead, specifying the models as Kalman Filters allowed us to take a rigorous
 777 approach to estimating the uncertainty on each trial. In addition to tracking a mean value, Kalman
 778 Filters keep an estimate of the variance around that mean, which embodies the uncertainty inherent
 779 to the estimate. Similar to the feature and weight vectors, the variance estimates were saved in a
 780 variance vector \mathbf{v}_t , which was a concatenation of slow and fast variance vectors: $\mathbf{v}_t = [\mathbf{v}_{t,S}, \mathbf{v}_{t,F}]$. The
 781 exploration bonus was the inner product of the feature vector with the variance vector:

$$U_t = \mathbf{x}_t^T \cdot \mathbf{v}_t \quad (12)$$

782 While the features shown on each trial changed, the mapping between the feature and the reward
 783 was stationary within each block. Therefore, the uncertainty was highest at the beginning of each
 784 block and steadily reduced with each observed outcome.

785 When predicting the next choice, the models compared the value of accepting $V_{a,t}$ with the value
 786 of a rejecting, by testing for the probability of $V_{a,t}$ under a cumulative normal distribution centred on
 787 50, with a standard deviation σ :

$$\begin{aligned} p(\text{accept}) &= P[X \leq V_{a,t}] \\ X &\sim N(50, \sigma^2) \end{aligned} \quad (13)$$

788 Here a smaller σ means a steeper increase in accept probability with increasing $V_{a,t}$.

789 After an 'accept' choice the reward outcome R_t of the trial t is used to update the value and
 790 uncertainty estimates. The reward prediction error is used to update weight vector with a learning
 791 rate α_t , as follows:

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha_t \mathbf{x}_t (R_t - V_t) \quad (14)$$

792 The variance vector is reduced by an amount proportional to the learning rate α_t :

$$\mathbf{v}_{t+1} = \mathbf{v}_t - \alpha_t \mathbf{x}_t \mathbf{v}_t \quad (15)$$

793 Finally, the Kalman Filters also update the learning rate on each trial, as with decreasing uncertainty
794 about the value estimates, smaller updates to the weight vector are needed.

$$\alpha_{t+1} = \frac{U_t}{U_t + M} \quad (16)$$

795 where M is the constant measurement noise.

796 All four learning models included the three free parameters, κ , c and σ , as specified in the equations
797 above, but they differed in their ability to adapt their learning rates to the slowness of the features
798 (Fig 3). A one learning rate ($1LR$) model used the same learning rate α regardless of feature speed
799 and thus was indifferent to feature variability and could not account for a difference in performance
800 between the slow and fast blocks. A two learning rates model sensitive to feature variability ($2LR_f$)
801 used different learning rates for the slow α_S and fast α_F changing feature across all blocks, irrespective
802 of whether they were relevant or irrelevant. Another two learning rates model, this one sensitive to
803 block condition ($2LR_c$), used different learning rates, depending on whether the relevant feature was
804 changing slowly α_S or quickly α_F (but used the same learning rate for both features within the block).
805 Finally, a four learning rates ($4LR$) model had learning rates sensitive to both the feature variability
806 and the block condition. Meaning it had separate learning rates for the slow and fast-changing features
807 when they were relevant ($\alpha_{S,R}$, $\alpha_{F,R}$) and irrelevant ($\alpha_{S,I}$, $\alpha_{F,I}$).

808 In models with separate learning rates for the slow and fast feature ($2LR_f$ and $4LR$), the uncertainty
809 U_t (equation 12) and learning rates α (equation 16) were calculated separately for the slow $\mathbf{x}_{t,S}$ and
810 fast $\mathbf{x}_{t,F}$ feature vector. Accordingly, the weight and variance vectors for the slow and fast features
811 were updated with their respective learning rates. To keep comparable magnitudes of learning rates
812 between models, in models with the same learning rate for both features in a block ($1LR$ and $2LR_c$),
813 we calculated the uncertainty separately for the slow and fast feature and used their mean to update
814 the learning rate according to equation 16.

815 **Control models** We implemented a control model with the same Kalman Filter machinery, which
816 treated the task as a single, stationary bandit for which it estimated a mean and variance (Bandit
817 model). By ignoring the stimulus features, this model could only learn from the reward outcomes.
818 This model was critical to rule out that learning might be easier on slow blocks, simply due to the
819 reward on the current trial being more predictive of the reward on the next trial, irrespective of the
820 variability of the features. Equations were similar to the models of interest, obviating the need for
821 vectors. A single value V and uncertainty U estimate were kept. These were combined as in equation
822 11 to the value of accepting V_a with the mediating parameter c . The same choice rule as in equation
823 13 was used. The value and uncertainty estimates, and the learning rate were updated according to:

$$V_{t+1} = V_t + \alpha_t (R_t - V_t) \quad (17)$$

$$U_{t+1} = U_t - \alpha_t U_t \quad (18)$$

$$\alpha_{t+1} = \frac{U_t}{U_t + M} \quad (19)$$

824 where M is the constant measurement error.

825 To account for a choice perseverance strategy, which could selectively benefit performance in slow
826 blocks where the correct choice on the previous trial was likely the same as the correct choice on the
827 current trial, we included a win-stay-lose-shift model (WSLS model). When the choice on the previous
828 trial was ‘accept’ and the received reward was equal to or above the default value of 50, this was
829 counted as a win and the model was likely to choose ‘accept’ again. In contrast, if the outcome of an
830 ‘accept’ choice lay below 50, this was counted as a loss and the model was likely to choose ‘reject’ on
831 the next trial. In both cases the model could instead make the less likely choice with probability ϵ .
832 As ‘reject’ choices always resulted in a reward of exactly 50 no wins or losses as such were possible,
833 so the model continued to make ‘reject’ choices and switched to ‘accept’ with probability ϵ . The first
834 choice was made randomly. The WSLS model can be described as follows:

$$p(\text{accept}) = \begin{cases} 1 - \epsilon, & \text{if choice}_{t-1} = \text{accept and } R_{t-1} \geq 50. \\ \epsilon, & \text{otherwise.} \end{cases} \quad (20)$$

835 In addition, we set up models which did not learn and responded randomly, with either a bias to
836 ‘accept’ or ‘reject’ (Random Choice model), with choices given by:

$$p(\text{accept}) = b_a \quad (21)$$

837 or a bias for the left or right response key (Random Key model), with choices given by:

$$p(\text{accept}) = \begin{cases} b_r, & \text{if right key is ‘accept’}. \\ 1 - b_r, & \text{otherwise.} \end{cases} \quad (22)$$

838 **Model fitting** Models were fit to each participant’s data in the training trials using the `nloptr`
839 package version 2.0.3 in R by minimising the log likelihood with the `NLOPT_GN_DIRECT_L` optimisation
840 function run for 10.000 evaluations. We initialised the learning models and the Bandit model, so
841 that on the first trial of each block, the value estimate of the stimulus V_t was 50 (the same as the
842 value of rejecting), and the uncertainty bonus U_t was 5 for each feature. At the start of fitting, the
843 measurement error M was adjusted so that the learning rate α_t on the first trial would be equal to
844 the fit learning rate (equations 16 and 19).

845 We quantified the reliability of parameter estimates through parameter recovery for the learning
846 rates of the learning models (see S6 Fig). The fitting procedure provided fair to excellent reliability,
847 with a high correspondence between ground truth and recovered learning rates.

848 **Model comparison** We simulated model choices given the parameter values obtained from maxi-
849 mum likelihood fitting and obtained the predicted likelihoods for participant choices. These likelihoods
850 were used to calculate the Akaike Information, corrected for small samples [74]:

$$AICc = 2k - 2LL + \frac{2k(k+1)}{N - k - 1}$$

851 Where k is the number of free parameters of the model, LL is the log likelihood of the data given
852 the model and fit parameters and N is the sample size.

853 We then calculated $AICc$ weights, which provide a measure of goodness of fit of a model relative
854 to a baseline model (for which we chose the 1LR model) [50], as follows:

$$AICc \text{ weight} = \frac{e^{-\frac{1}{2} \Delta AICc}}{\sum_{m \in M} e^{-\frac{1}{2} \Delta AICc_m}} \quad (23)$$

855 where $AICc_\delta$ is the difference in $AICc$ between the $AICc$ of the model and the baseline model,
856 and M is the set of all models m . $AICc$ weights are normalised to sum to one for each partici-
857 pant, with larger values indicating a better fit. Finally, we used $AICc$ weights as an approxima-
858 tion of model evidence to calculate protected exceedance probabilities with the `bmsR` package in R
859 (<https://github.com/mattelisi/bmsR>) [51].

860 We tested model identifiability through model recovery, using the same fitting and model compar-
861 ison procedure as for participants (see S6 Fig). Model recovery proved to be reliable, identifying the
862 model that had generated the data correctly for most simulations.

863 Acknowledgements

864 The authors would like to thank Ondrej Zika and Christoph Koch for helpful discussions on the models
865 and Anika Löwe and Georgy Antonov for their valuable feedback on the manuscript.

866 Data and Code availability

867 The code and data used to produce the results and analyses presented in this manuscript will be made
868 freely available upon publication.

869 Supporting Information

- 870 • **S1** Figure. Participant behaviour in the pilot experiment.
- 871 • **S1** Text. Preregistration description.
- 872 • **S1** Table. Overview of analyses in the pre-registration (PR) and the paper.
- 873 • **S2** Figure. Effect of additional parameters on the slowness prior effect.
- 874 • **S2** Text. Effect sizes and confidence intervals for the mixed effects models.
 - 875 – **S2** Table. Best model predicting predicting cumulative reward in the learning
 - 876 phase.
 - 877 – **S3** Table. Best model predicting correct choices in the learning phase
 - 878 – **S4** Table. Mixed effects model predicting participant learning phase choices
 - 879 from the feature positions in the slow blocks.
 - 880 – **S5** Table. Mixed effects model predicting participant learning phase choices
 - 881 from the feature positions in the fast blocks.
 - 882 – **S6** Table. Best model predicting correct choices in the test phase.
 - 883 – **S7** Table. Best model predicting choices from the reward difference in the test
 - 884 phase.
- 885 • **S3** Figure. Models fit individual participant learning curves.
- 886 • **S4** Figure. Model performance in the slow and fast condition.
- 887 • **S5** Figure. Development of the learning rates of the four learning rate model.
- 888 • **S6** Figure. Parameter and model recovery.