

1 Comparing rapid rule-learning strategies in humans
2 and monkeys

3 Vishwa Goudar*¹, Jeong-Woo Kim *¹, Yue Liu¹, Adam J. O. Dede²,
4 Michael J. Jutras², Ivan Skelin^{4,5}, Michael Ruvalcaba⁷, William Chang⁷,
5 Adrienne L. Fairhall², Jack J. Lin^{4,5}, Robert T. Knight^{6,7}, Elizabeth A.
6 Buffalo^{2,3}, and Xiao-Jing Wang †¹

7 ¹*Center for Neural Science, New York University, NY, USA*

8 ²*Department of Physiology and Biophysics, University of Washington, Seattle, WA, USA*

9 ³*Washington Primate Research Center, University of Washington, Seattle, WA, USA*

10 ⁴*Department of Neurology, University of California, Davis, Davis, CA, USA*

11 ⁵*The Center for Mind and Brain, University of California, Davis, Davis, CA, USA*

12 ⁶*Department of Psychology, University of California Berkeley, Berkeley, CA, USA*

13 ⁷*Helen Wills Neuroscience Institute, University of California Berkeley, Berkeley, CA,*

14 *USA*

*equal contribution

†Corresponding Author: xjwang@nyu.edu

15

Abstract

16 Inter-species comparisons are key to deriving an understanding of the behavioral
17 and neural correlates of human cognition from animal models. We perform a
18 detailed comparison of macaque monkey and human strategies on an analogue of
19 the Wisconsin Card Sort Test, a widely studied and applied multi-attribute measure
20 of cognitive function, wherein performance requires the inference of a changing
21 rule given ambiguous feedback. We found that well-trained monkeys rapidly infer
22 rules but are three times slower than humans. Model fits to their choices revealed
23 hidden states akin to feature-based attention in both species, and decision processes
24 that resembled a Win-stay lose-shift strategy with key differences. Monkeys and
25 humans test multiple rule hypotheses over a series of rule-search trials and perform
26 inference-like computations to exclude candidates. An attention-set based learning
27 stage categorization revealed that perseveration, random exploration and poor
28 sensitivity to negative feedback explain the under-performance in monkeys.

29 Introduction

30 Animal models are essential for mechanistic investigations of the circuit underpinnings of
31 complex computation. New frontiers in the training of non-human animals to perform
32 computationally challenging tasks while simultaneously recording from large neural
33 populations across several brain regions promise rapid advances in our understanding of
34 the neural substrates of cognition. However, our ability to extrapolate any findings to an
35 understanding of human cognition relies on an overlap between the computational and
36 neurocognitive means used to carry out complex tasks across species [1, 2]. The prefrontal
37 cortex, which plays an essential role in higher cognitive functions [3], is disproportionately
38 enlarged in humans compared to macaque monkeys [4, 5]. It has been argued that
39 the resulting increase in the number of neurons may underlie superior human cognitive
40 abilities [6, 7]. Thus, interpretation of findings from animals demands systematic and
41 rigorous comparisons between cognitive computations in humans and non-human animals.

42 Towards this end, we compared the behavioral strategies employed by macaque
43 monkey and humans on the same task: a variant of the Wisconsin Card Sorting Test
44 (WCST), widely used to evaluate the cognitive functions involved in abstract thinking,
45 rule search, cognitive set shifting and the effective use of feedback [8, 9]. The WCST and
46 its variants have long been employed in the study of prefrontal function and dysfunction
47 [10, 11, 12, 13, 14], lending support to the presence of abstract thinking and computation
48 in the monkey brain [15]. In the task, subjects must match or select items composed of
49 multiple visual features based on an uncued or hidden rule. Feedback at the end of each
50 trial indicates whether the response was correct, but does not unambiguously reveal the
51 rule identity. Instead it helps narrow down the rule identity, which must then be inferred
52 from the collective outcome of multiple trials. Critically, the hidden rule changes in an
53 uncued manner across trial blocks, requiring detection of and adaptation to these uncued
54 rule switches based solely on positive or negative feedback.

55 We implemented a version of the task in which each card has three feature dimensions
56 (color, shape and texture), each of which can take one of four values, defining twelve

57 possible “rules” (Fig. 1a). What strategy might one use to identify the current rule
58 from the outcome of choosing an object in each trial? It is clearly extremely inefficient
59 to simply learn the value of 64 individual object-reward associations; each object is a
60 conjunction of multiple visual features, and learning the value of one object does not
61 generalize to any of the other objects. Learning the value of features to identify the rule
62 is far more efficient as it allows generalization. In reinforcement learning, the problem of
63 attributing binary feedback in situations when there are many features (high-dimensional
64 environments), referred to as the “curse-of-dimensionality”, is effectively resolved through
65 such abstract reasoning [16, 17, 18]. However, a rule-based strategy still poses challenges
66 of cognitive capacity. Simultaneously tracking and updating the value of twelve features
67 can impose prohibitive working-memory demands and be computationally daunting. On
68 the other hand, selectively attending to and evaluating a single feature at a time is
69 inefficient as it discards relevant feedback regarding other features. The strategy that
70 humans or monkeys use to address this tradeoff between computational complexity and
71 information efficiency remains to be elucidated.

72 To analyze how individuals handle the cognitive complexity and demands of this task,
73 we developed a detailed behavioral model that allowed us to compare the strategy and
74 performance of humans and trained macaque monkeys on the same WCST analogue.
75 This allowed us to identify differences in the types of errors in the two species, revealing
76 the underlying cognitive differences between them. Indeed, differences in the types and
77 prevalence of errors on the WCST, demonstrated via similar comparisons between healthy
78 and patient populations, has played an essential role in establishing behavioral markers
79 of various forms of cognitive dysfunction [10, 19, 20, 21]. We found that while monkeys
80 rapidly identified new rules and learned several rules over an individual session, the rule
81 learning in monkeys was 3-4 times slower than humans. To understand this difference,
82 we fit a behavioral model that predicts upcoming choices based on choices and their
83 outcomes on previous trials. We modeled the transformation of the trial history to an
84 upcoming choice through inferred hidden behavioral states. While the model is not
85 constrained to do so, the best-fit models for all subjects of both species developed hidden
86 states that aligned with a feature-based attention strategy wherein some visual features

87 are selectively examined over others while making a choice and attributing feedback.
88 The decision process governing each species' rule search strategy is characterized by the
89 statistics of the inferred transitions between these states. This strategy bore a striking
90 resemblance to the Win-Stay Lose-Shift strategy but with a few important differences.
91 First, our model revealed that both species often explore more than one features at a time.
92 Second, both species perform inference-like computations – the model reported changes
93 in the attentional state towards one feature based on the outcome of choosing another.
94 We developed a novel approach to identify distinct stages of rule-learning. Analysis of
95 these stages revealed three key reasons for slower monkey rule-learning performance: (*i*)
96 following a rule switch, monkeys persevere on the previous rule more than humans; (*ii*)
97 monkeys often make seemingly random choices that do not involve any of the features
98 under exploration, even after finding the rule, which delays the expression of the learnt
99 rule; and (*iii*) poorer attention to negative feedback in monkeys particularly when they
100 simultaneously explore the rule and non-rule features poses a credit-assignment challenge
101 which delays rule learning.

102 Results

103 Monkeys are slower rule learners than humans

104 We compared the ability of monkeys and humans to rapidly adapt to changes in task
105 contingencies and learn new rules in a modified version of the WCST. On each trial,
106 subjects were presented with 4 objects and received feedback upon selecting one of them
107 (Fig. 1a, middle). Each object was composed of one unique feature from each of three
108 dimensions - visual pattern, shape and color (Fig. 1a, top). Every feature appeared in
109 one of the objects on each trial, but object compositions changed across trials. On a
110 given block of trials, subjects received positive feedback (monkeys: food reward; humans:
111 the word “CORRECT” displayed on screen) for selecting the object with one of the
112 twelve features (e.g. red) — the current hidden rule — and negative feedback (monkeys:
113 timeout; humans: the word “INCORRECT” displayed on screen) otherwise. After
114 subjects demonstrated that they had learned the current rule by reaching criterion
115 performance on the current block, a new block was initiated through an uncued switch
116 to a randomly chosen new rule (Fig. 1a, bottom). Parameter differences in the task
117 implementation for the two species, including the response type, trial epoch durations
118 and learning criteria, are outlined in Supplementary tables 1 and 2.

119 Remarkably, well-trained monkeys learned new rules within only tens of trials. Yet,
120 they were over three times slower than humans (Fig. 1b; monkeys: 27.84 ± 2.92 trials
121 (mean \pm SEM); humans: 5.98 ± 0.52 trials), a learning deficit that was significant following
122 a correction for the inter-species difference in the learning criterion (Supplementary Fig.
123 1; monkeys: 20.61 ± 1.52 trials; humans: 5.98 ± 0.52 trials; bootstrap test with t -statistic,
124 $p < 0.005$). We then sought to explain the inter-species computational differences that
125 produce this rule learning slowing in monkeys. Specifically, we focused on inferring
126 individuals’ rule-learning strategies from behavior, and identifying the species differences
127 that contribute to the learning speed difference. One possible strategy, Win-Stay Lose-
128 Shift (WSLS), is widely-reported during rule learning in many species, particularly in
129 the two-armed bandit problem where the identity of the more rewarding arm must be

130 learned and can change over trials. Here, one of the arms is repeatedly chosen as long
131 as this produces positive feedback (win-stay). When negative feedback is received the
132 other arm is chosen on the next trial (lose-shift). This strategy can be cast as a decision
133 process comprised of two behavioral states — persist and avoid — where the choice of
134 the currently rewarded arm is in the persist state and it transitions to the persist or
135 avoid states subject to positive or negative feedback, respectively, while the choice of the
136 other arm is in the avoid state and transitions to the persist state when that arm was
137 not chosen on the previous trial and negative feedback was received (Fig. 1c).

138 The WSLS strategy is computationally efficient, and requires that the subject attend
139 to and maintain only a single arm’s identity in working memory. By replacing arm iden-
140 tity with feature identity, the approach is readily adapted to solve WCST problems and
141 always finds the rule. However, feedback is equally informative about all three features
142 in the chosen object, not just the attended one. Due to this neglect of information about
143 unattended features, a simulated WSLS agent learns rules much slower than optimal:
144 indeed humans learn more rapidly (Supplementary Fig. 1; WSLS agent mean: 13.31
145 trials, std. dev.: 12.85 trials; humans: 5.98 ± 0.52 trials). This underscores a tradeoff
146 between computational and information efficiency in multi-dimensional environments.
147 Simultaneously maintaining and updating beliefs about multiple features is more informa-
148 tion efficient, but increases computational complexity and working memory demands. In
149 contrast, attending to a single feature at a time is computationally simpler but inefficient
150 in its integration of trial outcomes. In the following sections, we address how the two
151 species solve this tradeoff.

152 **Dynamic model uncovers hidden states during rule learning**

153 Prior cognitive model comparisons of human behavior in WCST variants provide evi-
154 dence for rule-learning strategies wherein subjects selectively attend to and learn about
155 individual features or dimensions, rather than feature configurations (i.e. objects) [17,
156 18, 20]. It is argued that such a mental representation of stimuli in terms of features
157 resolves the curse-of-dimensionality which impairs learning efficiency in high-dimensional

158 environments. For example, it is far more efficient to learn the value of 12 features than
159 the dozens of objects they can be combined into. Drawing on these findings, we developed
160 a behavioral model to predict the probability of choosing individual features given their
161 choices and outcomes on previous trials. However, in contrast to earlier work, our model
162 does not postulate a specific internal belief structure and update rule, thus making fewer
163 assumptions regarding the learning algorithm underlying a subject's behavior. Instead,
164 it aims to discover in an unbiased manner how the decision making process evolves as a
165 function of feedback. Recently, this approach has been successful at revealing previously
166 unobserved behavioral states underlying human, rodent and fruit-fly decision making [22,
167 23, 24].

168 For each feature, we model whether the feature is chosen or not (denoted as c) as a
169 function of past choices and their outcomes (h) via a Bernoulli Generalized Linear Model
170 (GLM) (Fig. 2a; see *Methods*). The choice outcome on an earlier trial is represented
171 by a one-hot four dimensional vector where the dimensions represent whether positive
172 feedback was received after choosing the feature on the trial (C^+), negative feedback
173 was received after choosing the feature (C^-), positive feedback was received after not
174 choosing the feature on the trial (NC^+), or negative feedback was received after not
175 choosing the feature (NC^-). This allows us to assess separately how the present choice
176 depends on past choice outcomes both when that feature was chosen (direct) and when it
177 was not (indirect). Furthermore, the model permits dynamic changes in how past choices
178 and outcomes are transformed into a present choice via hidden states (s). A feature's
179 associated hidden-state also undergoes a transition at the end of each trial depending
180 on past choices and outcomes, which may reflect updates to the feature's value based
181 on past choice outcomes, or a change in the level of attention to the feature, or even a
182 shift in strategy (i.e. how a feature's history is weighted in determining its choice). Note
183 that while the model permits these possibilities and others, it does not prescribe the
184 nature and function of the states. Rather, the states and their dynamics emerge upon
185 fitting the model to behavioral data. These hidden state dynamics are modeled as an
186 input-dependent or Input-Output Hidden Markov Model (IOHMM) [25].

187 The IOHMM-GLM's goodness of fit to behavior depends both on the number of

188 previous trials determining a subject's choice (lag), and on the number of possible hidden
189 states. Accordingly, we fit IOHMM-GLMs to each subject's behavior while systematically
190 varying these two parameters (Fig. 2b). Across subjects in both species, model accuracy
191 showed a stronger dependence on the number of states than lag. Crucially, accuracy
192 plateaued as the number of states increased, and exhibited over-fitting at higher lags.
193 In this task, subjects choose objects rather than individual features. Therefore, we
194 extended our model to compute the probability of choosing each object in a trial, based
195 on the model's predicted probability of choosing individual features on that trial (see
196 *Methods*). Fits of this model extension to each subject's behavior based on each of the
197 IOHMM-GLMs in Figure 2b revealed a qualitatively similar relationship between model
198 accuracy and the underlying parameters (Supplementary Fig. 2a). For each subject,
199 the best-fit model comprised of 4 states and lag 1 (history from the previous trial only)
200 does not overfit the data while producing prediction accuracies at or very close to the
201 performance plateau. Therefore, we selected these models (dashed-blue box, Fig. 2b,
202 Supplementary Fig. 2a) for further analysis.

203 Figure 2c (left) shows the choice probability predicted by these 4 state lag 1 models
204 for the chosen object at each trial after a rule switch, averaged over rule blocks; averaging
205 across blocks is achieved by normalizing the trial number by the block length. The
206 results show that the model's prediction of the chosen object is significantly above chance
207 (0.25) in both species (monkeys: 0.47 ± 0.02 ; humans: 0.63 ± 0.02). Also, prediction
208 accuracy improves as the rule is learned over the block's time course. Our primary
209 goal, however, is to find the most accurate explanation for each subject's rule learning
210 behavior, rather than predict their future choices. For this, we consider the most likely
211 sequence of states across trials inferred by the model for each feature. Rather than
212 predict the most likely state on each upcoming trial given only past choice outcomes, the
213 most likely sequence of states is the maximum a posteriori probability (MAP) estimate
214 of the sequence of states across all trials in an experimental session — each estimated or
215 inferred state best explains not only the present choice but also past and future choices
216 subject to the model's choice probabilities in the inferred past/future states and its
217 state transition probabilities between the inferred present and past/future states (see

218 *Methods*; Figure 2d). The model is generally quite confident in its MAP estimates of
219 the most likely sequence of states, as evidenced by the the cumulative density of their
220 posterior probabilities (Supplementary Fig. 2b). Moreover, since the inferred states
221 for each feature are estimated from past, present and future choices, they yield more
222 accurate estimates of the choice probabilities for chosen objects (Fig. 2c, right). For this
223 reason, we rely on the inferred states to identify the rule learning strategy in each species
224 and to interpret the inter-species differences therein.

225 To gain insight into the interpretation of our model fits, we similarly analyzed
226 the choices of a simulated Win-Stay Lose-Shift agent (Supplementary Fig. 1). By
227 construction, we know that the model's choices only rely on the previous trial. As
228 expected, higher lag models tend to overfit the agent's choices (Supplementary Fig. 3a).
229 While the agent's true behavior has only two states (Fig. 1c), we find that a three-state
230 model provides a better fit. Our model splits the avoid state into two states — random
231 and avoid. This is due to a combination of the task's structure and our model setup. By
232 picking one feature consistently across trials, the agent necessarily avoids other features
233 in the same dimension. However, the agent's choices of features in other dimensions
234 appear random, since an object is composed of one feature from each dimension and
235 objects compositions are generated randomly on each trial. Thus, the appearance of
236 this additional state results from our model's treatment of each feature independent
237 of its relationship to other intra-dimensional features, a simplifying assumption that
238 allows for tractable fitting. Nevertheless, the model largely recovers the hidden states
239 and state-transitions that drive the agent's behavior — it correctly identifies when a
240 feature is associated with the persist state 57.6% of the time and accurately determines
241 the underlying decision process (Supplementary Fig. 3c-d). Collectively, these results
242 show the reliability of this modeling approach to explain rule-learning in both species.

243 **Hidden states reflect feature-based attention and reveal qualita-** 244 **tively similar strategies in the two species**

245 Learning is often conceptualized as updates to a decision-making schema based on
246 past decisions and their outcomes [18, 26]. We sought to identify hidden states that
247 capture this decision-making process and to explore what they reveal about the *dynamics*
248 of human and monkey rule learning in the WCST. In both humans and monkeys, a
249 comparison of the choice probability of features associated with each state — calculated
250 by marginalizing the model’s predicted choice probability under each state and history
251 (Supplementary Fig. 4a) over the choice outcome histories (Supplementary Fig. 4b) —
252 revealed that the model determines states based on distinct probabilities of choosing
253 the associated feature, ranging from below chance (*avoid*) to chance (*random*) to above
254 chance (*preferred*) to very high (*persist*) (Fig. 3a). That is, the model states correspond
255 to levels of attention paid to each feature. Moreover, this result was consistently observed
256 in the majority of the models fit to the behavior in both species, as well as in a simulated
257 WSLs agent (Supplementary Fig. 3). Since features associated with the preferred or
258 persist state are favored during rule-learning, we refer to them as being under exploration.
259 We will show that the estimation of the attentional state towards each feature at each
260 trial permits a systematic analysis of when features are selected for or withdrawn from
261 exploration, and how the choice outcome history informs these decisions. This exercise
262 fosters an exposition of the decision-making process that describes the rule learning
263 strategy in both species, the resulting learning dynamics between rule switch and rule
264 learning, and the identification of those differences in the decision-making process that
265 most prominently explain the learning performance difference between the two species.

266 Since these analyses rely heavily on the most-likely state estimates, we validated
267 the consistency of these estimates with the model’s parameters. First, we compared
268 the feature choice probability per history and state computed directly from the fit
269 parameters (model) and measured based on the state estimate for each feature on each
270 trial (empirical). The two measurements yield consistent results demonstrating that the
271 most-likely states are estimated not only to best explain the sequence of choices but also

272 to conform with the model's parameters. Next, we similarly compared state transition
273 probabilities per history computed directly from the fit parameters (model) and measured
274 based on the state estimate for each feature on each trial (empirical)(Supplementary Fig.
275 5). Here again, we find that the transition probabilities computed from the fit parameters
276 (model) are consistent with empirical measurements of the transition statistics based
277 on the most-likely state estimates. Figure 3b schematizes the decision process in the
278 two species derived from their state transition probabilities. The thickness of an arrow
279 indicates the probability of the respective transition; extremely rare transitions have been
280 pruned. Similar to the WSLs agent, we find that a feature is most often associated with
281 the avoid state while an intra-dimensional feature is simultaneously under exploration
282 (Supplementary Fig. 4c). Since the avoid state likely emerges due to this interdependence
283 between the choices of inter-dimensional features, which our model forgoes for tractability,
284 we do not treat it as distinct from the random state.

285 We would like to compare observed behavior in the WCST with a WSLs strategy. To
286 do so, we must take into account task structure differences between the WCST and the
287 2-armed bandit task (Fig. 1c). The composition of a chosen object by three features forces
288 the choice of features in the avoid state of the WSLs strategy. Thus a WSLs decision
289 process for the WCST must define transitions for such features when they are chosen.
290 Moreover, the multi-dimensional environment of the WCST offers multiple alternatives
291 for a subject to shift attention to during lose-shift, compared to the 2-armed bandit task.
292 An updated WSLs strategy that accounts for these differences is depicted in Figure 3b
293 (right). We can now compare the decision process inferred by the model for the two
294 species (Fig. 3b, left-middle) to this WSLs decision process, revealing salient differences
295 that are delineated by dashed lines. Key among these is the existence of a *preferred* state
296 where items are not chosen with certainty (or near-certainty) as in the *persist* state, but
297 above chance. The effect of direct feedback (as a result of choosing the feature) on these
298 states and the random/avoid states is similar to those in the WSLs decision process.
299 For example, both species select a feature in the random/avoid state for exploration
300 (by promoting it to the preferred state) seemingly at random after receiving negative
301 feedback for choosing other features. However, an interesting exception is that humans

302 sometimes choose to explore such a feature upon receiving positive feedback for choosing
303 it.

304 Larger differences emerge with regards to indirect effects of feedback. A feature may
305 not be chosen on a trial when it is associated with the preferred state (feature choice
306 probability in preferred state < 1 , Fig. 3a). However, its state may still transition subject
307 to the feedback received at the end of the trial — an indirect effect. For example, humans
308 and, to a lesser extent, monkeys demote features from the preferred to the random/avoid
309 state upon receiving positive feedback for choosing a different feature. Consequently,
310 their probability of subsequently choosing an unchosen feature that was associated with
311 the preferred state decreases (Fig. 3d, right). They also promote features from the
312 preferred to the persist state upon receiving negative feedback for choosing a different
313 feature. Consequently, their probability of subsequently choosing an unchosen feature
314 that was associated with the preferred state increases (Fig. 3d, left). This is striking
315 because it is the only way a feature can transition into the persist state, which appears
316 to be reserved mainly for a feature that the subject determines to be the rule (Fig.
317 2d). Receiving positive feedback for choosing a feature in the preferred state does not
318 definitively confirm that it is the rule, since the rule may be among the other two features
319 in the chosen object. Confidence in the rule's identity may be increased based on the
320 consistency of receiving such direct positive feedback across many trials. Alternatively, it
321 may be done by ruling out other candidates, that is, after receiving negative feedback for
322 choosing an object with a different candidate feature. Consistent with this interpretation,
323 measurements show that when a feature under exploration is not chosen, the object that
324 is chosen often contains a different feature that is also under exploration (Fig. 4d).

325 This approach of promoting a feature to the persist state as an inferred consequence of
326 ruling out an alternative candidate, rather than integrate direct positive feedback across
327 trials in favor of the feature, may be favored by both species due to its computational
328 simplicity — it relies on the outcome of just the previous trial rather than multiple
329 trials and thereby reduces working memory demands. It is possible that these inference-
330 like computations are not deliberate but an inadvertent consequence of demoting or
331 promoting an intra-dimensional chosen feature. For example, given that the probability

332 of choosing all shapes must sum to 1, when one shape is demoted after its choice produces
333 negative feedback, the probability of choosing another shape that was associated with
334 the preferred state may automatically increase, forcibly promoting it to the persist
335 state. Measurements of the probability of demoting (promoting) the chosen feature
336 while promoting (demoting) an unchosen intra-dimensional feature in the preferred state
337 are mixed: monkeys do so at chance levels; humans always (seldom) demote (promote)
338 the chosen feature (Fig. 3c). Nevertheless, these indirect-effect transitions directly and
339 significantly alter the subsequent choice probability of the unchosen feature (Fig. 3d). In
340 summary, the best-fit models discover feature-based attentional states whose dynamics
341 show marked deviations from a Win-Stay Lose-Shift strategy.

342 **Both species simultaneously evaluate multiple features over sev-** 343 **eral trials during rule-learning**

344 The explore-exploit dilemma pits the benefit of continuing to select a recently rewarded
345 option (exploit) against the benefit of selecting a different and potentially more rewarding
346 (but possibly less rewarding) option (explore). While much work has been done to
347 determine how humans and other animals navigate this dilemma [27, 28, 29], how they
348 deal with it in a multi-dimensional environment with transiently overlapping options
349 remains unclear. Which of the three features of a chosen and rewarded object should
350 be exploited on the next trial, given that they are unlikely to appear co-located in the
351 same object on the following trial? How should the tradeoff between the computational
352 complexity and information efficiency of exploring several features at once be resolved?

353 The model finds that both species continuously explore one or more features (Fig.
354 4a). In the process, they explore multiple features over the course of a block before
355 ultimately identifying the rule (Fig. 4b). Moreover, each feature is often explored for
356 a series of several trials in both species (Fig. 4c). But the number of these trials is
357 substantially larger in monkeys, a finding we analyze more closely in the following sections.
358 The model also indicates that both species often explore multiple, but not all, features

359 at a time (Fig. 4e). This is consistent with the theory of selective attention [30, 31]
360 wherein objects are selectively attended to (or filtered for higher processing) subject to
361 an internally-maintained set of relevant perceptual features (or attentional filters). It also
362 underscores the solution of both species to the computational complexity-information
363 efficiency tradeoff. Since it is computationally challenging to simultaneously attend to
364 and evaluate all twelve features over several trials but inefficient to attend to one feature
365 at a time, both species evaluate a small subset of all features at a time. Indeed, during
366 exploration it is uncommon for either species to select an object where none of the
367 features is in the preferred or persist states (Fig. 4d). However, monkeys do engage in
368 such random exploration much more frequently than humans.

369 From these results, we conclude that both species exhibit a deliberate form of
370 exploration to address the challenges inherent in the task environment. Features, often
371 more than one at a time, are selected for exploration via promotion to the preferred
372 state after choices of other features produce negative feedback (Fig. 3b). They are
373 then continuously explored so long as they produce positive feedback, until alternatives
374 are ruled out at which point they are then promoted to the persist state, or they are
375 themselves ruled out after choosing them produces negative feedback (or, in the case of
376 humans, choosing other features produces positive feedback).

377 **Categorization of feature attentional states characterizes learning** 378 **dynamics**

379 Rule learning proceeds through a sequential process that progressively reduces ambiguity
380 regarding the rule's identity until it is ultimately determined. Our model reveals elemen-
381 tary feature-specific computations that individuals in each species apply to maintain and
382 update a small subset of candidate features, one of which may determine the rule. In
383 order to elucidate the resulting over-arching learning dynamics that governs the sequential
384 rule learning process, we developed a simple approach to categorize individual trials
385 based on the features under exploration and the true rule (Fig. 5a). The categories are

386 mutually exclusive and exhaustive – each trial falls into one and only one category. These
387 are:

- 388 • “perseveration”: a continuous series of trials following a rule switch when the feature
389 governing the previous rule is associated with the persist state,
- 390 • “random search”: trials when none of the features are under exploration (i.e.
391 associated with the preferred or persist states),
- 392 • “non-rule exploration”: trials when one or more features are under exploration not
393 including the rule feature,
- 394 • “rule-favored exploration”: trials when one or more features including the rule are
395 under exploration,
- 396 • “rule preferred”: trials when only the rule is associated with the preferred state,
- 397 • “rule exploitation”: trials when only the rule is associated with the persist state.

398 We compare the distribution of categories for trials across the course of a rule block
399 between species (Fig. 5b). Humans show a progression from perseveration to non-
400 rule exploration, where non-rule features are explored and ruled out, to rule-favored
401 exploration, where the rule feature is simultaneously explored with non-rule features, to
402 rule exploitation, once other candidates are ruled out and the rule is identified. Monkey
403 rule learning is described by similar dynamics except for a much higher incidence of
404 the rule preferred category for most of the block and a larger (smaller) proportion of
405 the blocks ending in the rule-favored exploration (rule exploitation) category. That
406 is, a significant subset of blocks end with the model indicating that the monkey is
407 simultaneously exploring the rule and at least one other non-rule feature, even if the
408 monkey has met a learning criterion. These results show that our categorization approach
409 expresses human and monkey rule learning dynamics in terms of behaviorally interpretable
410 learning stages; for example, an increase in the reward rate following a rule switch in
411 both species is marked by the onset of rule exploration with the rule-favored exploration
412 category (Fig. 5c, bottom).

413 Examining the number of trials spent in each category determined bottleneck cate-
414 gories that produce the rule learning performance deficit in monkeys (Fig. 5c). Specifically,

415 monkeys spend much longer perseverating on the previous rule, in disambiguating the
416 rule feature from non-rule features (rule-favored exploration) and demonstrating that
417 they have learned the rule (rule preferred or exploitation). The latter two sources of the
418 learning performance deficit in monkeys also explain a majority of the variance in their
419 performance across blocks (Supplementary Fig. 7, bottom). In contrast, the number
420 of trials humans spend exploring non-rule features before selecting the rule feature for
421 exploration (non-rule exploration) largely determines the variance in their rule-learning
422 performance.

423 **Random exploration prolongs the expression of learning in mon-** 424 **keys**

425 A key difference between the two species identified via this trial categorization is that
426 monkeys spend many more trials than humans in the rule preferred or exploitation
427 categories. These extra trials spent demonstrating or expressing that the rule has
428 been learned significantly increases both the block length mean and variance (Fig. 5c,
429 Supplementary Fig. 7). The inter-species difference in learning criteria explains a portion
430 of the difference in the mean length of the rule exploitation category. However, the
431 remaining difference in the mean length as well as the difference in the variance of the
432 category's length is unexplained. We hypothesized that the larger mean and variance
433 of the rule exploitation category's length in monkeys compared to humans (Fig. 6a)
434 may result from their random exploration of other features when a feature is already
435 associated with the persist state (Fig. 3a). This behavior is unique to monkeys and is
436 prevalent even during rule exploitation trials (Fig. 6b) — after they have identified the
437 rule, monkeys occasionally choose objects that do not include the rule feature.

438 To test our hypothesis, we simulated agents that select the rule feature with the
439 same probability as monkeys and humans do during the rule exploitation category and
440 asked how many trials it would take these agents to reach a learning criterion. The
441 results revealed that the trial count distributions of the simulated agents were nearly

442 identical to the corresponding subjects (Fig. 6c), thus confirming our hypothesis. Similar
443 “random errors” have been observed in humans with focal lateral prefrontal lesions on the
444 WCST [19], where they were attributed to distraction, or a failure to maintain the rule in
445 working memory. However, it remains unclear whether the monkeys in our experiments
446 were more distractable than their healthy human counterparts, or deliberately adopted
447 occasional random exploration as part of their strategy — for example to prolong a
448 highly rewarded state.

449 **Reduced sensitivity to negative feedback increases perseverative** 450 **errors in monkeys**

451 Perseverative errors occur when the feature that governed the rule on the previous block
452 continues to be chosen following the rule switch despite receiving negative feedback
453 for the choice. These errors are characteristic of frontal lobe damage and dysfunction
454 [10, 21] and are believed to reflect a cognitive deficit in adapting to changes in task
455 contingencies. Pronounced perseveration error rates are also observed in patients with
456 neuropsychiatric [32, 33, 34] and substance abuse disorders [20, 33]. Interestingly, our
457 model’s association of the persist state with the previous rule feature during consecutive
458 trials immediately following a rule switch suggests that monkeys persevere on the previous
459 rule for several more trials than humans (Fig. 6d). Indeed, direct measurements showed
460 that the probability of choosing the previous rule after a rule switch is consistent with
461 such a state estimate in both species (Fig. 6e).

462 In order to determine the cause underlying the elevated perseveration in monkeys, we
463 asked which choice outcome(s) most explained the difference in continued persistence
464 with the previous rule between the two species. The analysis showed that humans were far
465 more likely to demote the chosen previous rule feature from the persist state in response
466 to negative feedback compared to monkeys (Fig. 6f). Monkeys’ weaker sensitivity to
467 negative feedback parallels that of humans with substance abuse disorders and prefrontal
468 lesions, who also persevere more than healthy controls [20, 21].

469 **Reduced negative feedback sensitivity compromises efficient** 470 **credit assignment and prolongs rule learning in monkeys**

471 The largest contribution to the inter-species difference in rule learning performance is
472 from trials in the rule-favored exploration category where the rule feature is concurrently
473 explored with one or more non-rule features (Fig. 7a - Fig. 7b, left). While it is reasonable
474 to explore the rule for several consecutive trials as it produces rewards, what must be
475 explained is why non-rule features are concurrently explored for many more trials by
476 monkeys. Indeed, monkeys continuously explore individual non-rule features for many
477 more trials during the rule-favored exploration category (Fig. 7b, right). This explains
478 the lengthier duration of the category in monkeys, and is caused by a higher probability
479 of a non-rule feature transitioning back into an exploration state during rule-favored
480 exploration trials (Fig. 7c). Analysis further showed that this inter-species difference in
481 transition probability is explained by a lower sensitivity of monkeys to either form of
482 negative feedback – direct, when the non-rule feature is chosen and negative feedback is
483 received, and indirect, when it is not chosen and positive feedback is received (Fig. 7d).

484 While both species also explore non-rule features during non-rule exploration trials,
485 this category is relatively short in both humans and monkeys (Fig. 5c). So what explains
486 the difference in duration between the two categories in monkeys? Since the non-rule
487 exploration category is followed by rule-favored exploration trials, one possibility is that
488 it is cut short by the onset of exploration of the rule feature as the non-rule feature
489 continues to be concurrently explored for many more trials. However, measurements
490 in monkeys showed that non-rule feature exploration only occasionally spans the two
491 categories (probability = $0.27\% \pm 0.04\%$). Therefore, the number of trials during which
492 a non-rule feature is explored by monkeys is usually much smaller when it happens in
493 the non-rule exploration category than in the rule-favored exploration category.

494 This difference in duration is reflected in a higher probability of a non-rule feature
495 transitioning back into an exploration state during rule-favored exploration trials (Fig.
496 7e). Since the probability of transitioning back into the explore state is a marginalization

497 of its joint probability with the choice outcome it follows, we asked what choice outcome
498 history best explains the transition probability difference between the two categories.
499 Measurements showed that receiving positive feedback for choosing the non-rule feature
500 (C^+) is the key differentiator between the joint probabilities for the two categories (Fig.
501 7f, left). This could either be because the transition probability in response to this
502 history is different under the two categories, or because the frequency of the history
503 is different under them. We found that while the transition probabilities are similar
504 (Fig. 7f, left), monkeys are more likely to have received positive feedback for choosing a
505 non-rule feature under exploration during the rule favored, exploration category (Fig.
506 7f, middle). When the rule feature is concurrently explored with a non-rule feature
507 (rule favored, exploration) the probability of selecting them both when they co-locate
508 in an object is higher. This increases the probability of receiving positive feedback for
509 choosing the non-rule feature, which makes appropriately assigning credit to the rule
510 feature challenging. This underscores the importance of negative feedback sensitivity in
511 demoting non-rule features from exploration states, in the absence of which the duration
512 of concurrent exploration of the rule and non-rule feature(s) is prolonged.

513 Discussion

514 Methodological and technological advances in training and recording from animal models
515 now allow for the study of increasingly complex behaviors in non-humans. However,
516 before interpreting their brain activity as a human-like model of neural computation, it
517 is important to ascertain whether their computational algorithms are human-like. Indeed
518 macaque monkeys and humans learn the structure of tasks in different ways (monkeys
519 via impoverished reward-based feedback, and humans via rich verbal instruction plus
520 feedback), raising the possibility that while they both learn the same tasks, they may
521 enlist different abstractions, cognitive operations and neural mechanisms [2]. Our study
522 aims to assess whether macaques and humans employ similar mental representations and
523 operations to perform a cognitively complex task that relies on several interdependent
524 cognitive processes, such as the Wisconsin Card Sorting Test. The results of such studies
525 can play a crucial role in interpreting inter-species comparisons of the neural correlates of
526 these representations and operations. Our findings demonstrate that both species employ
527 similar overall strategies to perform the task (Fig. 3a-b). However, key differences in the
528 decision criteria of these strategies explain monkey performance deficits on the task.

529 The Wisconsin Card Sorting Test was originally developed to test cognitive flexibility,
530 i.e. the ability to rapidly adapt to a change in the task contingency, in the context of
531 abstract reasoning [8]. Early studies utilizing and developing scoring conventions for the
532 test [35] focused on perseverative errors. However, it has since become clear that the test
533 does not engage just a single cognitive process for task set switching; rather, it relies on
534 a variety of cognitive functions throughout the test including working memory, attention,
535 decision making, inhibitory control and reasoning [21, 36, 37, 38, 39]. For example, the
536 “failure to maintain set” error occurs when a subject applies an incorrect rule after they
537 have learned the correct rule and is associated with an error in working memory.

538 This has inspired systematic studies on WCST performance with two related goals.
539 First, research has focused on an accurate characterization of rule-learning strategies
540 and/or the cognitive processes that support their underlying computations [17, 20, 21].

541 Here we developed a relatively hypothesis-free approach to identify the rule-learning
542 strategy in humans and monkeys based on hidden behavioral states. The best-fit models
543 for both species ascribe these hidden states to varying levels of attention to individual
544 task-relevant visual features (Fig. 3a). These results are consistent with the conclusions of
545 earlier studies that humans contend with the “curse-of-dimensionality” which is inherent
546 in the WCST with selective attention towards individual features during exploration
547 [17, 20, 21]. Our findings clarify these results by showing that in the high-dimensional
548 version of the task (twelve instead of three possible rules) both humans and monkeys
549 must further contend with a tradeoff between computational complexity and information
550 efficiency while exploring for the rule, and they do so by selectively attending to a few,
551 but not all, features at a time (Fig. 4e).

552 Our approach differs from these earlier studies in that it does not postulate a specific
553 learning algorithm [17, 20]. Rather, it discovers the decision process that determines
554 the rule-learning strategy. In doing so, it illustrates important differences between
555 human/monkey rule learning strategies on the WCST and the commonly observed win-
556 stay lose-shift learning strategy (Fig. 3b). For example, a key function of the preferred
557 state, which is not part of the win-stay lose-shift strategy, is to support the simultaneous
558 exploration of multiple features at a time over many trials. Moreover, this state is also
559 associated with inference-like computations that support a computationally efficient
560 strategy of narrowing down the rule by eliminating other candidates using unambiguous
561 negative feedback. Indeed, a reinforcement-learning inspired description of rule learning
562 in the WCST by Bishara et al. [20] was parameterized to update the value of unchosen
563 options facilitating inference-like computations. However, the prevalence of the behavior
564 itself was not reported by the authors.

565 The second goal, with stronger clinical implications, is a comparison of error types in
566 healthy and diseased populations towards identifying more accurate behavioral markers
567 for different types of neuropsychiatric disorders and for dysfunction or lesions of different
568 brain regions [12, 19, 20, 21, 38, 39, 40, 41, 42, 43]. For example, a detailed analysis
569 of non-perseverative errors led to the differentiation between “efficient” errors that are
570 expected to occur during hypothesis testing from “random” errors that reflect a failure in

571 maintaining the cognitive set, and demonstrated a prevalence of the latter in patients with
572 frontal lobe pathology [19]. In support of this goal, we have developed a learning-stage
573 categorization method that delineates learning stages by the features under exploration
574 and their relationship to the rule (Fig. 5a). Intuitively, this approach tracks how far
575 along a subject is from learning the rule and reflects this in the reward rates across
576 categories (Fig. 5c, bottom). Furthermore, the categories are mutually exclusive and
577 exhaustive, allowing us to precisely ascribe differences in learning performance between
578 subjects or even between rule-blocks for the same subject to differences in individual
579 categories (Fig. 5a; Supplementary Fig. 7).

580 Importantly, the approach identifies various known classes and sub-classes of error
581 types, but also newer ones that may prove useful in future investigations of behavioral
582 markers for neuropsychiatric disorder and cognitive impairment. It distinguishes perse-
583 verative errors (made during the perseveration category) from non-perseverative ones.
584 Consistent with earlier work [19], it further sub-categorizes the latter into random and
585 efficient errors. It identifies two forms of random errors: one occurs during rule search
586 (before the rule preferred or exploitation categories) when subjects occasionally choose
587 none of the features they are currently exploring (Fig. 4d); the other occurs after they
588 have found the rule and while they are demonstrating this (Fig. 6b). However, it remains
589 unclear if either of these random errors are a feature of cognitive flexibility and result
590 from random exploration, or a bug caused by the failure to maintain the attention set
591 in working memory. Indirect evidence in humans has been found in favor of the latter
592 interpretation [44]. If in fact it is a result of higher distractability in monkeys, monkey
593 performance may be improved by imposing stronger controls on potential environmental
594 distractors [45]. Efficient errors arise instead when subjects test hypotheses regarding
595 rule identity and occur during the random search and non-rule exploration categories.
596 However, most errors during the rule-favored exploration category are repeated despite
597 unambiguous (direct or indirect) negative feedback (Fig. 7). These “disambiguation”
598 errors are neither random nor efficient but arise from a deficit in disambiguating the rule
599 feature that is under exploration from a simultaneously explored non-rule feature. This
600 newly identified error type bears further exploration in patient populations.

601 The higher incidence of these error types in monkeys may have more to do with
602 how they learn rather than some fundamental cognitive constrains — since they cannot
603 receive a rich verbal description of the task’s structure as humans do and must learn
604 about it via trial-and-error, it is possible that monkeys misinterpret uncued rule switches
605 as stochasticity in the environment resulting in a maladaptive strategy. Nevethless, it
606 is noteworthy that many of the errors we have identified as contributing to deficits in
607 monkey rule-learning performance have also been implicated in the poor performance
608 of humans with cognitive impairment. A higher incidence of perseverative errors in
609 patients with prefrontal cortex pathology was first reported in a landmark study by
610 Milner [10]. Random errors, while rare in healthy humans, are more frequently observed
611 in patients with frontal lobe dysfunction [19]), similar to the present study finding in
612 monkeys. Moreover, poor sensitivity to negative feedback, which underlies perseverative
613 and disambiguation errors in monkeys, is more pervasive in patients with schizophrenia
614 and substance abuse as well [20, 21].

615 Ultimately, these questions can be resolved by inter-species comparisons of neural data.
616 For example, perseverative and random errors have distinct neural signatures in humans
617 [46]; are similar signatures observed in monkeys? More generally, our work produces
618 several testable neural hypotheses in both species. First, does the neural representation
619 of the current rule persist during and across rule exploitation trials (i.e. after its identity
620 has been learned) [47, 48, 49, 50]? Second, since the set of explored features must be
621 maintained across several trials, are they represented by neural activity during and across
622 trials? Our model indicates that this attention set is typically small (Fig. 4e) and longer
623 bouts of exploring multiple features simultaneously (Fig. 7b) requires a choice alternation
624 between these features. This drives the need to maintain the explored features in working
625 memory, particularly to support recall of one of them after it is not chosen on one or
626 more previous trials. Third, are the distinct error types (perseverative, random, efficient,
627 disambiguation) differentially represented in the brain? Error coding neurons have been
628 reported in the prefrontal cortex of monkeys performing a WCST analogue [50, 51].
629 Moreover, perseverative, random and disambiguation errors signal the need to disengage
630 from the previous rule, address a working memory error and remove a non-rule feature

631 from the attention set, respectively. This difference in their function raises the possibility
632 that they are represented differently, either eliciting stronger responses in different brain
633 regions or eliciting differential responses in the same region [46]. Fourth, is the strength
634 of these error signals or their modulation of the attention set representation [50] larger
635 on trials when they serve their function? For example, are perseverative error signals
636 stronger on trials after which perservation halts compared to those that are followed
637 by continued perseveration? A potential reason for inter-species performance differences
638 may be found in these analyses: what inter-species neurocognitive differences explain
639 the relative prevalence of perseverative, random and disambiguation errors in monkeys
640 compared to humans, and are they also observed in humans with cognitive impairment?

641 There exist several avenues to clarify and improve upon our modeling approach. A
642 key difference between earlier models and ours is our assumption that each feature is
643 associated with discrete states, which our model relates to feature-based attentional
644 states. In contrast, Bayesian and reinforcement learning approaches posit that subjects
645 reason about features by assigning continuous-valued functions such as belief [17] and
646 value [18, 20] to them, respectively. In future work, we will test whether a model with
647 continuous-value states provide a better fit to the behavior of the two species, which
648 may offer a different interpretation of the latent variable used by them to reason about
649 features. Our model has also been simplified to keep subsequent analysis tractable — it
650 does not explicitly account for interactions between features. This had the unintended
651 consequence of discovering the “phantom” avoid state. Future improvements to our model
652 will carefully incorporate such interactions explicitly, while retaining high interpretability.

653 In conclusion, we have applied a hypothesis-free state-characterization method to
654 identify and compare the rule-learning strategy on the Wisconsin Card Sorting Test in
655 humans and monkeys. The hidden attentional states and state transitions inferred by the
656 model facilitated the determination of the decision process underlying this strategy as well
657 as the various stages of rapid rule learning. The inferred states perfectly (substantively)
658 explain human (monkey) choice behavior (Fig. 2c). Our overall approach reveals
659 differences in cognitive strategy between the two species and isolates the identity and
660 relative contribution of various error types to the performance difference between the

661 two species. It shows that random exploration or distraction and poorer sensitivity
662 to negative feedback underlies a higher incidence of these error types in monkeys thus
663 leading to their under-performance. The high fidelity demonstrated by the model in
664 inferring hidden attentional and decision states holds promise in advancing the search
665 for more accurate behavioral markers of various types of cognitive dysfunction and in
666 motivating targeted analyses to determine and compare the neural correlates of the
667 various cognitive processes engaged by the WCST.

668 **Acknowledgements:** We thank S.W. Linderman for generously sharing his code and
669 advice on fitting HMM-GLM models to data; I.R. Stone, J. Ferre and E.Y. Walker for
670 fruitful discussions. This work was supported by the National Institute of Health U-19
671 program grant no. 5U19NS107609-03, R01 grant no. R01MH062349 and the Office of
672 Naval Research grant no. N00014-17-1-2041.

673 **Data and Code Availability:** All training and analysis codes will be available at
674 publication on GitHub (<https://github.com/xjwanglab>). We will also provide data files
675 in Python readable formats for further analyses. Pre-trained models will be stored in a
676 Google Drive folder with its link provided on the same GitHub repository.

677 References

- 678 [1] Jonathan Birch, Alexandra K Schnell, and Nicola S Clayton. “Dimensions of animal
679 consciousness”. In: *Trends in cognitive sciences* 24.10 (2020), pp. 789–801.
- 680 [2] Lucia Melloni et al. “Computation and its neural implementation in human cogni-
681 tion”. In: *Strüngmann Forum Reports*. Vol. 27. 2019, pp. 323–46.
- 682 [3] Joaquin Fuster. *The prefrontal cortex*. Academic press, 2015.
- 683 [4] Richard E Passingham and Jeroen B Smaers. “Is the prefrontal cortex especially
684 enlarged in the human brain? Allometric relations and remapping factors”. In:
685 *Brain, behavior and evolution* 84.2 (2014), pp. 156–166.
- 686 [5] Chad J Donahue et al. “Quantitative assessment of prefrontal cortex in humans
687 relative to nonhuman primates”. In: *Proceedings of the National Academy of
688 Sciences* 115.22 (2018), E5183–E5192.
- 689 [6] Terrence W Deacon. “What makes the human brain different?” In: *Annual Review
690 of Anthropology* (1997), pp. 337–357.
- 691 [7] Suzana Herculano-Houzel. “The human brain in numbers: a linearly scaled-up
692 primate brain”. In: *Frontiers in human neuroscience* (2009), p. 31.
- 693 [8] David A Grant and Esta Berg. “A behavioral analysis of degree of reinforcement
694 and ease of shifting to new responses in a Weigl-type card-sorting problem.” In:
695 *Journal of experimental psychology* 38.4 (1948), p. 404.
- 696 [9] Bruno Kopp, Florian Lange, and Alexander Steinke. “The reliability of the Wis-
697 consin card sorting test in clinical practice”. In: *Assessment* 28.1 (2021), pp. 248–
698 263.
- 699 [10] Brenda Milner. “Effects of different brain lesions on card sorting: The role of the
700 frontal lobes”. In: *Archives of neurology* 9.1 (1963), pp. 90–100.
- 701 [11] RE Passingham. “Non-reversal shifts after selective prefrontal ablations in monkeys
702 (*Macaca mulatta*)”. In: *Neuropsychologia* 10.1 (1972), pp. 41–46.
- 703 [12] EA Drewe. “The effect of type and area of brain lesion on Wisconsin Card Sorting
704 Test performance”. In: *Cortex* 10.2 (1974), pp. 159–170.

- 705 [13] Hazel E Nelson. “A modified card sorting test sensitive to frontal lobe defects”. In:
706 *Cortex* 12.4 (1976), pp. 313–324.
- 707 [14] James M Gold et al. “Auditory working memory and Wisconsin Card Sorting
708 Test performance in schizophrenia”. In: *Archives of general psychiatry* 54.2 (1997),
709 pp. 159–165.
- 710 [15] Farshad Alizadeh Mansouri, David J Freedman, and Mark J Buckley. “Emergence
711 of abstract rules in the primate brain”. In: *Nature Reviews Neuroscience* 21.11
712 (2020), pp. 595–610.
- 713 [16] Samuel J Gershman and Yael Niv. “Learning latent structure: carving nature at
714 its joints”. In: *Current opinion in neurobiology* 20.2 (2010), pp. 251–256.
- 715 [17] Robert C Wilson and Yael Niv. “Inferring relevance in a changing world”. In:
716 *Frontiers in human neuroscience* 5 (2012), p. 189.
- 717 [18] Yael Niv et al. “Reinforcement learning in multidimensional environments relies on
718 attention mechanisms”. In: *Journal of Neuroscience* 35.21 (2015), pp. 8145–8157.
- 719 [19] Francisco Barceló and Robert T Knight. “Both random and perseverative errors
720 underlie WCST deficits in prefrontal patients”. In: *Neuropsychologia* 40.3 (2002),
721 pp. 349–356.
- 722 [20] Anthony J Bishara et al. “Sequential learning models for the Wisconsin card
723 sort task: Assessing processes in substance dependent individuals”. In: *Journal of*
724 *mathematical psychology* 54.1 (2010), pp. 5–13.
- 725 [21] Jan Gläscher, Ralph Adolphs, and Daniel Tranel. “Model-based lesion mapping of
726 cognitive control using the Wisconsin Card Sorting Test”. In: *Nature communica-*
727 *tions* 10.1 (2019), pp. 1–12.
- 728 [22] Adam J Calhoun, Jonathan W Pillow, and Mala Murthy. “Unsupervised identifica-
729 tion of the internal states that shape natural behavior”. In: *Nature neuroscience*
730 22.12 (2019), pp. 2040–2049.
- 731 [23] Nicholas A Roy et al. “Extracting the dynamics of behavior in sensory decision-
732 making experiments”. In: *Neuron* 109.4 (2021), pp. 597–610.

- 733 [24] Scott S Bolkan et al. “Opponent control of behavior by dorsomedial striatal
734 pathways depends on task demands and internal state”. In: *Nature neuroscience*
735 25.3 (2022), pp. 345–357.
- 736 [25] Yoshua Bengio and Paolo Frasconi. “An input output HMM architecture”. In:
737 *Advances in neural information processing systems* 7 (1994).
- 738 [26] Timothy EJ Behrens et al. “Learning the value of information in an uncertain
739 world”. In: *Nature neuroscience* 10.9 (2007), pp. 1214–1221.
- 740 [27] Thomas T Hills et al. “Exploration versus exploitation in space, mind, and society”.
741 In: *Trends in cognitive sciences* 19.1 (2015), pp. 46–54.
- 742 [28] Samuel J Gershman. “Deconstructing the human algorithms for exploration”. In:
743 *Cognition* 173 (2018), pp. 34–42.
- 744 [29] Robert C Wilson et al. “Balancing exploration and exploitation with information
745 and randomization”. In: *Current opinion in behavioral sciences* 38 (2021), pp. 49–
746 56.
- 747 [30] Jon Driver. “A selective review of selective attention research from the past century”.
748 In: *British journal of psychology* 92.1 (2001), pp. 53–78.
- 749 [31] Maurizio Corbetta and Gordon L Shulman. “Control of goal-directed and stimulus-
750 driven attention in the brain”. In: *Nature reviews neuroscience* 3.3 (2002), pp. 201–
751 215.
- 752 [32] James Everett et al. “Performance of patients with schizophrenia on the Wisconsin
753 Card Sorting Test (WCST).” In: *Journal of Psychiatry and Neuroscience* 26.2
754 (2001), p. 123.
- 755 [33] Edith V Sullivan et al. “Factors of the Wisconsin Card Sorting Test as measures of
756 frontal-lobe function in schizophrenia and in chronic alcoholism”. In: *Psychiatry*
757 *research* 46.2 (1993), pp. 175–199.
- 758 [34] Sally Ozonoff and Robin E McEvoy. “A longitudinal study of executive function
759 and theory of mind development in autism”. In: *Development and psychopathology*
760 6.3 (1994), pp. 415–431.

- 761 [35] Robert K Heaton. “Wisconsin card sorting test manual”. In: *Psychological assess-*
762 *ment resources* (1981).
- 763 [36] Stanislas Dehaene and Jean-Pierre Changeux. “The Wisconsin Card Sorting Test:
764 Theoretical analysis and modeling in a neuronal network”. In: *Cerebral cortex* 1.1
765 (1991), pp. 62–79.
- 766 [37] Francisco Barceló. “Does the Wisconsin card sorting test measure prefrontal func-
767 tion?” In: *The Spanish journal of psychology* 4.1 (2001), pp. 79–100.
- 768 [38] Chuh-Hyoun Lie et al. “Using fMRI to decompose the neural processes underlying
769 the Wisconsin Card Sorting Test”. In: *Neuroimage* 30.3 (2006), pp. 1038–1049.
- 770 [39] Bradley R Buchsbaum et al. “Meta-analysis of neuroimaging studies of the Wiscon-
771 sin Card-Sorting task and component processes”. In: *Human brain mapping* 25.1
772 (2005), pp. 35–45.
- 773 [40] Adrian M Owen et al. “Extra-dimensional versus intra-dimensional set shifting
774 performance following frontal lobe excisions, temporal lobe excisions or amygdalo-
775 hippocampectomy in man”. In: *Neuropsychologia* 29.10 (1991), pp. 993–1006.
- 776 [41] Y Nagahama et al. “The cerebral correlates of different types of perseveration
777 in the Wisconsin Card Sorting Test”. In: *Journal of Neurology, Neurosurgery &*
778 *Psychiatry* 76.2 (2005), pp. 169–175.
- 779 [42] Trevor William Robbins. “Dissociating executive functions of the prefrontal cortex”.
780 In: *Philosophical Transactions of the Royal Society of London. Series B: Biological*
781 *Sciences* 351.1346 (1996), pp. 1463–1471.
- 782 [43] Mark J Buckley et al. “Dissociable components of rule-guided behavior depend on
783 distinct medial and prefrontal regions”. In: *Science* 325.5936 (2009), pp. 52–58.
- 784 [44] Ivonne J Figueroa and Robert J Youmans. “Failure to maintain set: A measure of
785 distractibility or cognitive flexibility?” In: *Proceedings of the human factors and*
786 *ergonomics society annual meeting*. Vol. 57. 1. 2013, pp. 828–832.
- 787 [45] Robert B Malmo. “Interference factors in delayed response in monkeys after removal
788 of frontal lobes”. In: *Journal of Neurophysiology* 5.4 (1942), pp. 295–308.

- 789 [46] Francisco Barceló. “Electrophysiological evidence of two different types of error in
790 the Wisconsin Card Sorting Test”. In: *Neuroreport* 10.6 (1999), pp. 1299–1303.
- 791 [47] Juri Minxha et al. “Flexible recruitment of memory-based choice representations
792 by the human medial frontal cortex”. In: *Science* 368.6498 (2020), eaba3313.
- 793 [48] Silvia Bernardi et al. “The geometry of abstraction in the hippocampus and
794 prefrontal cortex”. In: *Cell* 183.4 (2020), pp. 954–967.
- 795 [49] Jonathan D Wallis, Kathleen C Anderson, and Earl K Miller. “Single neurons in
796 prefrontal cortex encode abstract rules”. In: *Nature* 411.6840 (2001), pp. 953–956.
- 797 [50] Farshad A Mansouri, Kenji Matsumoto, and Keiji Tanaka. “Prefrontal cell activities
798 related to monkeys’ success and failure in adapting to rule changes in a Wisconsin
799 Card Sorting Test analog”. In: *Journal of Neuroscience* 26.10 (2006), pp. 2745–2756.
- 800 [51] Masaru Kuwabara et al. “Cognitive control functions of anterior cingulate cortex in
801 macaque monkeys performing a Wisconsin Card Sorting Test analog”. In: *Journal*
802 *of Neuroscience* 34.22 (2014), pp. 7531–7547.

803 **Methods**

804 **Task Description**

805 Human and monkey subjects were tested on an analogue of the Wisconsin Card Sorting
806 Test (WCST). On each trial, they were simultaneously presented with an array of four
807 objects on a computer screen. Each object was comprised of a stimulus feature from
808 each of three stimulus dimensions: color, pattern, and shape (Fig. 1a), e.g., a blue
809 polka-dotted triangle. For each trial, these four objects were chosen from a pool of 64
810 unique objects, each containing a possible combination of individual features from each of
811 the three dimensions, such that there was no feature overlap between them. Accordingly,
812 for each possible array, all four features of each dimension appeared on the screen, but
813 the combination of features represented by each individual object varied across trials.
814 Within a single rule-learning block of trials, one color, texture, or shape was designated as
815 the target, resulting in 12 possible rules. The identity of this rule was not cued, but had
816 to be learned by trial and error, based on the feedback received at the end of each trial.
817 Upon meeting a rule-learning criteria for the current rule, the rule feature changed on the
818 next trial in an uncued manner, initiating a new rule-learning block. This rule shift could
819 be either intradimensional, where the dimension of the new rule feature matched that of
820 the previous rule feature (e.g., changing from triangle to square), or extradimensional,
821 where the dimensions of the old and new rule features did not match (e.g., changing from
822 triangle to yellow);

823 **Monkeys**

824 All procedures were carried out in accordance with the National Institutes of Health
825 guidelines and were approved by the University of Washington Institutional Animal
826 Care and Use Committee. Subjects were four adult female rhesus monkeys (*Macaca*
827 *mulatta*) with mean age 12.5 ± 2.5 years and mean weight 7.5 ± 0.6 kg at the start of the
828 experiment. Prior to testing, a titanium post for holding the head was surgically affixed

829 to each monkey. During testing, each monkey was head-fixed in a dimly illuminated
830 room and positioned 60 cm away from a 19-inch CRT monitor with a screen refresh rate
831 of 120 Hz noninterlaced. The monitor had a resolution of 800×600 pixels, subtending
832 33 degrees by 25 degrees of visual angle (dva). Eye movements were recorded using a
833 noninvasive infrared eye-tracking system (EyeLink 1000 Plus, SR Research). Stimuli were
834 presented using experimental control software (NIMH Cortex or NIMH MonkeyLogic).
835 Calibration of the infrared eye tracking system was accomplished using a nine-point
836 manual calibration task.

837 Following the calibration task, the monkey was tested on the WCST analogue. The
838 monkey initiated each trial by fixating a white cross (0.5°) at the center of the computer
839 screen. Following 500 ms of successful fixation, the cross disappeared and was replaced by
840 an array of four objects. During the self-paced decision epoch that followed, the monkey
841 was free to explore the array of objects; her response was defined as maintaining her gaze
842 within a $9^\circ \times 9^\circ$ window centered on the object for 800 ms. The monkey received a food
843 slurry reward over a 1.4-second duration for selecting the object that contained the rule
844 feature. A time-out period (either 1-second or 5-seconds) occurred on trials where the
845 monkey did not choose the object containing the rule feature or where she did not make
846 a choice within 4 seconds. The feedback period was immediately followed by a 400 ms or
847 1 s inter-trial interval. We classified a rule as learned either when the monkey made eight
848 consecutive correct responses or when she made 16 correct responses in 20 trials or fewer.

849 The type of rule shift that followed (intradimensional or extradimensional) was
850 determined pseudo-randomly to occur with equal probability. A block consisted of all the
851 trials from the initial rule shift to the final trial of criterion performance. We analyzed a
852 total of 1305 blocks in 81 recording sessions from monkey B, 872 blocks in 29 recording
853 sessions from monkey C, 805 blocks in 29 recording sessions from monkey S, and 224
854 blocks in 13 recording sessions from monkey T. Only completed blocks were included in
855 the analysis.

856 **Humans**

857 The studies involving human participants were reviewed and approved by the Institutional
858 Review Boards of University of California, Berkeley. All participants provided their
859 written informed consent to participate in this study and received a small compensation.
860 These subjects were four adult males and one adult female with mean age 26.4 ± 4.1 years.
861 Subjects were brought into a room where they sat and completed a computer-adapted
862 version of WCST analogue on a recording laptop after receiving the following instructions:
863 “In this experiment, you will see 4 cards on each trial. Each card has 3 unique features
864 (color, shape and texture). No feature is shown on more than one card, so you will see 12
865 different features on each trial (4 colors, 4 shapes, 4 textures). The card containing the
866 correct feature (1 out of 12 possible) will be correct choice. The correct feature might
867 change during the task. The answer is given by pressing one of the four arrow keys that
868 corresponds with the selected card position on the screen (up, down, left or right). You
869 have 4 sec to provide the answer, or the trial times out. The task goes on for 200 trials
870 or about 15 minutes.”

871 Individual trials consisted of the following epochs: cross fixation (black cross displayed
872 in the center of the screen on a gray background for 300 ms), choice (four objects displayed
873 on the screen at locations corresponding to up, down, left or right positions, for up to
874 4000 ms), feedback (‘correct’ or ‘incorrect’ feedback message displayed for 1500 ms) and
875 inter-trial interval (ITI, gray screen for 1000 ms). Subjects indicated their choice by
876 pressing the arrow key on the laptop keyboard, corresponding to the chosen object’s
877 position on the screen. If the choice was not indicated within the 4000 ms, the trial
878 was considered timed-out. After reaching the learning criteria, defined as 5 consecutive
879 correct trials or 8 correct out of the last 10 trials, the rule was switched and a new
880 rule-learning block began. The new rule was randomly determined. Each participant
881 completed five task sessions (300 trials/sessions for a total of 1500 trials). This spanned
882 between 107 and 138 blocks across the five subjects.

883 **Win-Stay Lose-Shift (WSLS) Agent**

884 The task structure (rule selection, learning criteria) for the WSLS agent was identical to
885 that of the humans, except for the trial structure – the agent’s algorithm determined its
886 choice immediately upon stimulus presentation. The algorithm (Supplementary Fig. 3d,
887 left) always maintained a single feature in the persist state and deterministically chose
888 the object with that feature at each trial. Positive feedback maintained the feature in the
889 persist state. Negative feedback demoted it to the avoid state and promoted, a randomly
890 selected feature from among the 11 others that were in the avoid state, to the persist
891 state. The agent completed 500 rule blocks.

892 **Input-Output Hidden Markov Model - Generalized Linear Model** 893 **(IOHMM-GLM) for the prediction of feature choices**

894 **Model Design**

895 The four objects presented during a trial consist of twelve visual features, $f \in \{1, \dots, 12\}$.
896 In support of feature-based mental representations, the model predicts the choice of
897 each feature f at the next trial t . This choice is represented by $c_t^f \in \{0, 1\}$, where
898 $c_t^f = 1$ indicates the f was part of the chosen object, and $c_t^f = 0$ indicates it was not.
899 Either choice can result in a reward or timeout for the trial, given by $r_t \in \{0, 1\}$. The
900 choice-outcome history of f given the past ℓ trials is denoted $h^f \in \{1, \dots, 2^{2\ell}\}$. We refer
901 to ℓ as the *lag* and it is a hyperparameter of the model. The value of h^f at trial t is
902 given by the binary vector $(r_{t-1}, c_{t-1}, \dots, r_{t-\ell}, c_{t-\ell})$ of size 2ℓ . Therefore, it can take
903 on $2^{2\ell}$ possible values. In all our analyses, we choose a lag 1 ($\ell = 1$) model for further
904 analysis. Such a model depends on a choice-outcome history that takes on one of four
905 possible values at trial t : $(r_{t-1} = 0, c_{t-1} = 0)$, $(r_{t-1} = 1, c_{t-1} = 0)$, $(r_{t-1} = 0, c_{t-1} = 1)$ or
906 $(r_{t-1} = 1, c_{t-1} = 1)$ which we refer to as NC^- , NC^+ , C^- and C^+ respectively.

907 The transformation of the choice-outcome history into a choice at trial t is mediated by

908 discrete hidden states $s^f \in \{1, \dots, K\}$ that determine the parameters of the transformation.
909 The maximum number of states K is a second model hyperparameter. The transformation
910 is modeled as a Bernoulli GLM:

$$p(c_t = 1 \mid s_t = k, h_t) = \frac{1}{1 + \exp(-w_k^T h_t)} \quad (1)$$

911 where the parameters $w_k \in \mathbb{R}^{1 \times 2^{2\ell}}$ are determined by the state $s_t = k$. We denote the
912 set of parameters across all K states as $w \in \mathbb{R}^{K \times 2^{2\ell}}$.

913 Transitions between states also depend on the choice-outcome history and are modeled
914 by multinomial logistic regression:

$$p(s_{t+1} = k \mid s_t = j, h_{t+1}) = \frac{\exp(\log(P_{jk}) + u_{jk}^T h_{t+1})}{\sum_{k'=1}^K \exp(\log(P_{jk'}) + u_{jk'}^T h_{t+1})} \quad (2)$$

915 where the parameters $P \in \mathbb{R}_+^{K \times K}$ and $u \in \mathbb{R}^{K \times K \times 2^{2\ell}}$ represent the bias or baseline
916 transition probability and history weights. This model design is schematized in Figure
917 2a.

918 Finally, the probability distribution of initial states π , is a model parameter that
919 specifies the state at the first trial of a session.

920 Model Fitting

We fit the parameter values for the choice GLM weights w , the baseline transition probability P , the transition GLM weights u and the initial state distribution π to the choices of each subject. To avoid over-fitting, the parameter values were shared across all features. In other words, all parameter values were the same for all 12 features. The likelihood of the data under a model is its probability subject to the model's parameters and inputs $p(c_{1..T} \mid w, P, u, \pi, h_{1..T})$, where T is the number of trials in the session. It is expressed in terms of these parameters as:

$$p(c_{1..T} \mid w, P, u, \pi, h_{1..T}) = \sum_{s_{1..T}} p(c_{1..T}, s_{1..T} \mid w, P, u, \pi, h_{1..T})$$

$$= \sum_{s_{1..T}} p(s_1 | \pi) \left[\prod_{t=2}^T p(s_t | P, u, h_t) \right] \left[\prod_{t=1}^T p(c_t | w, s_t, h_t) \right]$$

921 where the last two terms are given by equations 2 and 1 respectively.

922 The model parameters were fit by minimizing $-\log [p(c_{1..T} | w, P, u, \pi, h_{1..T})]$, i.e. the
923 negative log-likelihood of the data, via gradient descent with the ADAM optimizer.
924 The choice GLM weights for all k states were initialized to a single $2^{2\ell}$ -dimensional
925 vector drawn from a standard normal distribution. The baseline transition probability
926 was initialized to the sum of a diagonal matrix with value $0.9I$ where I is the identity
927 matrix, and a random matrix with elements drawn from a uniform distribution in the
928 interval $[0, 0.05)$. The larger diagonal values enforce “stickiness” that bias transitions
929 back into a state. The transition GLM weights were initialized to zero, and the initial
930 state distribution was initialized to $1/K$ for each state k . For each subject and each pair
931 of hyperparameters (ℓ, K) , the parameters were optimized over 10000 iterations with
932 5-fold cross validation (Fig. 2b).

933 The best fit model was sought for each human (monkey) subject and hyperparameter
934 setting across 10 (5) independent parameter initializations. Figure 2b shows the mean
935 negative log-likelihood taken over all initializations and cross-validation folds. The best-fit
936 model for each human (monkey) subject was selected for further analysis from these
937 50 (25) models at hyperparameter values $\ell = 1$ and $K = 4$. Although, we found that
938 a majority of these models produced very similar choice and transition probabilities.
939 However, fits to the WSLs agent varied much more. Since negative feedback immediately
940 demoted features from the persist to avoid state, exploration of non-rule features typically
941 lasted 1-2 trials. This likely makes it harder for the model to identify exploration and
942 introduces more variability across fits.

943 Once the best-fit model is identified, the most likely sequence of states, s^* , for each
944 subject, session and feature is determined by the Viterbi algorithm [1] (Fig. 2d). For
945 each trial t and feature f , the algorithm performs a forward pass across all past trials
946 and a backward pass across all future trials to determine the most likely state of f at
947 trial t that best explains past, present and future history-dependent choices under the

948 constraints of the model’s parameters and the choice and transition probabilities they
949 yield. Supplementary Figure 2b) shows the cumulative distribution of the posterior
950 probabilities ($p(s_t = s_t^* | c_{1..T}, h_{1..T})$) of these state estimates calculated for the Viterbi
951 algorithm.

952 All model fits and the most-likely state determination was performed with the State
953 Space Model (SSM) python package [2].

954 Model Extension for the Prediction of Object Choices

955 We extended the feature choice prediction model described in the previous section to
956 predict object choices at each trial t . Given the predicted choice probability ($p(c_t^{f_{i,j}} |$
957 $w, P, u, \pi, h_{1..t}^{f_{i,j}})$) for each feature $f_{i,j}, i \in \{1, \dots, 3\}$ in an object $o_j, j \in \{1, \dots, 4\}$ pre-
958 sented at trial t , the model predicts the object chosen at t . This transformation of
959 predicted feature choice probabilities $p(f_{i,j})$ into object choice probabilities $p(o_j | p(f))$
960 is modeled by multinomial logistic regression:

$$p(o_j | p(f)) = \frac{\exp [\sum_{i=1}^3 v_{ij} \log(p(f_{i,j})) + b_j]}{\sum_{j'=1}^4 \exp [\sum_{i=1}^3 v_{ij'} \log(p(f_{i,j'})) + b_{j'}]} \quad (3)$$

961 where the parameters $v \in \mathbb{R}^{3 \times 4}$ and $b \in \mathbb{R}^{1 \times 4}$ represents the feature choice probability
962 weights and biases in selecting each object, respectively. These values were fit to the
963 choices of each subject by minimizing the cross-entropy loss $-\sum_{t=1}^T \sum_{j=1}^4 y_{j,t} \log(p(o_{j,t} |$
964 $p(f)_t)$) where $y_{j,t} \in \{0, 1\}$ indicates whether object $o_{j,t}$ was chosen on trial t . Model fitting
965 was performed via stochastic gradient descent with the ADAM optimizer implemented by
966 the Pytorch python package [3]. The parameter values for v and b were initialized from
967 a uniform distribution in the interval $[-\frac{1}{\sqrt{12}}, \frac{1}{\sqrt{12}}]$ and optimized until convergence with a
968 maximum of 100000 iterations. Cross validation was performed with the same training
969 and test sets used while training the feature choice prediction models (Supplementary
970 Fig. 2a).

971 The accuracy of the object choice prediction model based on the best-fit feature choice
972 prediction model with 4 states and lag 1 is shown in Figure 2c, left. We also fit a model

973 to determine the chosen object in a similar fashion using the feature choice probabilities
 974 based on their most-likely state estimates ($p(c_t^{f,i,j} | s_t^{f,i,j} = s_t^{*,f,i,j}, h_t^{f,i,j})$) instead. The
 975 accuracy of this model is shown in Figure 2c, right.

976 Model Analysis

977 The probability distribution of histories in each state (Supplementary Fig. 4b) is:

$$p(h = i | s^* = j) = \frac{\sum_{f,t} \mathbf{1}(h_t^f = i, s_t^{*,f} = j)}{\sum_{f,t} \mathbf{1}(s_t^{*,f} = j)} \quad (4)$$

978 where $\mathbf{1}$ is the indicator function and $\sum_{f,t}$ is a sum over features and trials. The
 979 state and history dependent choice probability (Supplementary Fig. 4a) can be directly
 980 calculated from the model's parameters (Eqn. 1) or empirically as:

$$p(c = 1 | s^* = j, h = i) = \frac{\sum_{f,t} \mathbf{1}(c_t^f = 1, s_t^{*,f} = j, h_t^f = i)}{\sum_{f,t} \mathbf{1}(s_t^{*,f} = j, h_t^f = i)} \quad (5)$$

981 The choice probability of a feature in each state (Fig. 3a) can be computed by utilizing
 982 equations (4) and (5) or 1.

$$p(c = 1 | s^* = j) = \sum_{i \in \{1, \dots, 4\}} p(c = 1 | s^* = j, h = i) \cdot p(h = i | s^* = j) \quad (6)$$

983

984 Similarly, the state transition probabilities (Supplementary Fig. 5) can be directly
 985 calculated from the model's parameters (Eqn. 2) or empirically as:

$$p(s_{t+1}^* = k | s_t^* = j, h_{t+1} = i) = \frac{\sum_{f,t} \mathbf{1}(s_{t+1}^{*,f} = k, s_t^{*,f} = j, h_{t+1}^f = i)}{\sum_{f,t} \mathbf{1}(s_t^{*,f} = j, h_{t+1}^f = i)} \quad (7)$$

986 We approximated the decision process in each species (Fig. 3b) from the state transition
 987 probability, and the “reverse” state transition probability ($p(s_t^* = j | s_{t+1}^* = k, h_{t+1} = i)$).

988 The latter helps in conditions where transitions into a state are typically rare, such
989 as transitions from the random/avoid state into the preferred state. This quantity
990 (Supplementary Fig. 6) is calculated empirically as:

$$p(s_t^* = j \mid s_{t+1}^* = k, h_{t+1}^f = i) = \frac{\sum_{f,t} \mathbb{1}(s_t^{*,f} = j, s_{t+1}^{*,f} = k, h_{t+1}^f = i)}{\sum_{f,t} \mathbb{1}(s_{t+1}^{*,f} = k, h_{t+1}^f = i)} \quad (8)$$

991

992 Trial Categorization

Trials were categorized based on the identity of the rule feature and the most-likely state estimates for all 12 features as in Figure. 5a. Since each trial is always designated to one and only one category, the trial categories are mutually-exclusive and exhaustive. This facilitates a precise decomposition of the length of each rule block into the number of trials spent in each category (Fig. 5c). Moreover, since the categories are mutually exclusive, we can explain summary statistics (mean and variance) of the block length for each subject in terms of statistics of their category lengths (Supplementary Fig. 7):

$$\begin{aligned} \mathbb{E}[\text{block length}] &= \sum_{\text{category } c} \mathbb{E}[\text{no. trials in category } c] \\ \text{Var}[\text{block length}] &= \sum_{\text{category } c} \text{cov}[\text{no. trials in category } c, \text{block length}] \end{aligned}$$

993 Inter-species Comparison of Category Lengths

994 In Figure 7, the higher probability of continued exploration of non-rule features by
995 monkeys during the rule-favored exploration category is attributed to poor (direct and
996 indirect) negative feedback sensitivity (Fig. 7c-d). In addition, we attribute the higher
997 probability of continued exploration in monkeys during rule-favored exploration trials
998 compared to non-rule exploration trials, to a higher prevalence of direct positive feedback
999 during rule-favored exploration trials (Fig. 7f).

These determinations were made based on the following decomposition:

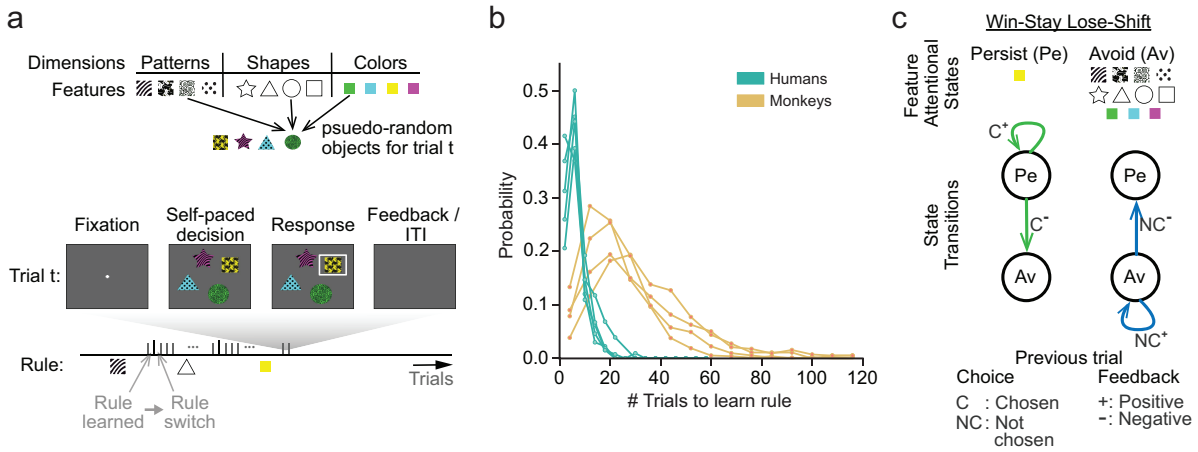
$$\begin{aligned} p(s_{t+1}^* \in \text{explore} \mid s_t^* \in \text{explore}, (t, t+1) \subseteq \text{category } c) \\ &= \sum_{i \in \{1, \dots, 4\}} p(s_{t+1}^* \in \text{explore}, h = i \mid s_t^* \in \text{explore}, (t, t+1) \subseteq \text{category } c) \\ &= \sum_{i \in \{1, \dots, 4\}} [p(s_{t+1}^* \in \text{explore} \mid h = i, s_t^* \in \text{explore}, (t, t+1) \subseteq \text{category } c) \\ &\quad \times p(h = i \mid s_t^* \in \text{explore}, (t, t+1) \subseteq \text{category } c)] \end{aligned}$$

1000 The joint probability above is shown in Figure 7f, left and in Supplementary Fig. 8), and
1001 quantities resulting from its decomposition below are shown in Figure 7f, middle-right.

1002 References

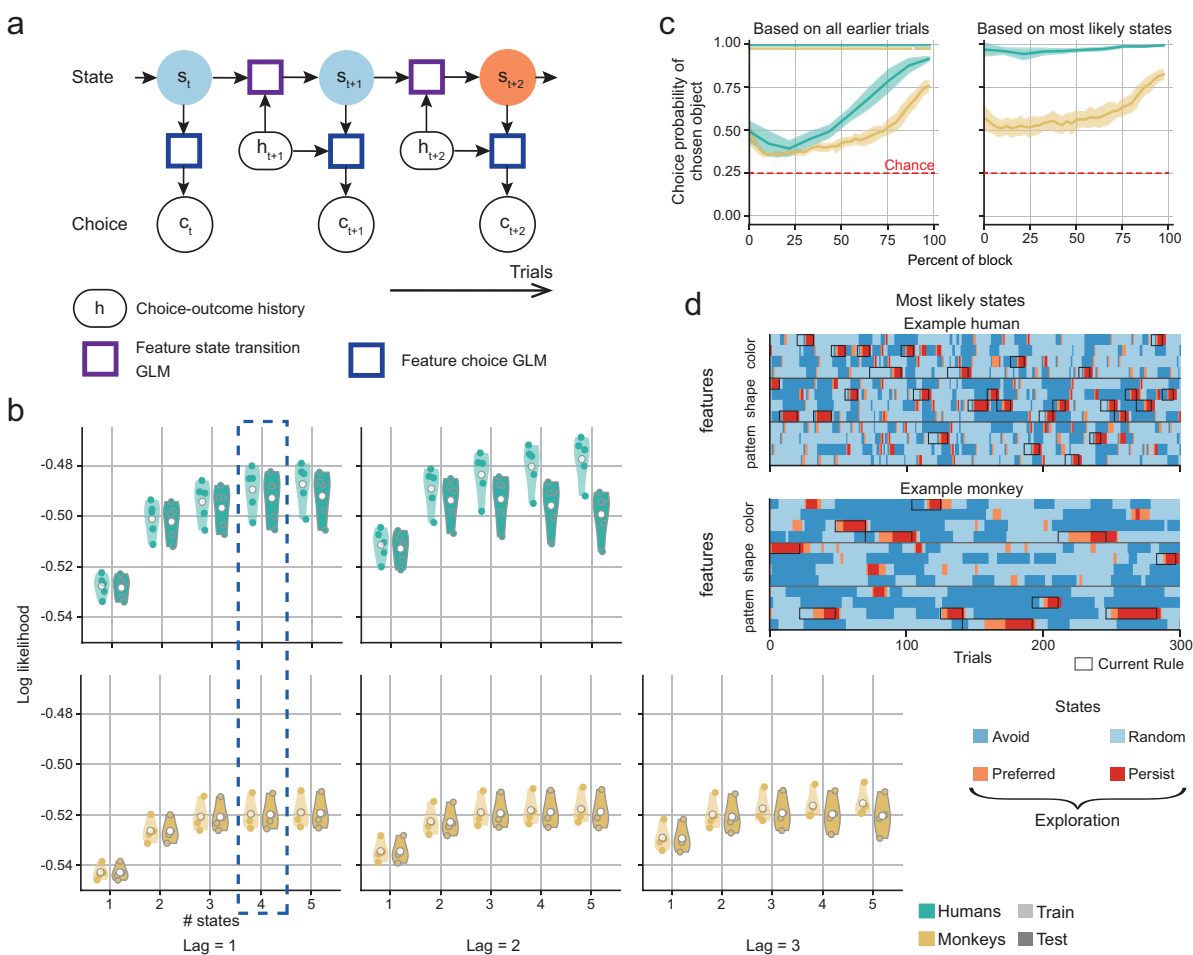
- 1003 [1] Andrew Viterbi. “Error bounds for convolutional codes and an asymptotically
1004 optimum decoding algorithm”. In: *IEEE transactions on Information Theory* 13.2
1005 (1967), pp. 260–269.
- 1006 [2] Scott Linderman et al. *SSM: Bayesian Learning and Inference for State Space*
1007 *Models (Version 0.0.1)*. <https://github.com/lindermanlab/ssm>. 2020.
- 1008 [3] Adam Paszke et al. “Pytorch: An imperative style, high-performance deep learning
1009 library”. In: *Advances in neural information processing systems* 32 (2019).

1010 **Figures**



1011

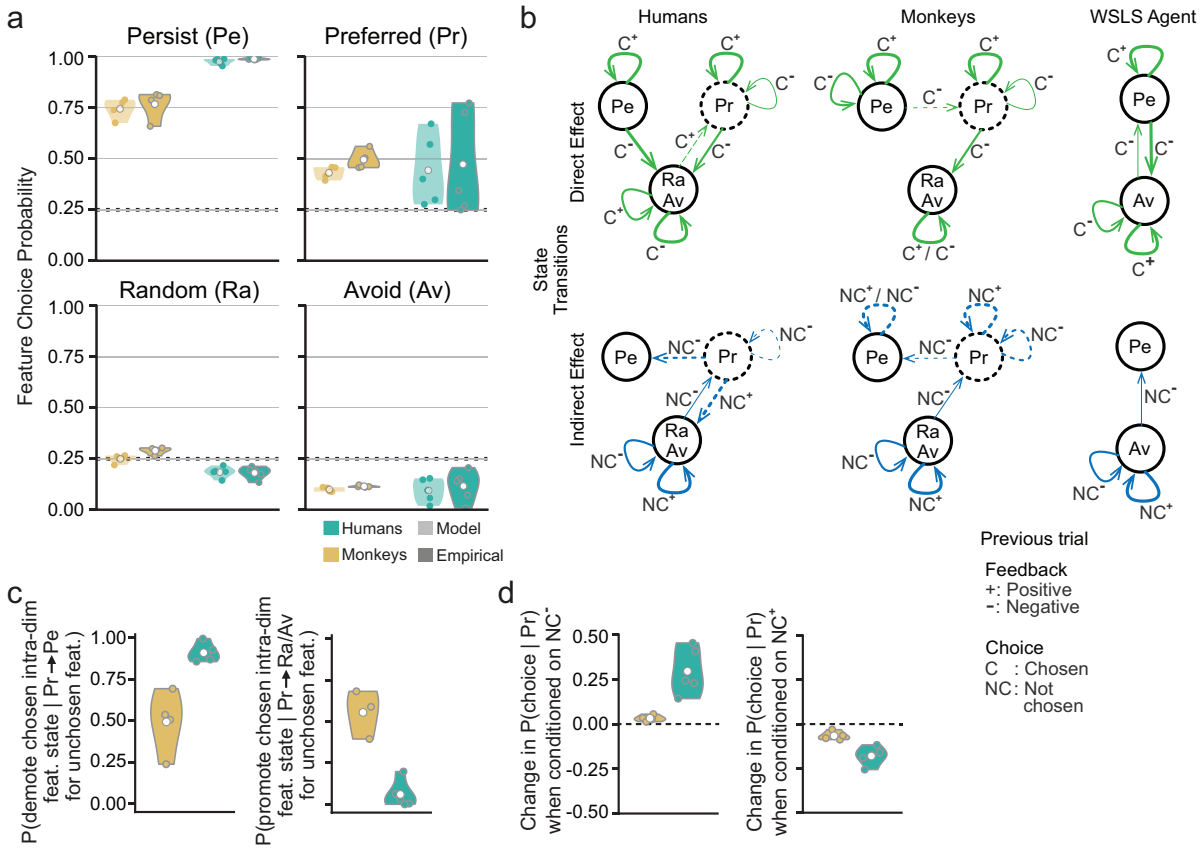
Figure 1: Monkeys rapidly learn rules in the WCST but are slower than humans. a. WCST task structure. Each trial is composed of fixation, decision, response, feedback and ITI epochs. After fixation, the subject is presented with 4 objects that are pseudo-randomly composed of 3 features - a pattern, shape and color. The features composing each object are mutually exclusive with respect to other objects. Each block of continuous trials is governed by a rule (one of the 12 features). The subject receives positive feedback only for choosing the object with that feature. The identity of the rule is hidden and must be discovered. An uncued rule switch to a random new feature occurs when the subject demonstrates they have learned the current rule. **b.** Distribution of trials-to-learning-criteria in 4 monkey (brown) and 5 human (green) subjects. All subjects rapidly learn the rule, but on average, monkeys are over 4 times slower than humans. **c.** Decision process for the Win-Stay Lose-Shift learning strategy in two-armed bandit problems. The decision to choose an arm can be in one of 2 states: persist when it is chosen and avoid when it is not. The decision to choose an arm stays in the persist state as long as positive feedback is received (win-stay) and switches to the avoid state otherwise. It then stays in the avoid state as long as positive feedback is received, and switches to the persist state when negative feedback is received (lose-shift).



1012

Figure 2: IOHMM-GLM model fits uncover dynamic changes in choice behavior during rule learning. *a.* IOHMM-GLM model architecture fit to data. The model predicts the choice of a feature c at each trial t from the choice-outcome trial history h via a GLM. Hidden states s determine the GLM's parameters. These states can transition at each trial also based on the choice outcome trial history via a separate state transition GLM. *b.* Model fit log-likelihoods on training and test datasets for each human (green) and monkey (brown) subject in models with varying numbers of states and that use choice outcomes from varying numbers of previous trials (lag) to determine the feature choice and state-transition probabilities. Each point represents a single subject's mean over a 5-fold cross-validation and over 5 (monkey) or 10 (human) different model initializations. Each subject's best-fit model with 4 states and lag 1 (dashed blue box) was chosen for further analysis. *c.* Probability of selecting the chosen object produced by a model extension based on feature choice probabilities predicted only from choice

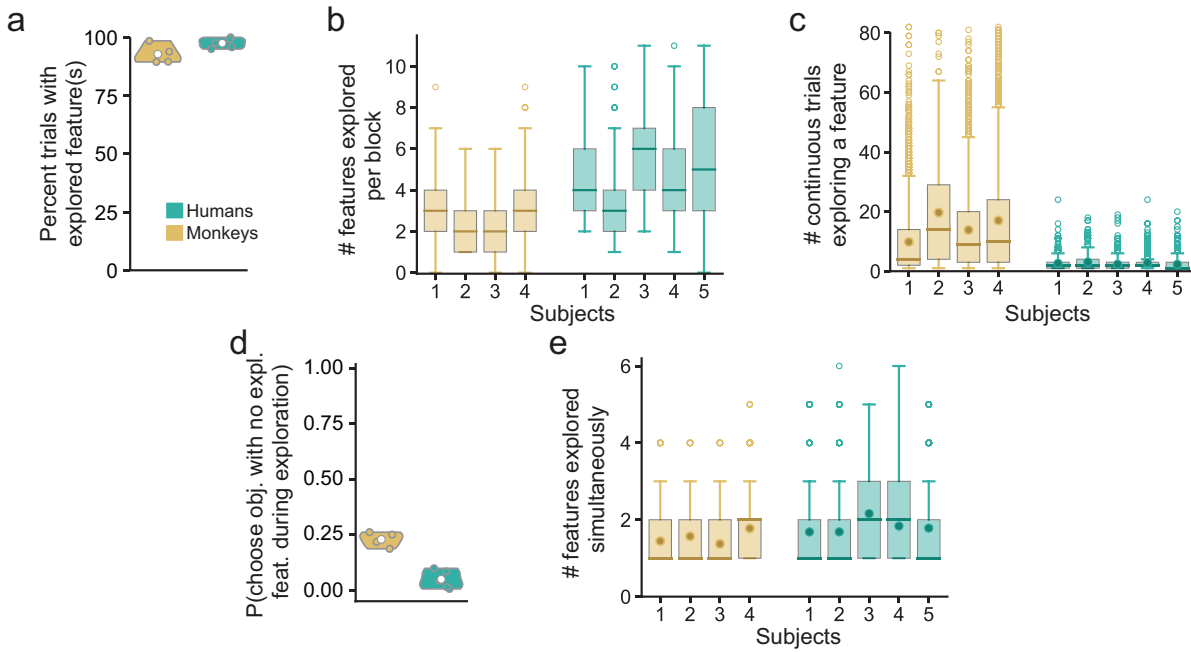
*outcomes on earlier trials (left), and on feature choice probabilities computed from most-likely state estimates derived from past, present and future choice outcomes (right). The probability on each trial was binned according to the trial's relative position in the rule block and averaged across blocks. Line and shading represent the mean and standard deviation across subjects for each species. Dots represent block percentiles at which the average object selection probability is significantly above chance (bootstrap test with t-statistic, $p < 0.05$). **d.** Most-likely states estimated by the model for 300 trials in an example human (top) and monkey (bottom) subject. The rule on each block is outlined in black.*



1013

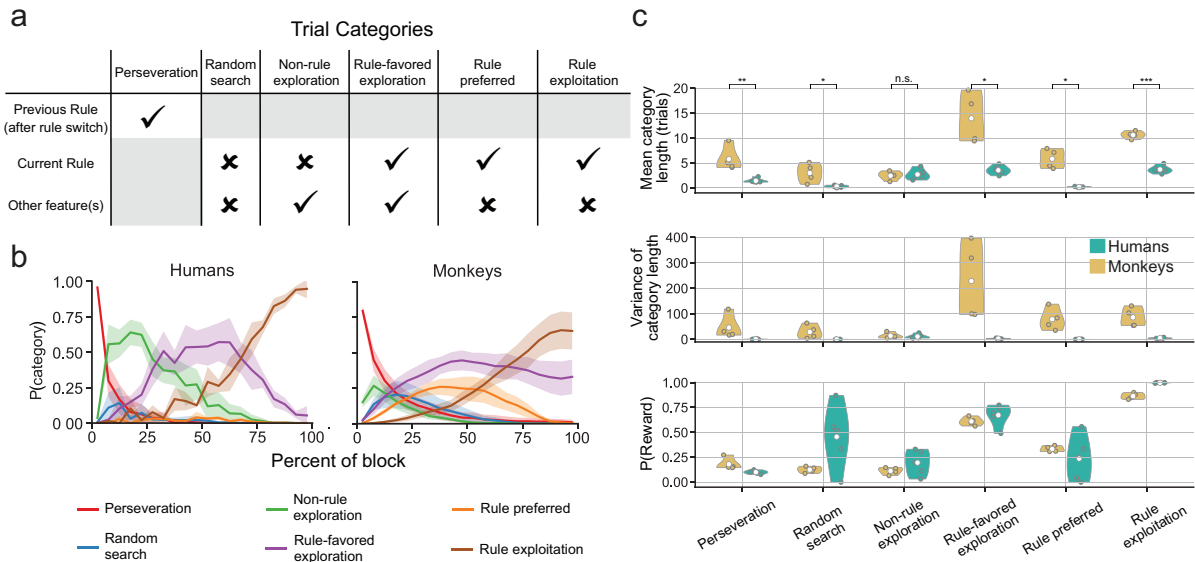
Figure 3: Model describes rule-learning dynamics in terms of changes in feature-attentional states. *a.* Choice probability of features associated with each state in human (green) and monkey (brown) subjects computed directly from model parameters and measured empirically based on most-likely state estimates. Choice probabilities order feature states akin to levels of attention. *b.* Decision process describing how humans, monkeys and the WSLs agent start, continue and stop exploring a feature, derived from their history-dependent state transition probabilities. Process is decomposed based on outcome-dependent transitions when the feature is chosen (direct effect) or not chosen (indirect effect). Arrow thickness indicates probability of the transition. Dashed lines highlight deviations from the WSLs strategy. *c.* Probabilities of demoting (left; promoting, right) the state of a chosen feature to a state with higher (lower) choice probability when an unchosen intra-dimensional feature is promoted (demoted) from the preferred to persist (random/avoid) state. Measurements test to what extent indirect effects of promoting or demoting features in the preferred state result from changing the state, and therefore the choice probability, of a chosen intra-dimensional feature. Perfect causality would

coincide with a probability of 1.0. d. Change in the choice probability of a feature in the preferred state after after receiving negative (left; positive, right) feedback for choosing a different feature. The indirect effect significantly increases (decreases) the feature choice probability.



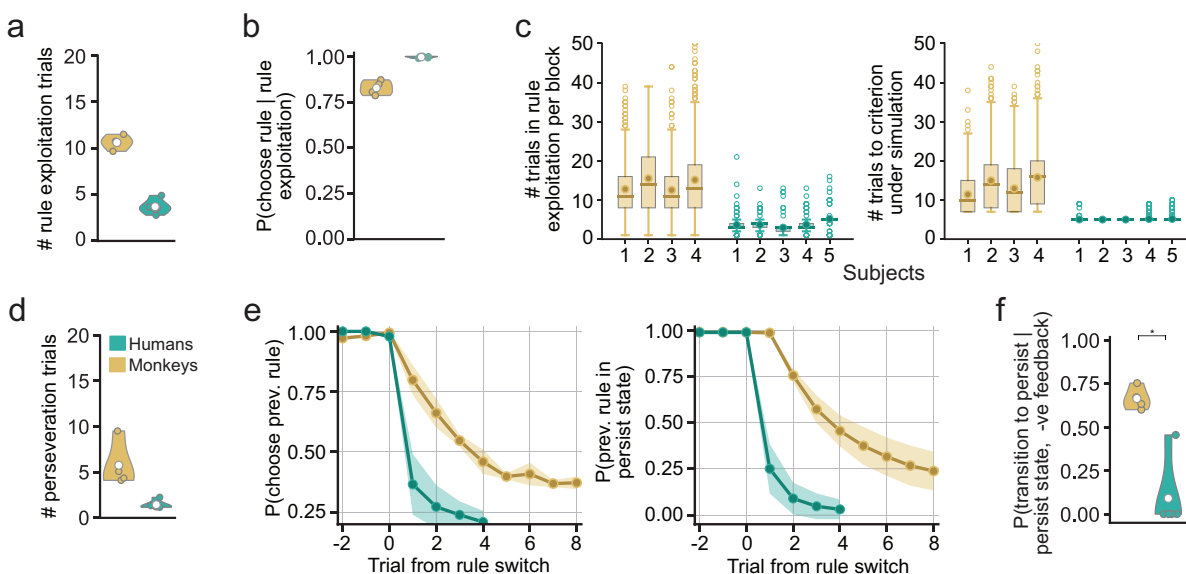
1014

Figure 4: Monkeys and humans explore multiple features for several trials in a row to evaluate them. **a.** Percent of all trials where at least one feature is under exploration by humans (green) and monkeys (brown). **b.** Distribution of the number of features explored by each monkey and human subject in a block. **c.** Distribution of number of continuous trials with a feature in an exploration state. **d.** Probability of choosing an object with all features in the random or avoid state, while at least one other feature is in the preferred or persist state. **e.** Distribution of number of features simultaneously explored by each monkey and human subject in trials where at least one feature is under exploration.



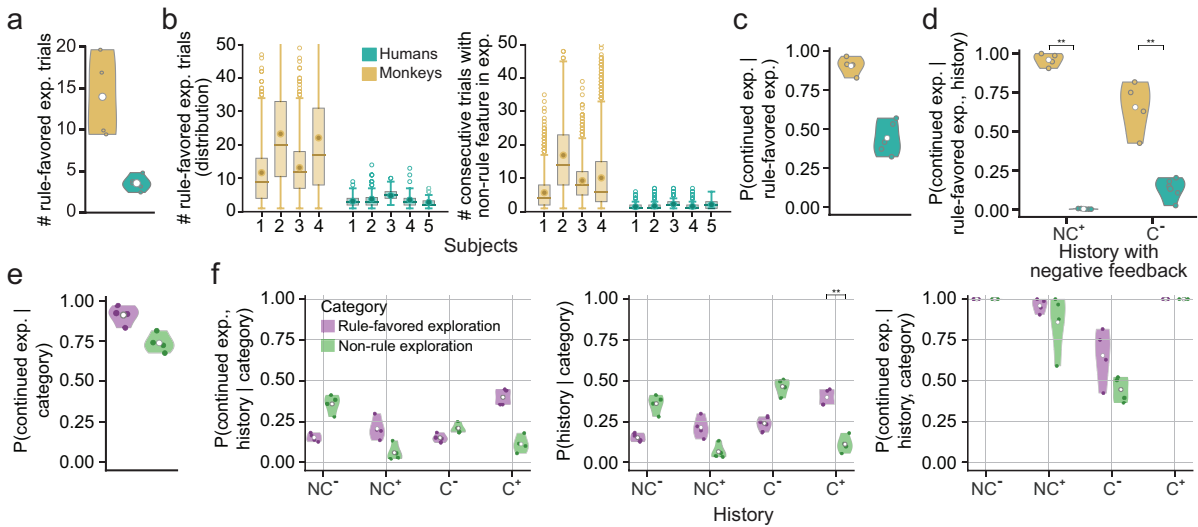
1015

Figure 5: Exploration-based trial categories reveal learning dynamics and identify causes for monkey learning performance deficit. **a.** Definition of the six trial categories based on whether the features under exploration during the trial include the rule feature. **b.** Distribution of trial categories at each percentile of rule block. Lines (shaded areas) reflect mean values (standard errors of the mean) across subjects. **c.** Trial category summary statistics (top: mean number of trials; middle: variance of number of trials; bottom: reward probability) across rule blocks for human (green) and monkey (brown) subjects. Inter-species comparisons of the mean number of trials per category reveal significant differences in the perseveration, random search, rule-favored exploration, rule preferred and rule exploitation categories (bootstrap test with t -statistic); *n.s.* - not significant; * $p < 0.1$; ** $p < 0.01$; *** $p < 0.001$.



1016

Figure 6: Random exploration and perseverative errors prolong monkey rule learning. **a.** Mean number of trials spent by human (green) and monkey (brown) subjects in the rule exploitation category per rule block. **b.** Probability of selecting an object with the rule feature across trials in the rule exploitation category. Monkeys occasionally explore other objects compared to humans. **c.** Distribution of the number of trials spent by human and monkey subjects in the rule exploitation category per rule block (left), and by simulated agents that select the rule feature with probabilities in (b) until they reach a learning criterion (right). **d.** Mean number of trials spent by human and monkey subjects in the perseveration category per rule block. **e.** The probability of humans and monkeys choosing the previous rule feature at each trial after a rule switch (left) is commensurate with the probability of the previous rule feature being associated with the persist state (right). **f.** The probability of the previous rule feature transitioning back into the persist state after its selection produces negative feedback is higher in monkeys (bootstrap test with t -statistic); $* p < 0.1$.



1017

Figure 7: Diminished negative feedback sensitivity prolongs concurrent exploration of rule and non-rule features. *a.* Mean number of trials spent by human (green) and monkey (brown) subjects in the rule-favored exploration category per rule block. *b.* Distribution across rule blocks of the number of trials spent in the rule-favored exploration category by each subject (left), and of the number of consecutive trials spent by them exploring individual non-rule features during this category (right). *c.* Probability of a non-rule feature transitioning back into an exploration state during rule-favored exploration trials. *d.* The probability of a non-rule feature transitioning back into an exploration state upon receiving negative feedback for choosing it (direct negative feedback) or positive feedback for choosing a different feature (indirect negative feedback) during rule-favored exploration trials is higher in monkeys (bootstrap test with *t*-statistic). *e.* Probability of a non-rule feature transitioning back into an exploration state during rule-favored exploration trials and non-rule exploration trials in monkeys. *f.* Joint probability of a non-rule feature transitioning back into an exploration state and each choice outcome history occurring during rule-favored exploration trials and non-rule exploration trials in monkeys (left); Probability of each choice outcome history occurring during either category (middle); Probability of a non-rule feature transitioning back into an exploration state in response to each choice outcome history during either category (right). The higher probability of a non-rule feature transitioning back into an exploration state during rule-favored exploration trials compared to non-rule exploration trials is explained by a higher incidence of direct positive feedback for choosing the non-rule feature in the former category (bootstrap test with *t*-statistic);

** $p < 0.01$.

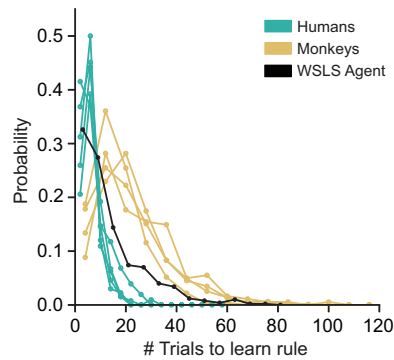
1018 **Supplementary Tables and Figures**

	Fixation	Decision	Response	Feedback	ITI
Monkeys	500 ms	≤ 4 s	Fixate (800 ms)	1.4 s (reward) 1 s / 5 s (timeout)	400 ms / 1 s
Humans	300 ms	≤ 4 s	Manual	1.5 s (text on screen)	1 s

Supplementary Table 1: Inter-species trial structure differences

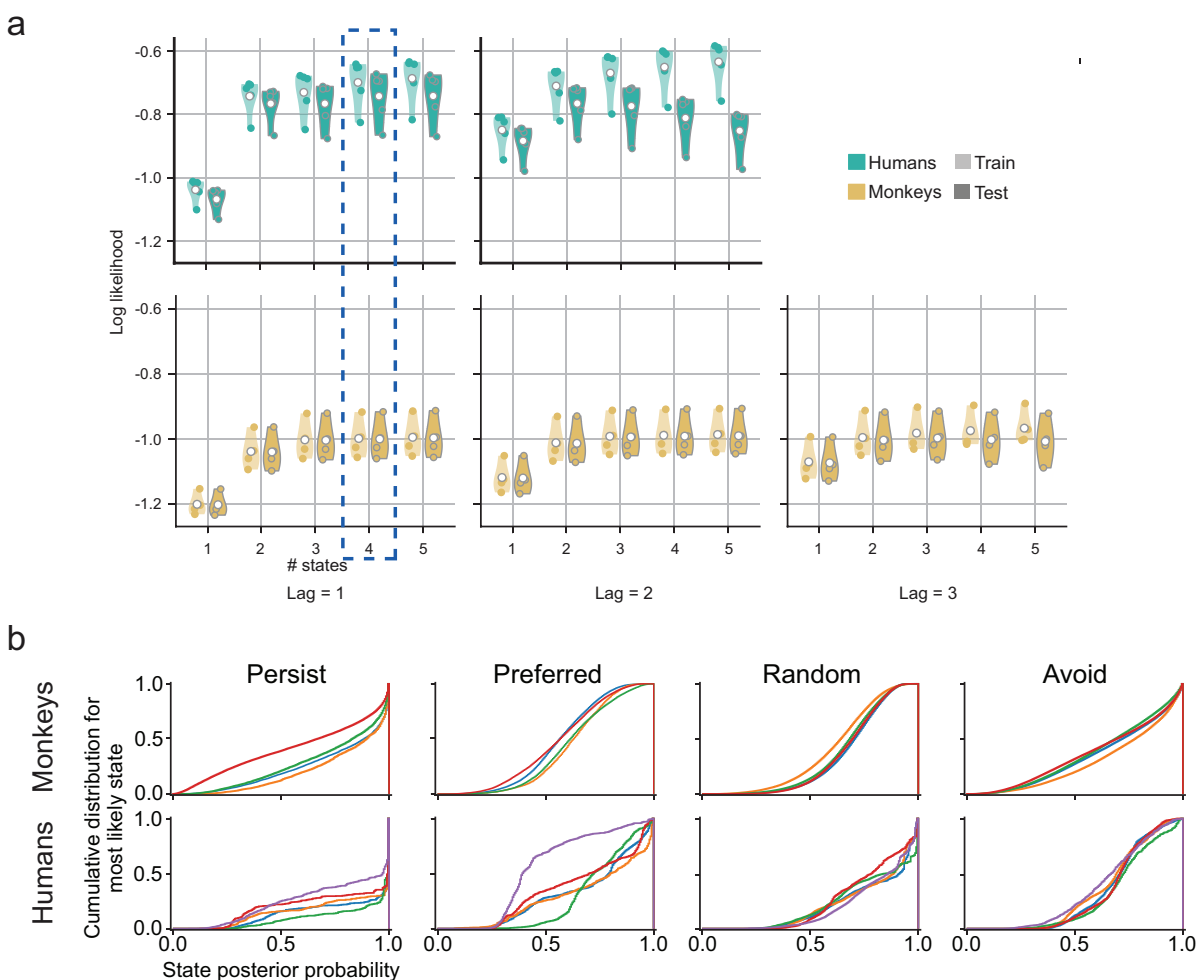
	Learning Criteria: Continuous	Learning Criteria: Proportion
Monkeys	8 correct trials	16/20 correct trials
Humans	5 correct trials	8/10 correct trials

Supplementary Table 2: Inter-species learning criteria differences



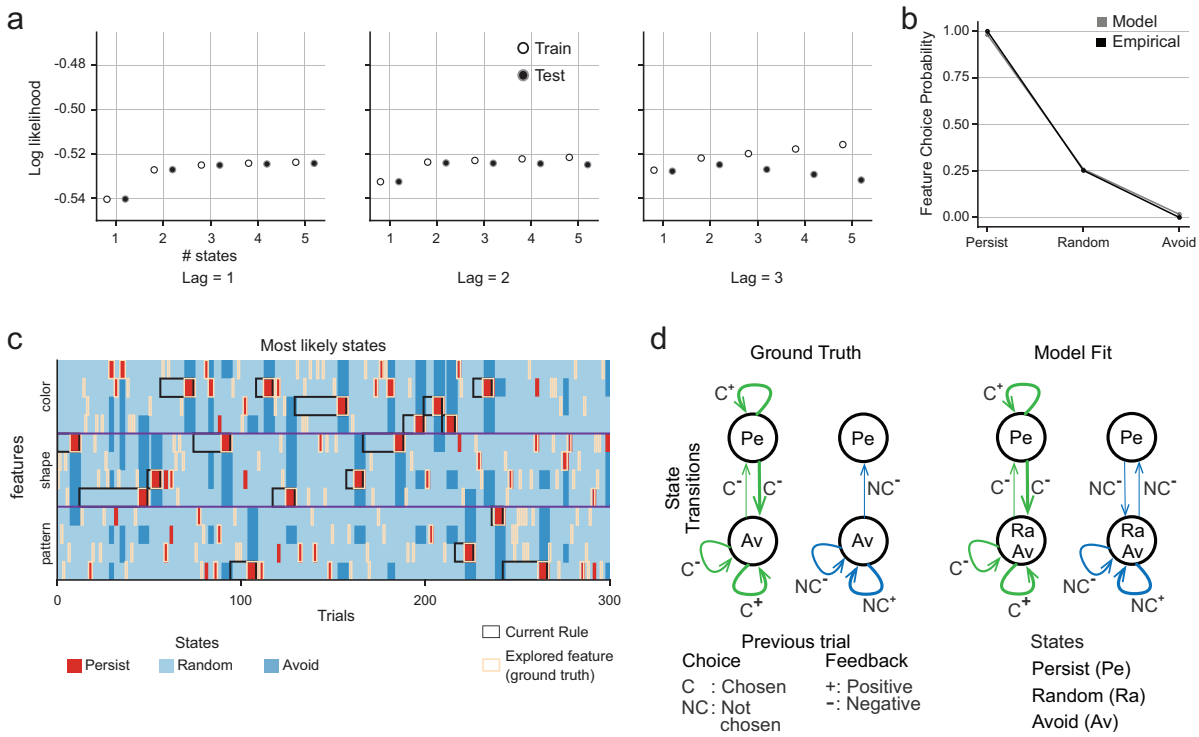
1019

Supplementary Figure 1: Monkey learning is slower than humans after correcting for learning criteria differences. Distribution of trials-to-learning-criteria in 4 monkey (brown) and 5 human (green) subjects. The less stringent learning criteria used in human subjects was applied to the monkey choices and outcomes to revise their trials-to-learning-criteria on each rule block. Yet, monkeys are over 3 times slower than humans on average. Humans are also faster than a simulated agent using the Win-Stay Lose-Shift strategy to learn WCST rules (black).



1020

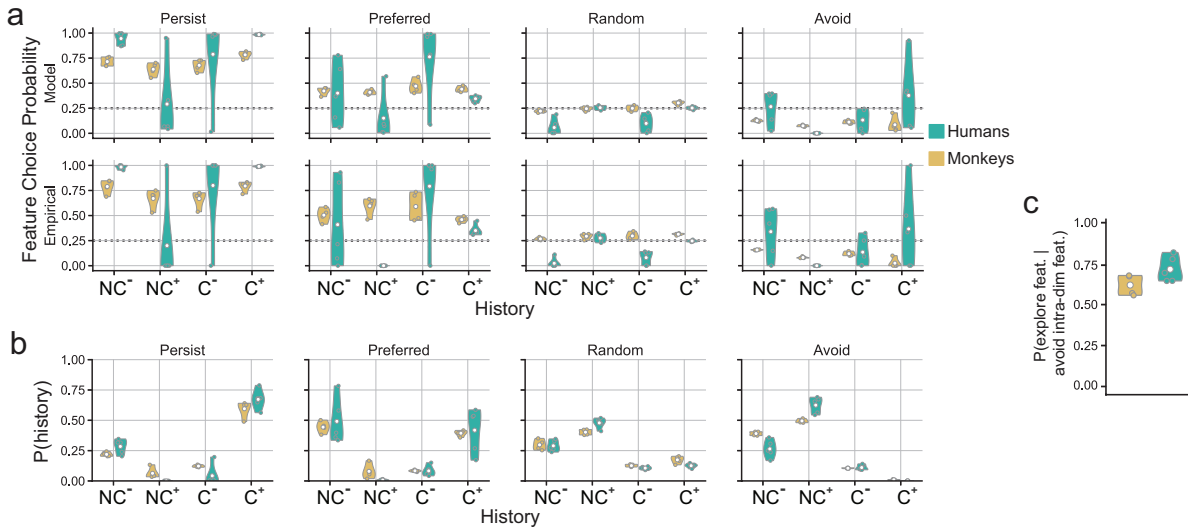
Supplementary Figure 2: Goodness of fit of a model extension to predict chosen objects is consistent with the underlying feature prediction model. a. Log-likelihood of a model extension that predicts the chosen object from feature choice probabilities. Log-likelihoods are shown separately on training and test datasets for each human (green) and monkey (brown) subject for extensions of feature choice models with varying numbers of states and that use choice outcomes from varying numbers of previous trials (lag). Each point represents the mean results over a 5-fold cross-validation and over 5 (monkey) or 10 (human) different initializations. **b.** Cumulative density of the posterior probabilities for most-likely state estimates by the model. Densities are presented separately for the 4 states and for each monkey (top) and human (bottom) subject.



1021

Supplementary Figure 3: Model recovers strategy of a simulated Win-Stay Lose-Shift agent. *a.* Model fit log-likelihoods for training and test datasets generated by a simulated agent using the Win-Stay Lose-Shift strategy on the WCST. The lag and number of states were parametrically varied across models. Each point represents the mean over a 5-fold cross-validation and over 10 different model initializations. *b.* Probability of choosing a feature associated with each state of the best-fit 3-state, lag-1 model. Plot shows probabilities generated by model parameters (gray) and measured empirically based on the most-likely state estimates (black). *c.* Most-likely states estimated by the model for 300 example trials. The rule on each block is outlined in black, and the ground-truth feature under agent exploration on each trial is outlined light brown. 57.6% of all ground-truth features under exploration are accurately assigned the persist state (choice probability = 1) by the model. *d.* Decision process for the Win-Stay Lose-Shift learning strategy employed by the agent (ground-truth; left) and fit to the agent's choices (right). The ground-truth process is modified from the one in Figure 1c to account for the composition of the chosen object by 3 features - the feature under exploration and two features that are extra-dimensional to it. Features in these other two dimensions are chosen even though they are not under exploration. The model differentiates these randomly

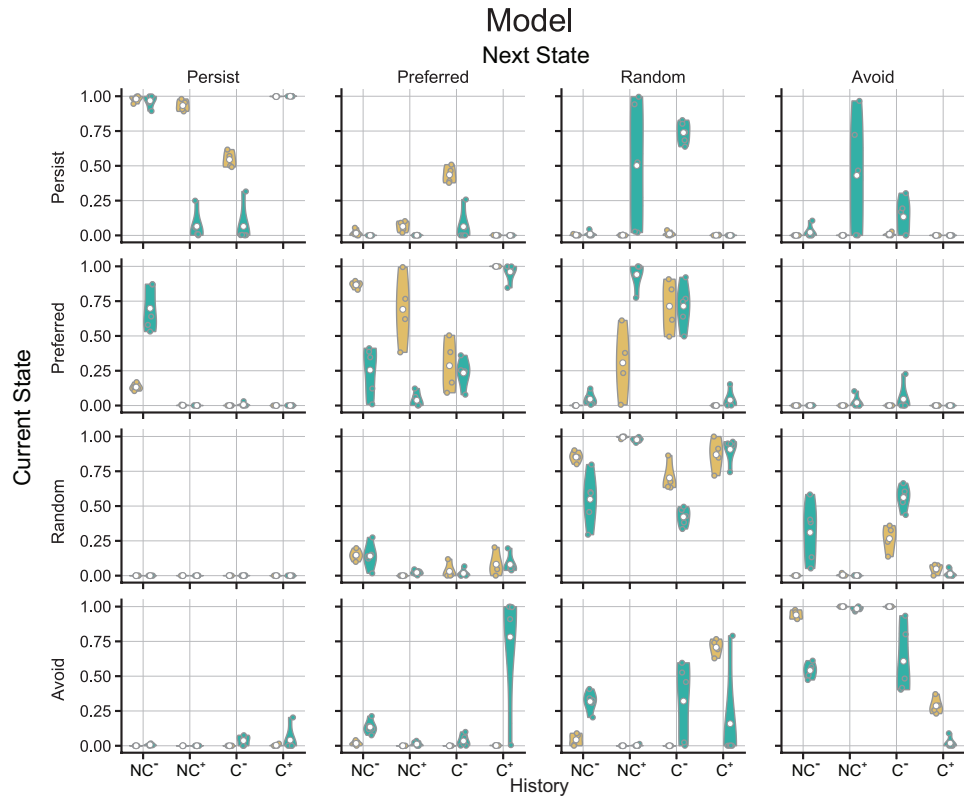
chosen features from those that are intra-dimensional to the explored feature and completely avoided. The fit recovers the underlying decision process.



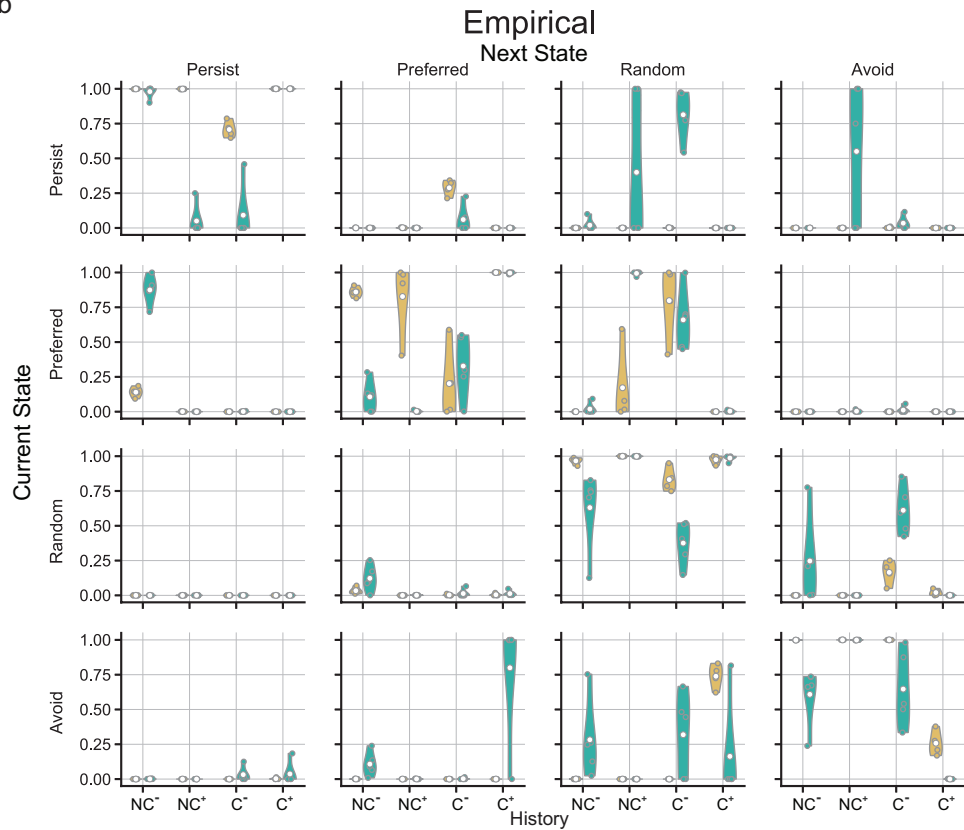
1022

Supplementary Figure 4: State- and history-dependent statistics. *a.* Feature choice probability given the feature's state and history in human (green) and monkey (brown) subjects. Values computed directly from model's parameters (above) are consistent with empirical measurements based on best-fit state estimates (below). *b.* Probability distribution of histories given the state estimate on the subsequent trial. *c.* Probability that a feature is associated with the preferred or persist state given one of its intra-dimensional counterparts is associated with the avoid state.

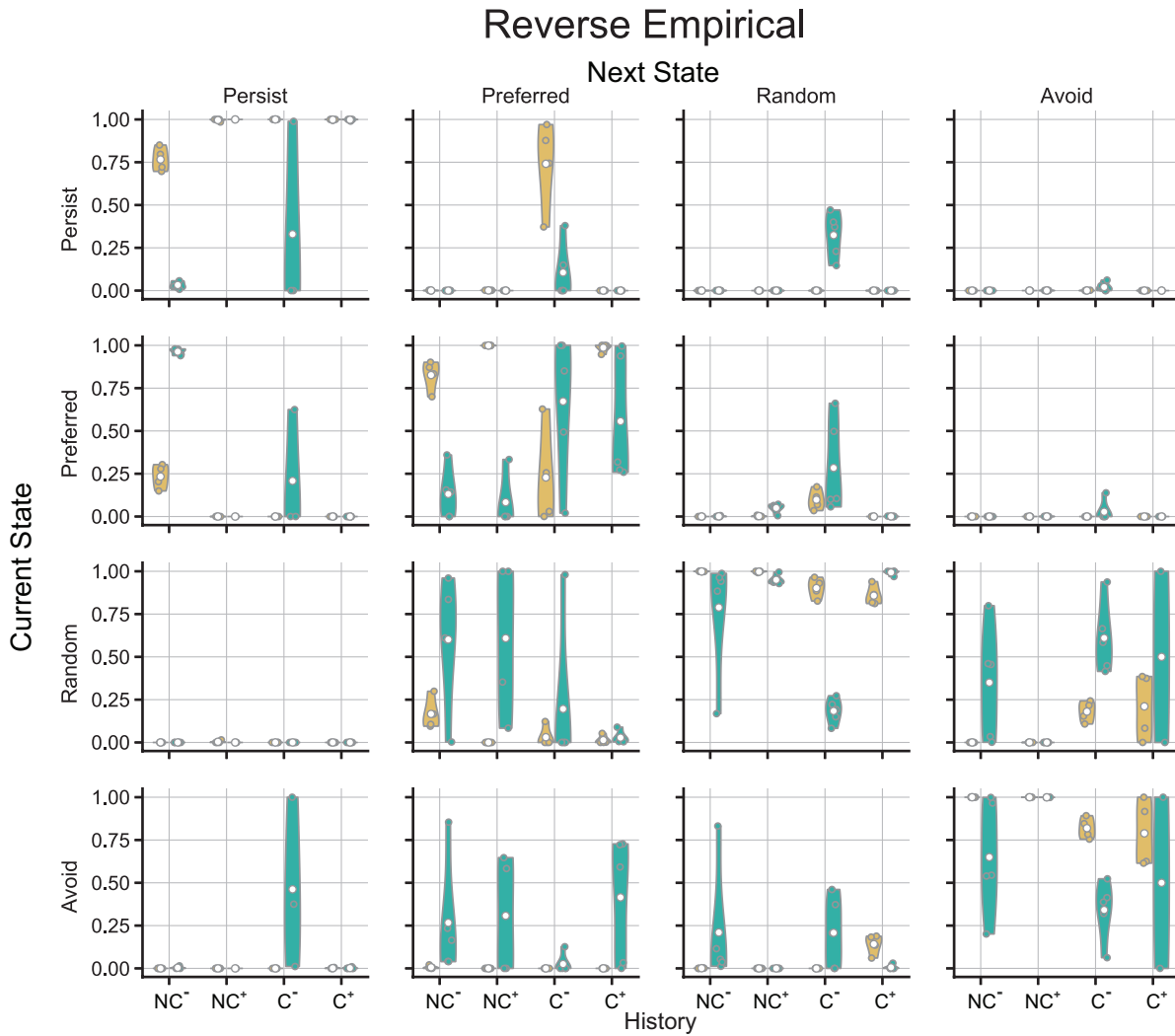
a



b

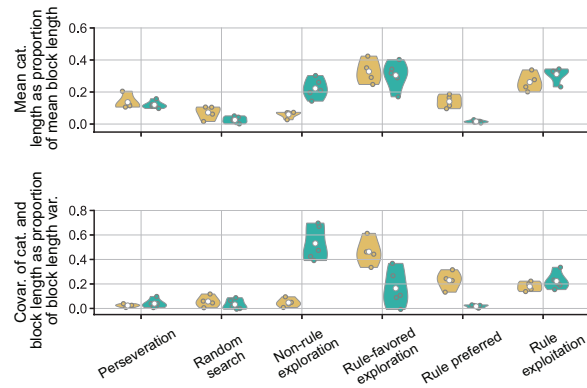


Supplementary Figure 5: State transition probabilities. a-b. Probability of a feature's state transitions given the state it was associated with on the previous trial and its choice-outcome history in human (green) and monkey (brown) subjects. Values computed directly from model's parameters (a) are consistent with empirical measurements based on best-fit state estimates (b).



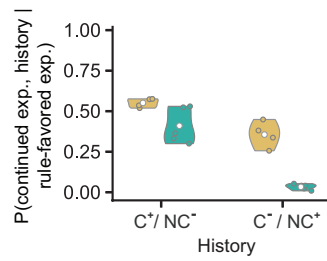
1024

Supplementary Figure 6: ‘Reverse’ state transition probabilities. Empirically measured probability of a feature’s state transitions given the state it is associated with on the next trial and its choice-outcome history in human (green) and monkey (brown) subjects.



1025

Supplementary Figure 7: Trial categories fully determine block length statistics. Mean category length as a fraction of mean block length (top) for human (green) and monkey (brown) subjects. Covariance of category length with block length as a fraction of the variance in block length (bottom). Since the categories are mutually exclusive and span all trials in a block, the sum over categories of each of these two fractions is 1. This allows an assessment of each category's contribution to the mean (top) and variance (bottom) of the block length.



1026

Supplementary Figure 8: Joint probability of a non-rule feature transitioning back into an exploration state and receiving direct/indirect positive (C⁺/NC⁻) or negative (C⁻/NC⁺) feedback during rule-favored exploration trials in human (green) and monkey (brown) subjects.