Interplay between rule learning and rule switching in a perceptual categorization task

F. Bouchacourt[*,1], S. Tafazoli[*,1], M.G. Mattar[1,2], T.J. Buschman[†,1], N.D. Daw[†,1]

[1] Princeton Neuroscience Institute and the Department of Psychology, Princeton, NJ USA

[2] Department of Cognitive Science, University of California, San Diego, San Diego, CA USA

Abstract

When performing a task in a changing world, sometimes we switch between rules already learned; at other times we must learn rules anew. Often we must do both, switching between known rules while also constantly re-estimating them. Here, we show these two processes, rule switching and rule learning, rely on distinct but intertwined computations, namely fast inference and slower incremental learning. To this end, we studied how monkeys switched between three rules. Each rule was compositional, requiring the animal to discriminate one of two features of a stimulus and then respond with an associated eye movement along one of two different response axes. By modeling behavior we found the animals learned the axis of response using fast inference (rule switching) while continuously re-estimating the stimulus-response associations within an axis (rule learning). Our results shed light on the computational interactions between rule switching and rule learning, and make testable neural predictions for these interactions.

Introduction

A crucial component of intelligence is learning from the environment, allowing one to modify their behavior in light of experience. Although a long tradition of research in areas like Pavlovian and instrumental conditioning has focused on elucidating general learning mechanisms – especially error-driven incremental learning rules, associated with dopamine and the basal ganglia[1–10] – it has also become increasingly clear that the brain's dynamics for learning can themselves be adapted[11,12]. For instance, these general-purpose incremental learning mechanisms can allow animals to gradually learn a new stimulus-response discrimination rule by trial and error. But this type of gradual adjustment seems unlikely to account for other more task-specialized learning effects: for instance, if two different stimulus-response rules are

repeatedly reinforced in alternation, animals can come to switch between them more rapidly[13,14]. Such rule switching has many formal similarities with *de novo* rule learning: here also, response behavior is modified in light of feedback, often progressively over several trials. However these more task-specialized dynamics are often modeled by a distinct computational mechanism - e.g., Bayesian inference, where animals accumulate evidence about which candidate rule currently applies[15–18]. Inference is associated with activity in the prefrontal cortex[19–30], suggesting also distinct neural mechanisms from *de novo* rule learning.

This type of inference process presupposes that the animal has previously learned about the structure of the task: which rules can apply, how often they switch, etc., and so it has typically been studied in well-trained animals[12,15,31]. However, there has been increasing theoretical interest – but relatively little direct empirical evidence – in the mechanisms by which the brain learns the broader structure of the task – thereby, in effect building task specialized inference mechanisms for rapid rule switching. For Bayesian inference models, this problem corresponds to learning the generative model of the task, e.g. inferring a mixture model over latent contexts or states (rules or task conditions, which are "latent" since they are not overtly signaled) and their properties (e.g., stimulus-response-reward contingencies)[15,16,32–36]. A more mechanistic account of rule switching, but not mutually exclusive, suggests it may be implemented by the dynamics of a recurrent neural network (RNN), in which case higher-level learning (here called "metalearning") corresponds to tuning the weights of this network[11,36–40].

Perhaps most intriguingly, these accounts often posit that a full theory of rule learning and rule switching ultimately involves an interaction between both major classes of learning mechanisms, inferential and incremental. Thus, in Bayesian inference models, it is often hypothesized that an inferential stage (e.g. prefrontal) decides which latent state is in effect, while the properties of each state are learned, conditional on this, by downstream (e.g. striatal) error-driven inferential learning[16,36]. Somewhat similarly, metalearning of RNN dynamics for rule switching has been proposed to be itself driven by incremental error-driven updates[7,11,38]. However, apart from a few interesting examples in human rule learning[16,32,41–44], this type of interaction has mostly been posited theoretically, while the two learning mechanisms have mostly been studied in regimes where they operate more or less in isolation[12,15,31,45–49].

To study rule switching and rule learning, we trained non-human primates to switch between three different category-response tasks. Depending on the rule in effect, the animals needed to attend to and categorize either the color or the shape of a stimulus, and then respond with a saccade along one of two different response axes. We observe a combination of both fast and slow learning during the task: monkeys rapidly switched into the correct response axis, consistent with inferential learning of the response state, while, within a state, the animals slowly learned category-response mappings, consistent with incremental (re)learning. To quantify the learning mechanisms underlying the animals' behavior, we tested whether inference or incremental classes of models, separately, could explain the behavior. Both classes of models produced learning-like effects - i.e., dynamic, experience-driven changes in behavior. However, neither model could, by itself, explain the combination of both fast and slow learning. Instead, we found that key features of behavior were well explained by a hybrid rule-switching and rule-learning model, which inferred which response axis was active while continually performing slower, incremental relearning of the consequent stimulus-response mappings within an axis. These results support the hypothesis that there are multiple, interacting, mechanisms that guide behavior in a contextually-appropriate manner.

Results

Task design and performance

Two rhesus macaques were trained to perform a rule-based category-response task. On each trial, the monkeys were presented with a stimulus that was composed of a color and shape (Fig. 1a). Each stimulus dimension was drawn from a subset of values along a continuous space. The animals' task was to categorize the stimulus according to either its color (red or green) or its shape ('bunny' or 'tee', Fig. 1b). Depending on the category of the stimulus, and the current rule, the animals made one of four different responses (an upper left, upper right, lower left, or lower right saccade).

Animals were trained on three different category-response rules (Fig. 1c). Rule 1 required the animal to categorize the shape of the stimulus, making a saccade to the upper-left location when

the shape was categorized as a 'bunny' and a saccade to the lower-right location when the shape was categorized as a 'tee'. These two locations – upper-left and lower-right – formed an 'axis' of response (*Axis 1*). Rule 2 was similar but required the animal to categorize the color of the stimulus and then respond on the opposite axis (*Axis 2;* red=upper-right, green=lower-left). Finally, Rule 3 required categorizing the color of the stimulus and responding on *Axis 1* (red=lower-right, green=upper-left). Note that these rules are compositional in nature, with overlapping dimensions (Fig. 1d). Rule 1 required categorizing the shape of the stimulus, while Rules 2 and 3 required categorizing the color of the stimulus. Similarly, Rules 1 and 3 required responding on the same axis (*Axis 1*), while Rule 2 required a different set of responses (*Axis 2*). In addition, the overlap in response axis for Rules 1 and 3 meant certain stimuli had congruent responses for both rules (e.g., red-tee and green-bunny stimuli) while other stimuli had incongruent responses between rules (e.g., red-bunny and green-tee). For all rules, when the animal made a correct response, it received a reward (an incorrect response led to a short 'time-out').

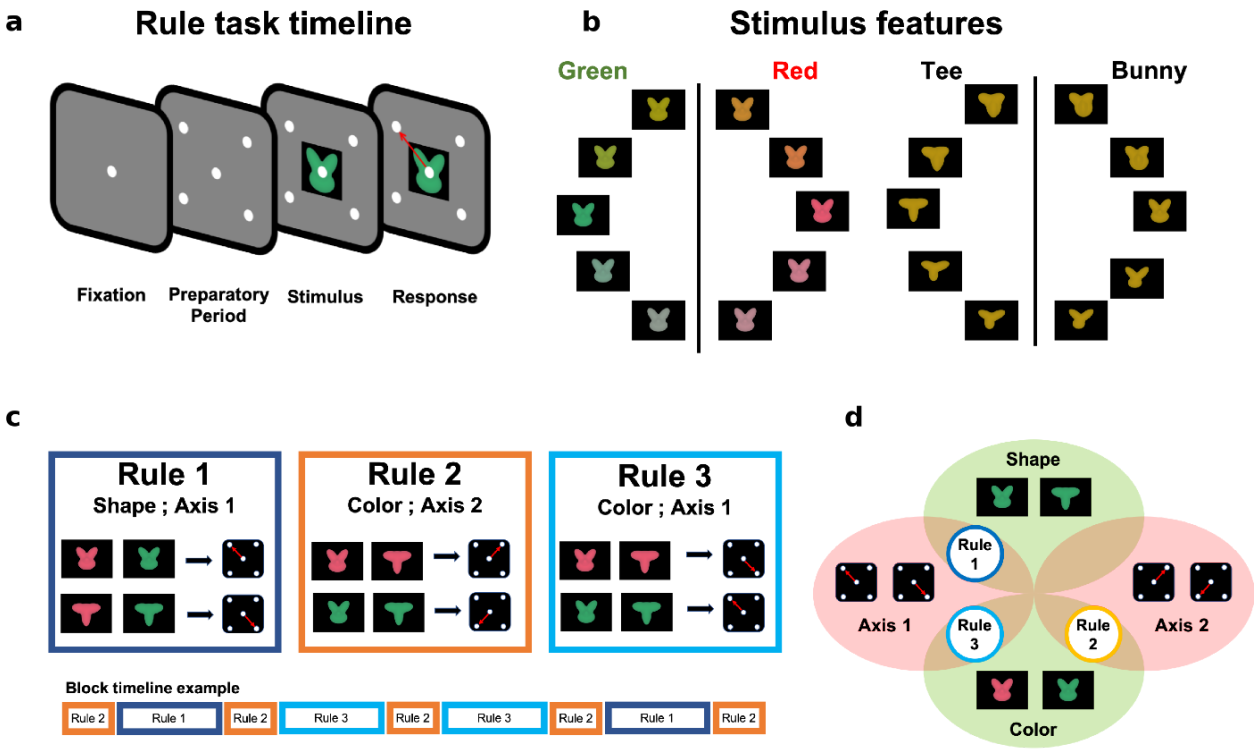
Animals had to perform the same rule during a block of trials. Critically, the animals were not explicitly cued as to which rule was in effect for that block. Instead, they had to use information about the stimulus, their response, and reward feedback, to infer which rule was in effect. After the animals discovered the rule and were performing it at a high level (defined as >70% on the past 100 trials, see Methods) the rule would switch. Although unpredictable, the moment of switching rules was cued to the animal (with a flashing screen). Importantly, this switch-cue did not indicate which rule was now in effect (just the switch itself). To facilitate learning and performance, the sequence of rules across blocks was semi-structured such that the axis of response always changed following a block switch (i.e., after a Rule 2 block the animal performed either Rule 1 or Rule 3, chosen pseudorandomly, and vice versa, see block timeline example in Fig. 1c).

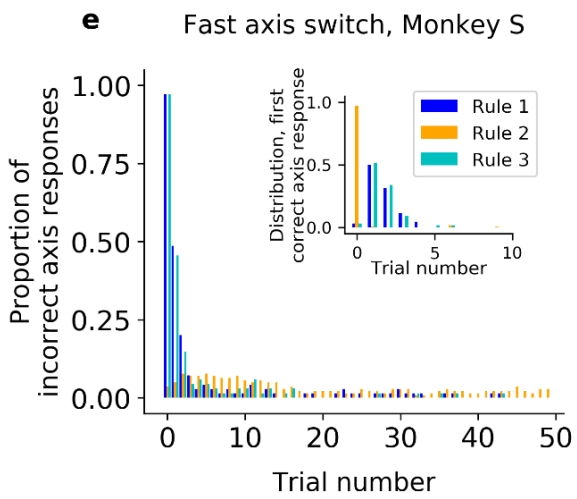
Learning the axis of response was fast: both monkeys switched into the correct axis nearly instantaneously. Indeed, Monkey S almost always responded on Axis 2 (the response axis consistent with Rule 2) immediately after each block switch cue (97%, CI=[0.90,0.99] in Rule 1 ; 97%, CI=[0.92,0.98] in Rule 2 ; 97%, CI=[0.90,0.99] in Rule 3 ; see Fig. 1e). Then, if this was

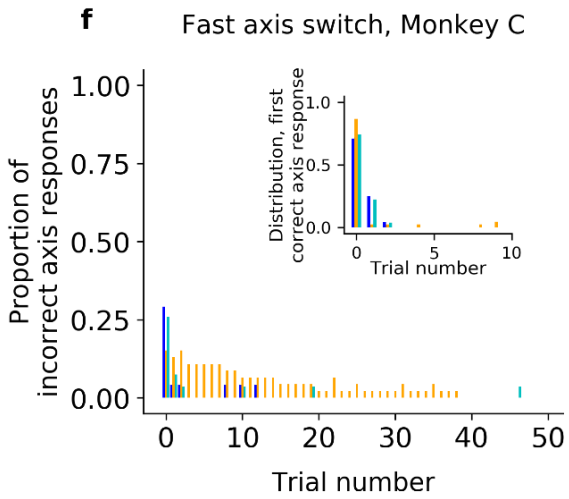
incorrect, it switched to the correct axis within 5 trials on 97%, CI=[0.90,0.99] of blocks of Rule 1, and 94% CI=[0.86,0.98] of blocks of Rule 3. Monkey C instead tended to alternate the response axis on the first trial following a switch cue (it made a response on the correct axis on the first trial with a probability of 71%, CI=[0.51,0.85] in Rule 1 ; 85%, CI=[0.72,0.92] in Rule 2 ; and 84%, CI=[0.55,0.87] in Rule 3), implying an understanding of the pattern of axis changes with block switches (Fig. 1f). Both monkeys maintained the correct axis with very few off-axis responses throughout the block (at trial 20, Monkey S: 1.4%, CI=[0.0025,0.077] in Rule 1 ; 2.1%, CI=[0.0072,0.060] in Rule 2 ; 0%, CI=[0,0.053] in Rule 3 ; Monkey C: 0%, CI=[0,0.14] in Rule 1 ; 4.3%, CI=[0.012,0.15] in Rule 2 ; 3.7%, CI=[0.0066,0.18] in Rule 3).

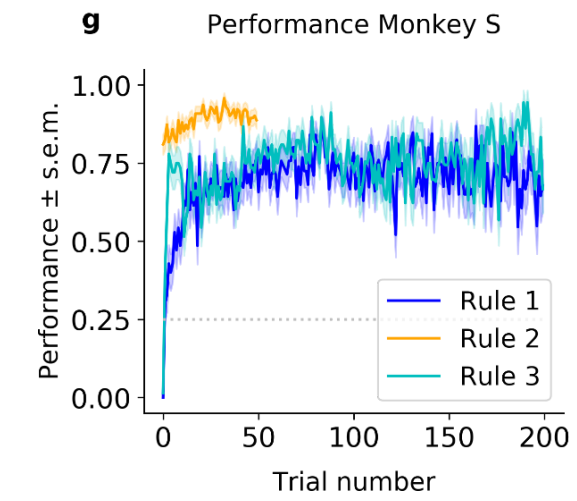
Overall, both monkeys performed the task well above chance (Fig. 1g,h). When the rule switched to Rule 2, the animals quickly switched their behavior: Monkey S responded correctly on the first trial in 81%, CI=[0.74,0.87] of Rule 2 blocks, and reached 91%, CI=[0.85,0.95] after only 20 trials (Monkey C being respectively at 78%, CI=[0.65,0.88] ; and 85%, CI=[0.72,0.92]). In Rule 1 and Rule 3, their performance also exceeded chance level quickly. In Rule 1, although the performance of Monkey S was below chance on the first trial (0%, CI=[0,0.052]; 46%, CI=[0.28,0.65] for Monkey C), reflecting perseveration on the previous rule, performance quickly climbed above chance (77% after 50 trials, CI=[0.66,0.85]; 63%, CI=[0.43,0.79] for Monkey C). A similar pattern was seen for Rule 3 (initial performance of 1.5%, CI=[0.0026,0.079] and 78%, CI=[0.67,0.86] after 50 trials for Monkey S; 41%, CI=[0.25,0.59] and 67%, CI=[0.48,0.81] for Monkey C, respectively).

Importantly, the monkeys were slower to switch to Rule 1 and Rule 3 than to switch to Rule 2. On the first 20 trials, the difference in average percent performance of Monkey S was Δ=35 between Rule 2 and Rule 1, and Δ=22 between Rule 2 and Rule 3 (significant Fisher test comparing Rule 2 to Rule 1 and Rule 3, with $p<10(-4)$ in both conditions ; respectively Δ=35, Δ=24 and $p<10(-4)$ for Monkey C).









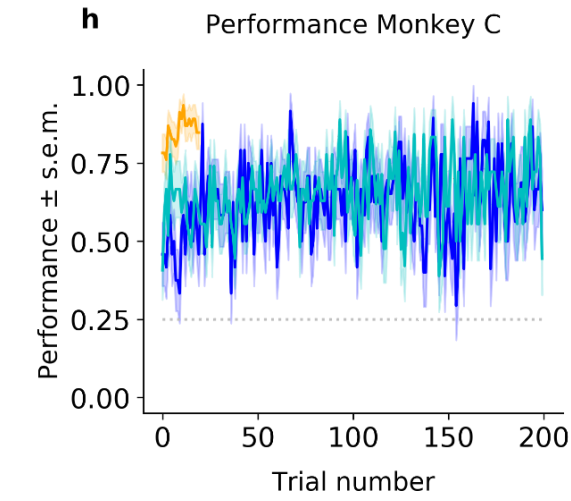


Figure 1: Task design and performance. (a) Schematic of a trial. (b) Stimuli were drawn from a two-dimensional feature space, morphing both color (left) and shape (right). Stimulus categories are indicated by vertical lines and labels. (c) The stimulus-response mapping for the three rules, and an example of a block timeline. (d) Venn diagram showing the overlap between rules. (e,f) Proportion of responses on the incorrect axis for the first 50 trials of each block for (e) Monkey S and (f) Monkey C. Insets: Trial number of the first response on the correct axis after a block switch, respectively for Monkey S and C. (g,h) Average performance (sample mean and standard error of the mean) for each rule, for (g) Monkey S and (h) Monkey C.

Learning rules *de novo* cannot capture the behavior

To perform the task, the animals had to learn which rule was in effect during each block of trials. This required determining both the response axis and the relevant feature. The central result from the monkeys' behavior above was that learning showed a mixture of fast switching, reminiscent of inference models, and slow refinement, as in error-driven incremental learning. In order to explain this behavior, we tested both inference and incremental classes of models, separately. As in previous work[50–54], all our models shared common noisy perceptual input and action selection stages (Fig. S1, and Methods). A single parameter was used to capture noisy perception: the concentration of a von Mises distribution around the true stimulus identity, with separate values for the color and shape features. A single lapse term accounted for noise at the choice stage. While all models shared the same mechanisms for perception and response, the intervening mechanism for mapping stimulus to action value differed between models (Fig. S1 and Methods).

First, the rapid switching of response axes cannot be explained by error-driven learning models, like Q learning, which would entail gradually relearning the stimulus-response mappings *de novo* at the start of each block. Such models work by learning the reward expected for different stimulus-response combinations, using incremental running averages to smooth out trial-to-trial stochasticity in reward realization – here, due to perceptual noise in the stimulus classification. To test the ability of incremental learning models to capture the animals' behavior, we fit a Q learning model to their behavior. In particular, we fit a variant of Q learning (model QL, see Fig.

S1 and Methods) that was elaborated to improve its performance in this task: for each action, model QL parameterized the mapping from stimulus to reward linearly using two basis functions over the feature space (one binary indicator each for color and shape), and used error-driven learning to estimate the appropriate weights on these for each block. This scheme effectively builds-in the two relevant feature-classification rules (shape and color). Also, this variant of the model resets the weights to fixed initial values to start over afresh at each block switch. Yet, even with these built-in advantages, the model was unable to rapidly switch axes, as it needed to relearn feature-response associations after each block switch (simulations under best-fitting parameters shown in Fig. 2 for Monkey S, Fig. S2 for Monkey C). Several tests show that the model learned more slowly than the animals (Fig. 2a,b). For instance, the model fitted on Monkey S's behavior responded on Axis 2 on the first trial of the block only 50% of the time in all three rules (Fig. 2a, Fisher test of model simulations against data: $p<10(-4)$ for the three rules). The model thus failed to capture the initial bias of Monkey S for Rule 2 discussed above. Importantly, the model switched to the correct axis within 5 trials on only 58% of blocks of Rule 1, and 57% of blocks of Rule 3 (Fig. 2b, Fisher test against monkey behavior: $p<10(-4)$). Finally, the model performed 24% of off-axis responses after 20 trials in Rule 1, 21% in Rule 2, and 21% in Rule 3, all much higher than what was observed in the monkey's behavior (Fisher test $p<10(-4)$).

In addition, because of the need to relearn feature-response associations after each block switch, the incremental QL model was unable to capture the dichotomy between the monkeys' slower learning in Rule 1 and 3 (which share Axis 1) and the faster learning of Rule 2 (using Axis 2). As noted above, the monkeys performed Rule 2 at near asymptotic performance from the beginning of the block but were slower to learn which feature to attend to on blocks of Rule 1 and Rule 3 (Fig. 1g,h). In contrast, the incremental learner performed similarly on all three rules (Fig. 2c for Monkey S, S2c for Monkey C). In particular, it performed correctly on the first trial in only 25% of Rule 2 blocks (Fisher test against behavior: $p<10(-4)$), and reached only 62% after 20 trials ($p<10(-4)$). As a result, on the first 20 trials, the difference in average percent performances was only $\Delta=4.0$ between Rule 2 and Rule 1, and was only $\Delta=0.76$ between Rule 2 and Rule 3 (similar results were seen when fitting the model to Monkey C, see Fig. S2). The same pattern of results was seen when the initial weights were free parameters (see Methods).

Thus simple incremental relearning of the axes and features *de novo* could not reproduce both the instantaneous relearning of the correct axis after a block switch and the discrepancy in learning speed for the different rules.

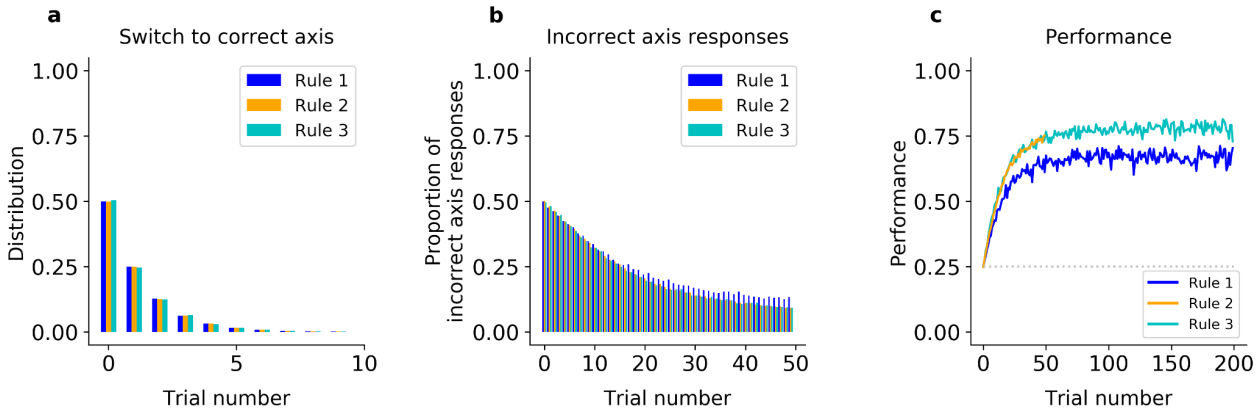


Figure 2: Incremental learner (QL) model fitted on Monkey S behavior (see Fig. S2 for Monkey C). (a) Trial number of the first response of the model on the correct axis after a block switch (compare to Fig. 1e, inset). (b) Proportion of responses of the model on the incorrect axis for the first 50 trials of each block (compare to Fig. 1e). (c) Model performance for each rule (averaged over blocks, compare to Fig. 1g).

Pure inference of previously learned rules cannot capture the behavior

The results above suggest that incremental learning is too slow to explain the quick switch between response axes displayed by the monkeys. The speed of learning suggests a different class of models that leverage Bayesian inference may be able to capture this aspect of the behavior. A fully informed Bayesian ideal observer model (IO, see Fig. S1 and Methods) uses statistical inference to continually estimate which of the three rules is in effect, accumulating evidence ("beliefs") over the history of previous stimuli, actions, and rewards. It can then choose the optimal action for any given stimulus, by averaging the associated actions' values under each rule, weighted by the estimated likelihood that each rule is in effect. Like incremental learning, the IO model learns and changes behavior depending on experience. However, unlike incremental models, these models leverage perfect knowledge of the rules to learn rapidly,

limited only by stochasticity in the evidence (here, noisy stimulus perception). Since perceptual noise is the limiting factor and is shared across rules, the IO model makes two characteristic predictions: the speed of learning should be shared across rules using the same features, and the speed of initial (re)learning after a block switch should be coupled to the asymptotic level of performance.

As expected, the IO model reproduced the animals' ability to rapidly infer the correct axis (Fig. S3), as the beliefs over rules were reset to fixed (fitted) values after a block switch. Fitted to Monkey S behavior, the model initially responded on Axis 2 almost always immediately after each block switch cue (96% in all rules, Fisher test against monkey's behavior $p>0.4$). Then, if this was incorrect, the model typically switched to the correct axis within 5 trials on 89% of blocks of Rule 1, and 95% of blocks of Rule 3 (Fisher test against monkey behavior: $p>0.2$ in both rules). The model maintained the correct axis with very few off-axis responses throughout the block (after trial 20, 1.3% in Rule 1 ; 1.1% in Rule 2 ; 1.2% in Rule 3, Fisher test against monkey's behavior: $p>0.6$ in all rules).

However, the IO model could not capture the discrepancy of learning speed for the different rules. To understand why, we looked at performance as a function of stimulus difficulty. The monkey's performance depended on how difficult it was to categorize the stimulus (i.e., the morph level; Fig. 3a-c for Monkey S, Fig. S4a-c for Monkey C). For example, in color blocks (Rule 2 and 3), the monkeys performed better for a 'prototype' red stimulus than for a 'morphed' orange stimulus (Fig. 3a-c). Indeed, on "early trials" (first 50 trials) of Rule 2, Monkey S correctly responded to 96% (CI=[0.95,0.97]) of prototype stimuli, and only to 91%, CI=[0.90,0.92] of 'morphed' stimuli ($p<10(-4)$; similar results for Monkey C in Fig. S4). Rule 3 had a similar ordering: Monkey S correctly responded to 80% (CI=[0.77,0.82]) of prototype stimuli, and only 62%, CI=[0.60,0.64] of 'morphed' stimuli ($p<10(-4)$). A similar trend was seen during "later trials" in Rule 3 (trials 50 to 200; 89%, CI=[0.88,0.90] and 74%, CI=[0.73,0.75], respectively for prototype and morphed stimuli, $p<10(-4)$).

Importantly, there was a discrepancy between the performance of 'morphed' stimuli in Rule 2 versus Rule 3, with a difference in average percent performance of $\Delta=28$ for the first 50 trials in

both rules ($p<10(-4)$), and still $\Delta=17$ if we considered Rule 2 against the last trials of Rule 3 ($p<10(-4)$). The same discrepancy was observed between the performance of 'prototype' stimuli in Rule 2 versus Rule 3, with a difference in average percent performance of $\Delta=16$ for the first 50 trials in both rules ($p<10(-4)$), and still $\Delta=6.8$ if we considered the last trials of Rule 3 ($p<10(-4)$).

The IO model captured the performance ordering on morphed and prototype stimuli for each rule separately (Fig. 3d-i, similar results for the model reproducing Monkey C, Fig. S4). However, because all errors were exclusively driven by perceptual noise (and lapses), the model performed similarly for morphed stimuli on Rule 2 and Rule 3. Indeed, because both rules involved categorizing color, the speed of learning was shared across Rule 2 and Rule 3, and initial learning in both rules on the first 50 trials was coupled to the asymptotic performance in later trials of Rule 3. The model thus had to trade-off between behavioral performance in each rule. So, while the IO model, using best-fit parameters, reproduced the animals' lower asymptotic performance in Rule 3 by increasing color noise (low concentration), it failed to capture the high performance on Rule 2 early on (Fig. 3e,f, Fig. S4e,f). The resulting difference in average percent performance for 'morphed' stimuli was only $\Delta=4.0$ for the first 50 trials and $\Delta=0.0044$ if we considered the last trials of Rule 3 (respectively $\Delta=4.2$ and $\Delta=0.032$ for 'prototype'). Conversely, if we forced the model to improve color perception (a high concentration parameter, Fig. 3h,i, and Fig S4h,i), then it was able to account for the monkeys' performance on Rule 2, but failed to match the animals' behavior on Rule 3. The resulting difference in average percent performance was again only $\Delta=3.4$ for the first 50 trials, and $\Delta=-0.17$ if we considered the last trials of Rule 3 (respectively $\Delta=3.2$ and $\Delta=-0.16$ for 'prototype').

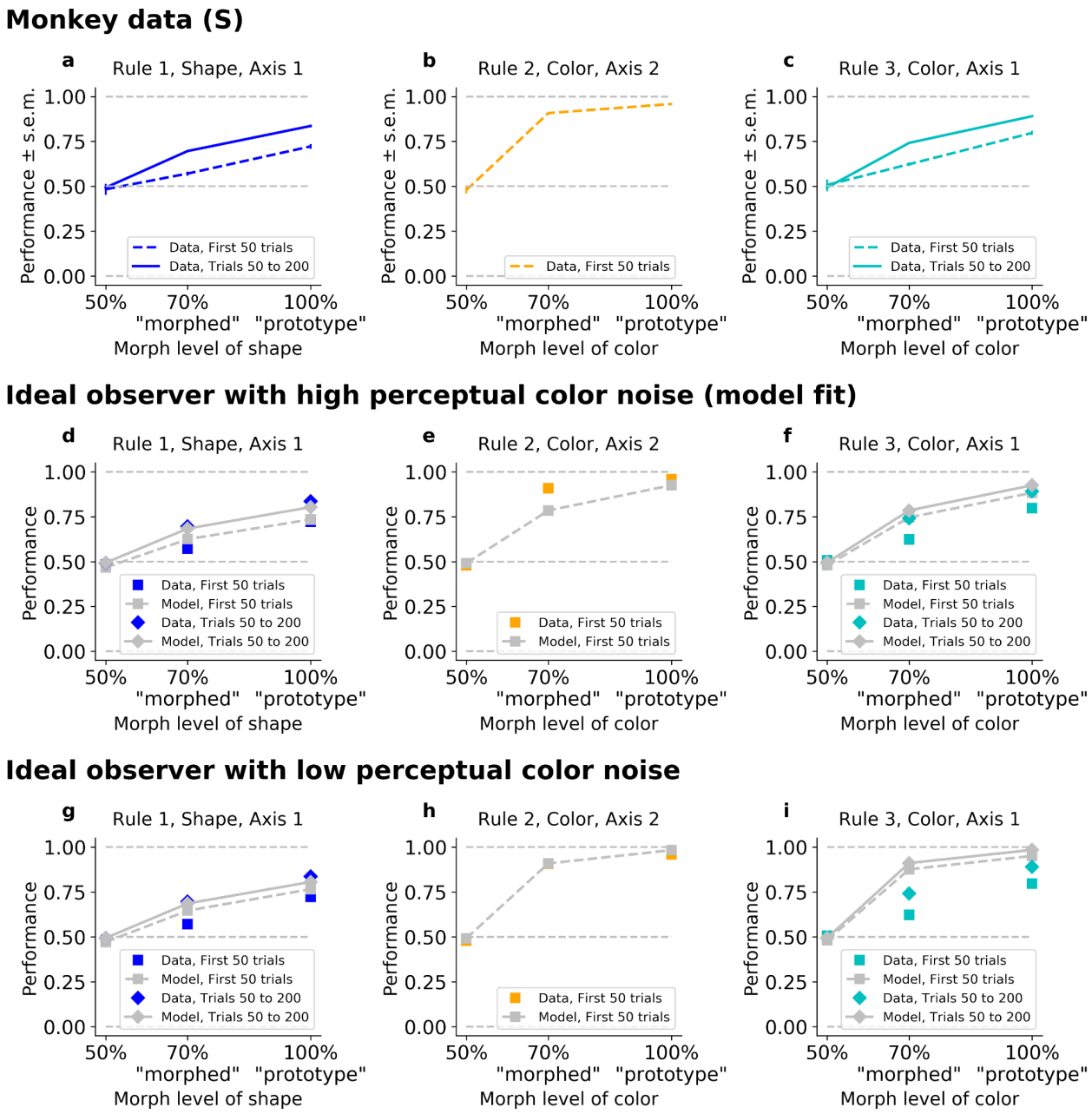


Figure 3: The ideal observer (IO), slow or fast, but not both. Fitted on Monkey S behavior (see Fig. S3 for Monkey C). (a-b-c) Performance for Rule 1, 2, and 3, as a function of the morphed version of the relevant feature. (d,e,f) Performance for Rule 1, 2, and 3, for IO model with high color noise. This parameter regime corresponds to the case where the model is fitted to the monkey's behavior (see Methods). (j,k,l) Performance for Rule 1, 2, and 3, for IO model with low color noise. Here, we fixed KC=6.

The key features of monkeys' behavior are reproduced by a hybrid model composing inference over axes and incremental relearning over features

To summarize, the main characteristics of the animals' behavior were 1) rapid learning of the axis of response after a block switch, 2) immediately high behavioral performance of Rule 2, the only rule on Axis 2, and 3) slower relearning of Rules 1 and 3, which were competing using different features on Axis 1. Altogether, these results suggest that the animals learned axes and features separately, with fast learning of the axes and slower learning of the features. One way to conceive this is as a Bayesian inference model (similar to IO), but relaxing the assumption that the animal had perfect knowledge of the underlying rules (i.e., all of the stimulus-action-reward contingencies). We propose that the animals maintained two latent states (e.g., one corresponding to each axis of response) instead of the three rules we designed. The stimulus-action-reward mappings would be stable for Axis 2, but subject to continual re-estimation between Rules 1 and 3 for Axis 1. To test this hypothesis, we implemented a hybrid model that inferred the axis of response while incrementally learning which features to attend for that response axis (Hybrid Q Learner, "HQL" in the Methods, Fig. S1). In the model, the current axis of response was inferred through Bayesian evidence accumulation (as in the IO model). Below that, at the feature level, the HQL model used incremental learning to learn a set of feature-response weights for each axis of response.

Intuitively, this model could explain all three core behavioral observations. First, inference allows for rapid switching between axes. Second, because the weights for Axis 2 did not change, the model was able to immediately perform well on Rule 2. Third, because Rules 1 and 3 shared an axis of response, and, thus, a single set of feature-response association weights, this necessitated relearning associations for each block, reflected in the animal's slower learning. Consistent with this intuition, the HQL model provided an accurate account of the animals' behavior.

First, unlike the QL model, the HQL model reproduced the fast switch to the correct axis (Fig. 4a,b and Fig. S5a,b and Fig. S6a-c,g-i). Fitted to Monkey S behavior, the model initially responded on Axis 2 almost always immediately after each block switch cue (91% in Rule 1, 89% in Rule 2 and Rule 3, Fisher test against monkey's behavior $p>0.05$). Then, if this was

incorrect, the model typically switched to the correct axis within 5 trials on 91% of blocks of Rule 1 and Rule 3 (Fisher test against behavior: $p>0.2$ in both rules). Similar to the animals, the model maintained the correct axis with very few off-axis responses throughout the block (on trial 20, 1.4% in Rule 1 ; 1.3%, in Rule 2 ; 1.5% in Rule 3, Fisher test against monkey's behavior: $p>0.7$ in all rules).

Second, contrary to the IO model, the HQL model could capture the animal's fast performance on Rule 2 and slower performance on Rules 1 and 3 (Fig. 4c and Fig. S5c). As detailed above, animals were significantly better on Rule 2 than Rules 1 and 3 on the first 20 trials. The model captured this difference: fitted on Monkey S's behavior, the difference in average percent performance on the first 20 trials was $\Delta=31$ between Rule 2 and Rule 1, and $\Delta=29$ between Rule 2 and Rule 3 (a Fisher test against monkey's behavior gave $p>0.05$ for the first trial, $p>0.1$ for trial 20).

Third, the HQL model captured the trade-off between the animals' initial and asymptotic behavioral performance in Rule 2 and Rule 3, for both 'morphed' and 'prototype' stimuli (Fig. 4d-f and S5d-f). Similar to the animals, the resulting difference in average percent performance for 'morphed' stimuli was $\Delta=26$ for the first 50 trials and $\Delta=16$ if we considered the last trials of Rule 3 (respectively $\Delta=19$ and $\Delta=9.9$ for 'prototype'). The model was able to match the animals' performance because the weights for Axis 2 did not change from one Rule 2 block to another (Fig. S6e,k), and the estimated perceptual accuracy of color was high (high concentration of the VonMises distribution) to account for the high performance of both morphed and prototype stimuli (Fig. 4e and Fig. S5e). To account for the slow re-learning observed for Rules 1 and 3, the best-fitting learning rate for feature-response associations was relatively low (Fig. 4d,f and Fig. S5d,f, Fig. S6d,f,j,l, Fig. S11).

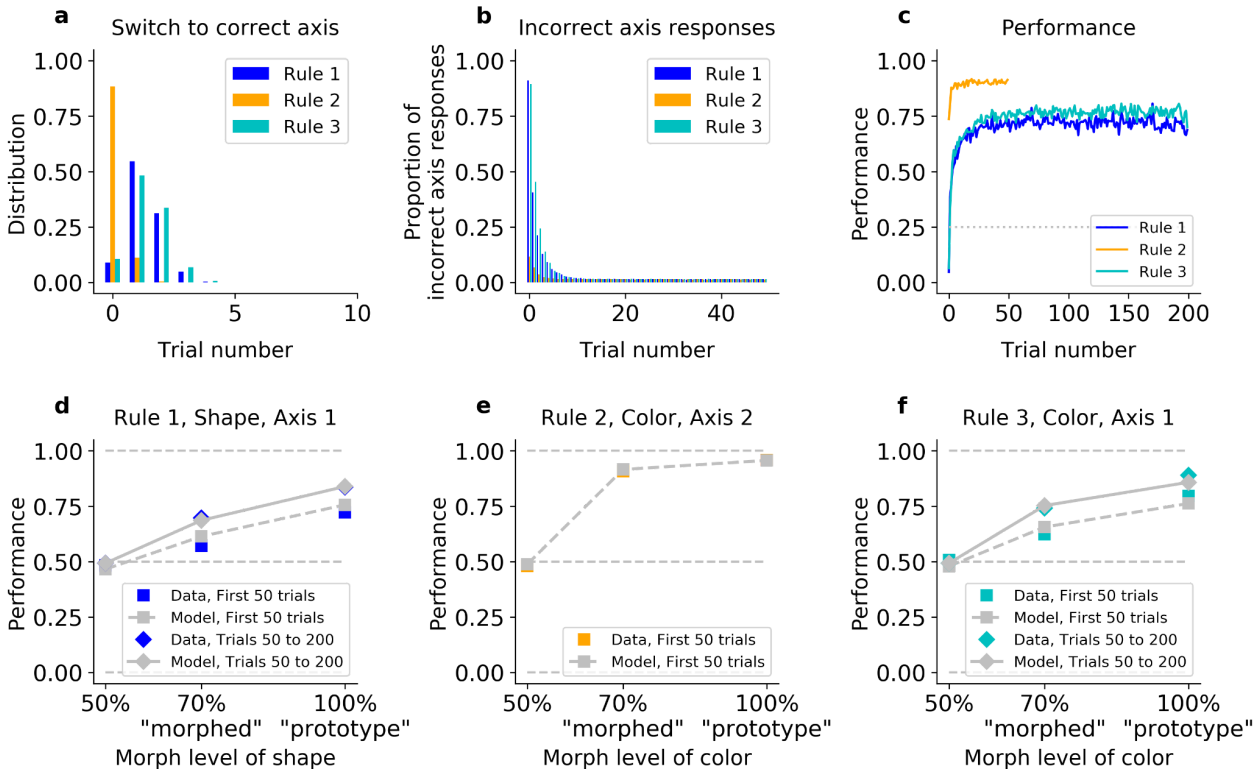


Figure 4: The hybrid learner (HQL) accounts both for fast switching to the correct axis, and slow relearning of Rule 1 and Rule 3. Model fit on Monkey S, see Fig. S5 for Monkey C. (a) Trial number for the first response on the correct axis after a block switch, for the model (compare to Fig. 1e inset). (b) Proportion of responses on the incorrect axis for the first 50 trials of each block, for the model (compare to Fig. 1e). (c) Performance of the model for the three rules (compare to Fig. 1g). (d,e,f) Performance for Rule 1, Rule 2 and Rule 3, as a function of the morphed version of the relevant feature.

The effect of stimulus congruency (and incongruency) provides further evidence for the hybrid model

To further understand how the HQL model outperforms the QL and IO models, we examined the animal's behavioral performance as a function of the relevant and irrelevant stimulus features. The orthogonal nature of the features and rules meant that stimuli could fall into two general groups. Congruent stimuli had features that required the same response for both Rule 1 and Rule 3 (e.g., a green bunny, Fig. 1) while incongruent stimuli had features that required opposite

responses between the two rules (e.g., a red bunny). Consistent with previous work[55–59], the animals performed better on congruent stimuli than incongruent stimuli (Fig. 5a for Monkey S, Fig. S7a for Monkey C). This effect was strongest during learning, but persisted throughout the block (Fig. S8a,e): during early trials of Rules 1 and 3, the monkeys' performance was significantly higher for congruent stimuli than for incongruent stimuli (gray vs. red squares in Fig. 5b; 94%, CI=[0.93,0.95] versus 57%, CI=[0.55,0.58] respectively ; with Δ=37 ; Fisher test p<10(-4); see Fig. S7b for Monkey C). Similarly, the animals were slower to respond to incongruent stimuli (Fig. S9, Δ=25ms between incongruent and congruent, t-test p<10(-4)). In contrast, the congruency of stimuli had no effect during Rule 2 – behavior depended only on the stimulus color, suggesting the monkeys ignored the shape of the stimulus during Rule 2, even when the morph level of the color was more difficult (gray vs. red squares in Fig 5c; performance was 92%, CI=[0.90,0.93], and 93%, CI=[0.92,0.93] for congruent and incongruent stimuli, respectively; with Δ=-0.73; Fisher test p=0.40 ; see Fig. S7c for Monkey C).

This incongruency effect provided further evidence for the HQL model. First, pure incremental learning by the QL model did not capture this result, but instead predicted an opposite effect. This is because incongruent trials were four times more likely than congruent trials (see Methods). As the QL model encodes the statistics of the task through error-driven updating of action values, the proportion of congruent/incongruent trials led to an anti-incongruency effect – the QL model fit to Monkey S predicted worse performance on congruent than incongruent trials (Fig. 5d,e ; 45% and 62%, respectively; Δ=-16 ; Fisher test p<10(-4) ; see Fig. S7d,e for Monkey C). Furthermore, for the same reason, the QL model produced a difference in performance during Rule 2 (Fig. 5f ; 54% for congruent versus 64% for incongruent ; Δ=-10 ; Fisher test p<10(-4); see Fig. S7f for Monkey C, see also Fig. S8b,f for this effect throughout the block).

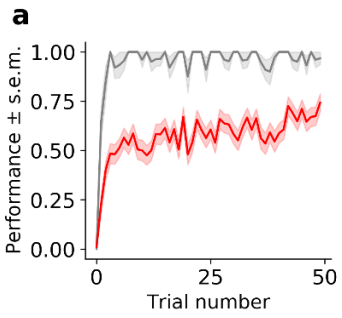
Second, the IO model also did not capture the incongruency effect. In principle, incongruency effects can be seen in this type of model when perceptual noise is large, because incongruent stimuli are more ambiguous when the correct rule is not yet known. But for the same reason this model could not explain slow learning of Rule 3, given the level of perceptual noise implied by asymptotic performance (Fig. 3f), it was able to rapidly determine which rule was in effect and execute it accurately, even for incongruent trials. Learning quickly reached a low asymptotic

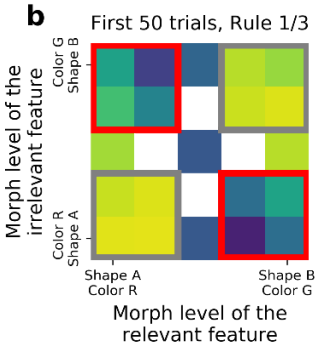
performance, for both congruent and incongruent trials (Fig. 5g,h ; 75% and 72% respectively ; Δ=3.8 only ; Fig. S7g,h for Monkey C).

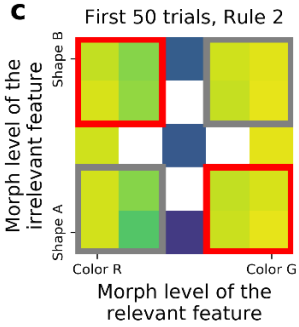
In contrast to the QL and IO models, the hybrid HQL model captured the incongruency effect. As the weights for congruent stimuli were the same for both Rules 1 and 3, the animals' performance was immediately high on those stimuli, while the associations for incongruent stimuli had to be relearned on each block (Fig. S10). The model fitted to Monkey S behavior reproduced the greater performance on congruent than incongruent stimuli (Fig. 5g,h ; 92% and 61%, respectively; Δ=31 ; see Fig. S7g,h for Monkey C). As with the monkey's behavior, this effect persisted throughout the block (Fig. S8d,h). Finally, the HQL model captured the absence of incongruency effect in Rule 2 (Fig. 5i, green versus red squares ; 92% and 93%, respectively; Δ=-1.0 ; see Fig. S7i for Monkey C), as there was no need to update the Axes 2 weights between blocks.

As a result, only a hybrid model performing simultaneously both rule switching of axis and rule-learning of features could account for the incongruency effect observed in the behavior.

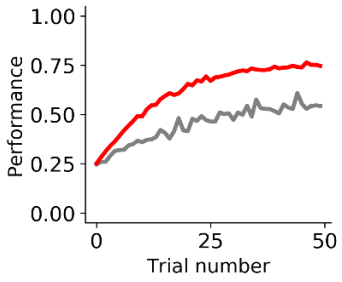
Monkey data (S)

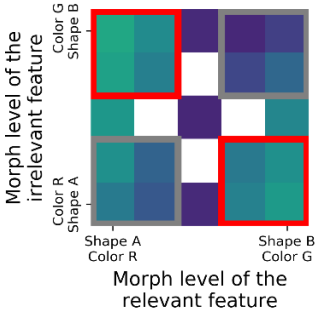


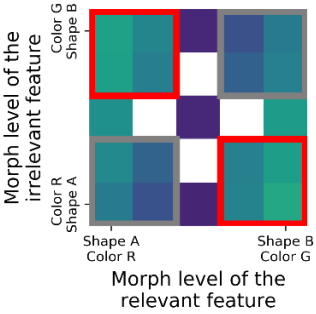




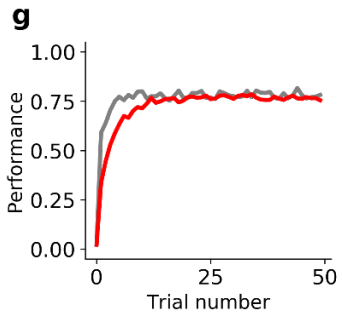
QI learner

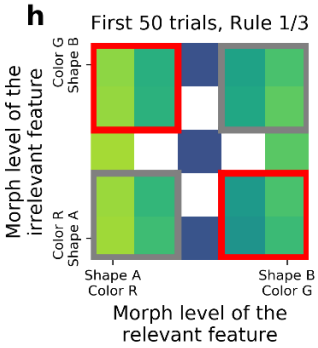


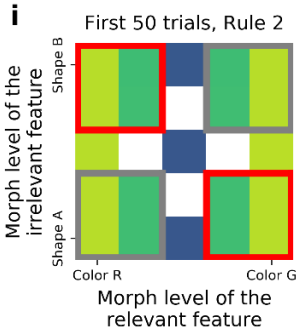




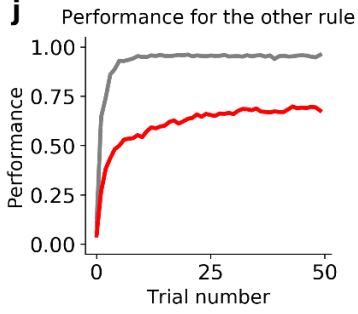
Ideal observer

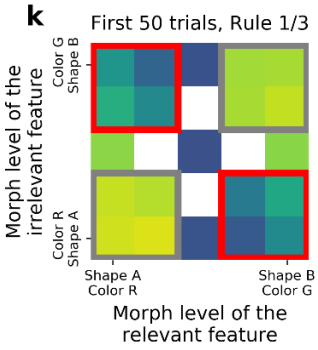






Hybrid learner





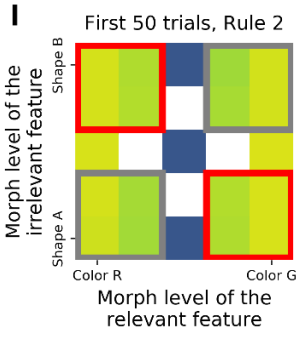


Figure 5: Comparison of incongruency effects in Monkey S and behavioral models (QL, IO, and HQL models). (a) Performance as a function of trial number for Rule 1 and Rule 3 (combined), for congruent and incongruent trials. (b) Performance for Rule 1 and 3 (combined, first 50 trials), as a function of the morph level for both color (relevant) and shape (irrelevant) features. Grey boxes highlight congruent stimuli, red boxes highlight incongruent stimuli. (c) Performance for Rule 2, as a function of the morph level for both color (relevant) and shape (irrelevant) features. Note the lack of an incongruency effect. (d,e,f) Same as a-c but for the QL model. (g,h,i) Same as a-c for the IO model. (j,k,l) Same as a-c but for the HQL model.

Discussion

In the present study, we investigated rule learning in two monkeys trained to switch between three category-response tasks. Critically, the animals were only informed of when the rule switched but had to learn which new rule was in effect. We compared two classes of models that were able to perform the task: incremental learning and inferential rule switching. Our results suggested that neither model fit the animals' performance well. Incremental learning was too slow to capture the monkeys' rapid learning of the response axis after a block switch. It was also unable to explain the immediately high behavioral performance on Rule 2, which was the only rule requiring responses along the second axis. Inference learning was unable to reproduce the difference in performance for two rules that required attending to the same feature of the stimulus, but responding on different axes (Rule 2 and Rule 3). Finally, neither of these two classes of models considered separately could explain the monkeys' difficulty for incongruent stimuli across rules that required a response on the same axes (Rule 1 and Rule 3). Instead, we found that a hybrid model that inferred axes quickly and relearned features slowly, was able to capture the monkeys' behavior. This suggests the animals were learning the current axis of response using fast inference while re-estimating continuously the stimulus-response mappings within an axis.

The superior explanatory power of the hybrid model suggests that the task induces animals to perpetually exercise both rule switching and rule learning – even in a well-trained regime in which they could, in principle, have discovered perfect rule knowledge. The model suggests that

they instead must perpetually relearn Rules 1 and 3, because they appear to be working with only two latent states (corresponding to the two axes of response) instead of the three rules we designed. These two latent states effectively encode Rule 2 (alone on its response axis) on one hand, and a combination of Rule 1 and Rule 3 (sharing a response axis) on the other hand. Within the second latent state, the monkeys continuously updated their knowledge of the rules' contingencies (different stimulus features to action mappings). Why animals fail to discover the correct rule structure (which would clearly support better performance in Rule 1 and 3 blocks) remains a question, but presumably reflects the brain's mechanisms for discovering, splitting, or differentiating different latent states on the basis of their differing stimulus-action-response contingencies. Clearly, the overlap between Rules 1 and 3 (sharing an axis of response) makes them harder to differentiate than either from Rule 2: for this, the axis is the most discriminatory feature (being discrete and also under the monkey's own explicit control) whereas the stimulus-reward mappings are noisier. Also, perhaps a two-latent state regime was resource-rational in this task, considering additional limitations relating to the cost of control (stability-flexibility trade-off) or working memory during training for multitasking [70,72,73]. Although more difficult, perhaps training the monkeys on the four possible rules permissible by the experimental design (and thus adding Rule 4, sharing the axis of Rule 2 but using the feature shape, cf. missing rule in Fig. 1d) would have forced encoding latent states from implicit information, including stimuli features, not restricted by the dimensionality of motor responses. Finally, with a different training protocol, the monkeys may have eventually encoded Rule 1 and Rule 3 as separate latent states (e.g., longer training, or with a higher ratio of incongruent versus congruent trials, or with less morphed and more prototyped stimuli).

Finally, our characterization of the computational contributions of rule switching and rule learning, and the ability to observe both interacting in a single task, leads to a number of testable predictions about their neural interactions. First, our results predict that there should be two latent states represented in the brain. This makes the prediction that the neural representation for the two rules competing on one axis (Rules 1 and 3) should be more similar to one another than to the neural representation of the rule alone on the other axis (Rule 2). This would not be the case if the neural activity was instead representing three latent causes. Furthermore, future neural work may be able to discriminate between cortical and sub-cortical networks for rule switching

and rule learning. Our hybrid model suggests there may be a functional dissociation for rule switching and rule learning, that may be represented in distinct networks. One hypothesis is that prefrontal cortex may carry information about the animal's trial beliefs (i.e. over the two latent states) in a similar manner as perceptual decision making when accumulating evidence from noisy stimuli[60–63]. Basal ganglia may, in turn, be engaged in the learning of rule-specific associations. Alternatively, despite their functional dissociation, future work may find both rule switching and rule learning are represented in the same brain regions (e.g., prefrontal cortex). Finally, inference and incremental learning may be distribution functions, requiring the cooperation of multiple brain regions.

Acknowledgments

This research was supported by U.S. Army Research Office ARO W911NF-16-1-047 (ND).

Author contributions

Flora Bouchacourt, Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing original draft, Writing reviewing and editing, Visualization.
Sina Tafazoli, Conceptualization, Methodology, Validation, Investigation, Data Curation.
Marcelo Mattar, Conceptualization, Methodology, Validation, Writing reviewing and editing.
Tim Buschman, Conceptualization, Methodology, Validation, Resources, Data Curation, Writing reviewing and editing, Supervision, Project administration, Funding acquisition.
Nathaniel Daw, Conceptualization, Methodology, Validation, Writing reviewing and editing, Supervision, Project administration, Funding acquisition.

Declaration of interests

The authors declare no competing interests.

References

1. Daw, N. D. & O'Doherty, J. P. Chapter 21 - Multiple Systems for Value Learning. in *Neuroeconomics (Second Edition)* (eds. Glimcher, P. W. & Fehr, E.) 393–410 (Academic Press, 2014). doi:10.1016/B978-0-12-416008-8.00021-8.

2. Daw, N. D. & Shohamy, D. The Cognitive Neuroscience of Motivation and Learning. *Soc. Cogn.* 26, 593–620 (2008).

3. Daw, N. D. & Tobler, P. N. Chapter 15 - Value Learning through Reinforcement: The Basics of Dopamine and Reinforcement Learning. in *Neuroeconomics (Second Edition)* (eds. Glimcher, P. W. & Fehr, E.) 283–298 (Academic Press, 2014). doi:10.1016/B978-0-12-416008-8.00015-2.

4. Dolan, R. J. & Dayan, P. Goals and Habits in the Brain. *Neuron* 80, 312–325 (2013).

5. Doya, K. Reinforcement learning: Computational theory and biological mechanisms. *HFSP J.* 1, 30–40 (2007).

6. O'Doherty, J. *et al.* Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* 304, 452–454 (2004).

7. O'Reilly, R. C. & Frank, M. J. Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Comput.* 18, 283–328 (2006).

8. Rescorla, R. A. Pavlovian conditioning: It's not what you think it is. *Am. Psychol.* 43, 151–160 (1988).

9. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* 275, 1593–1599 (1997).

10. Yin, H. H. & Knowlton, B. J. The role of the basal ganglia in habit formation. *Nat. Rev. Neurosci.* 7, 464–476 (2006).

11. Wang, J. X. *et al.* Prefrontal cortex as a meta-reinforcement learning system. *Nat. Neurosci.* 21, 860–868 (2018).

12. Asaad, W. F., Rainer, G. & Miller, E. K. Task-Specific Neural Activity in the Primate Prefrontal Cortex. *J. Neurophysiol.* 84, 451–459 (2000).

13. Asaad, W. F., Rainer, G. & Miller, E. K. Neural Activity in the Primate Prefrontal Cortex during Associative Learning. *Neuron* 21, 1399–1407 (1998).

14. Rougier, N. P., Noelle, D. C., Braver, T. S., Cohen, J. D. & O'Reilly, R. C. Prefrontal cortex and flexible cognitive control: Rules without symbols. *Proc. Natl. Acad. Sci.* 102, 7338–7343 (2005).

15. Sarafyazd, M. & Jazayeri, M. Hierarchical reasoning by neural circuits in the frontal cortex. *Science* 364, eaav8911 (2019).

16. Collins, A. & Koechlin, E. Reasoning, Learning, and Creativity: Frontal Lobe Function and Human Decision-Making. *PLOS Biol.* 10, e1001293 (2012).

17. Behrens, T. E. J., Woolrich, M. W., Walton, M. E. & Rushworth, M. F. S. Learning the value of information in an uncertain world. *Nat. Neurosci.* 10, 1214–1221 (2007).

18. Gershman, S. J., Radulescu, A., Norman, K. A. & Niv, Y. Statistical Computations Underlying the Dynamics of Memory Updating. *PLOS Comput. Biol.* 10, e1003939 (2014).

19. Durstewitz, D., Vittoz, N. M., Floresco, S. B. & Seamans, J. K. Abrupt Transitions between Prefrontal Neural Ensemble States Accompany Behavioral Transitions during Rule Learning. *Neuron* 66, 438–448 (2010).

20. Milner, B. Effects of Different Brain Lesions on Card Sorting: The Role of the Frontal Lobes. *Arch. Neurol.* 9, 90–100 (1963).

21. Boettiger, C. A. & D'Esposito, M. Frontal networks for learning and executing arbitrary stimulus-response associations. *J. Neurosci. Off. J. Soc. Neurosci.* 25, 2723–2732 (2005).

22. Nakahara, K., Hayashi, T., Konishi, S. & Miyashita, Y. Functional MRI of Macaque Monkeys Performing a Cognitive Set-Shifting Task. *Science* (2002) doi:10.1126/science.1067653.

23. Genovesio, A., Brasted, P. J., Mitz, A. R. & Wise, S. P. Prefrontal Cortex Activity Related to Abstract Response Strategies. *Neuron* 47, 307–320 (2005).

24. Boorman, E. D., Behrens, T. E. J., Woolrich, M. W. & Rushworth, M. F. S. How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. *Neuron* 62, 733–743 (2009).

25. Koechlin, E. & Hyafil, A. Anterior prefrontal function and the limits of human decision-making. *Science* 318, 594–598 (2007).

26. Koechlin, E., Ody, C. & Kouneiher, F. The architecture of cognitive control in the human prefrontal cortex. *Science* 302, 1181–1185 (2003).

27. Sakai, K. & Passingham, R. E. Prefrontal interactions reflect future task operations. *Nat. Neurosci.* 6, 75–81 (2003).

28. Badre, D., Kayser, A. S. & D'Esposito, M. Frontal cortex and the discovery of abstract action rules. *Neuron* 66, 315–326 (2010).

29. Miller, E. K. & Cohen, J. D. An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.*

24, 167–202 (2001).

30. Mansouri, F. A., Freedman, D. J. & Buckley, M. J. Emergence of abstract rules in the primate brain. *Nat. Rev. Neurosci.* 21, 595–610 (2020).

31. White, I. M. & Wise, S. P. Rule-dependent neuronal activity in the prefrontal cortex. *Exp. Brain Res.* 126, 315–335 (1999).

32. Collins, A. G. E. & Frank, M. J. Neural signature of hierarchically structured expectations predicts clustering and transfer of rule sets in reinforcement learning. *Cognition* 152, 160–169 (2016).

33. Schuck, N. W., Cai, M. B., Wilson, R. C. & Niv, Y. Human Orbitofrontal Cortex Represents a Cognitive Map of State Space. *Neuron* 91, 1402–1412 (2016).

34. Chan, S. C. Y., Niv, Y. & Norman, K. A. A Probability Distribution over Latent Causes, in the Orbitofrontal Cortex. *J. Neurosci.* 36, 7817–7828 (2016).

35. Hampton, A. N., Bossaerts, P. & O'Doherty, J. P. The Role of the Ventromedial Prefrontal Cortex in Abstract State-Based Inference during Decision Making in Humans. *J. Neurosci.* 26, 8360–8367 (2006).

36. Frank, M. J. & Badre, D. Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: computational analysis. *Cereb. Cortex N. Y. N 1991* 22, 509–526 (2012).

37. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* 503, 78–84 (2013).

38. Song, H. F., Yang, G. R. & Wang, X.-J. Reward-based training of recurrent neural networks for cognitive and value-based tasks. *eLife* 6, e21492 (2017).

39. Schweighofer, N. & Doya, K. Meta-learning in Reinforcement Learning. *Neural Netw.* 16, 5–9 (2003).

40. Duan, Y. *et al.* RL^2: Fast Reinforcement Learning via Slow Reinforcement Learning. *ArXiv161102779 Cs Stat* (2016).

41. Donoso, M., Collins, A. G. E. & Koechlin, E. Foundations of human reasoning in the prefrontal cortex. *Science* (2014) doi:10.1126/science.1252254.

42. Badre, D. & Frank, M. J. Mechanisms of hierarchical reinforcement learning in cortico-striatal circuits 2: evidence from fMRI. *Cereb. Cortex N. Y. N 1991* 22, 527–536 (2012).

43. Bouchacourt, F., Palminteri, S., Koechlin, E. & Ostojic, S. Temporal chunking as a mechanism for unsupervised learning of task-sets. *eLife* 9, e50469 (2020).

44. Franklin, N. T. & Frank, M. J. Compositional clustering in task structure learning. *PLOS Comput. Biol.* 14, e1006116 (2018).

45. Lak, A. *et al.* Reinforcement biases subsequent perceptual decisions when confidence is low, a widespread behavioral phenomenon. *eLife* 9, e49834 (2020).

46. Busse, L. *et al.* The Detection of Visual Contrast in the Behaving Mouse. *J. Neurosci.* 31, 11351–11361 (2011).

47. Fründ, I., Wichmann, F. A. & Macke, J. H. Quantifying the effect of intertrial dependence on perceptual decisions. *J. Vis.* 14, 9 (2014).

48. Gold, J. I., Law, C.-T., Connolly, P. & Bennur, S. The Relative Influences of Priors and Sensory Evidence on an Oculomotor Decision Variable During Perceptual Learning. *J. Neurophysiol.* 100, 2653–2668 (2008).

49. Tsunada, J., Cohen, Y. & Gold, J. I. Post-decision processing in primate prefrontal cortex influences subsequent choices on an auditory decision-making task. *eLife* 8, e46770 (2019).

50. Dayan, P. & Daw, N. D. Decision theory, reinforcement learning, and the brain. *Cogn. Affect. Behav. Neurosci.* 8, 429–453 (2008).

51. Pouget, A., Drugowitsch, J. & Kepecs, A. Confidence and certainty: distinct probabilistic quantities for different goals. *Nat. Neurosci.* 19, 366–374 (2016).

52. Pouget, A., Beck, J. M., Ma, W. J. & Latham, P. E. Probabilistic brains: knowns and unknowns. *Nat. Neurosci.* 16, 1170–1178 (2013).

53. Gold, J. I. & Shadlen, M. N. Neural computations that underlie decisions about sensory stimuli. *Trends Cogn. Sci.* 5, 10–16 (2001).

54. Bichot, N. P. & Schall, J. D. Effects of similarity and history on neural mechanisms of visual selection. *Nat. Neurosci.* 2, 549–554 (1999).

55. Noppeney, U., Ostwald, D. & Werner, S. Perceptual Decisions Formed by Accumulation of Audiovisual Evidence in Prefrontal Cortex. *J. Neurosci.* 30, 7434–7446 (2010).

56. Venkatraman, V., Rosati, A. G., Taren, A. A. & Huettel, S. A. Resolving Response, Decision, and Strategic Control: Evidence for a Functional Topography in Dorsomedial Prefrontal Cortex. *J. Neurosci.* 29, 13158–13164 (2009).

57. Bugg, J. M., Jacoby, L. L. & Toth, J. P. Multiple levels of control in the Stroop task. *Mem. Cognit.* 36, 1484–1494 (2008).

58. Carter, C. S., Mintun, M. & Cohen, J. D. Interference and Facilitation Effects during Selective

Attention: An H215O PET Study of Stroop Task Performance. *NeuroImage* 2, 264–272 (1995).

59. Musslick, S. & Cohen, J. D. Rationalizing constraints on the capacity for cognitive control. *Trends Cogn. Sci.* 25, 757–775 (2021).

60. Gold, J. I. & Shadlen, M. N. The neural basis of decision making. *Annu. Rev. Neurosci.* 30, 535–574 (2007).

61. Shadlen, M. N. & Kiani, R. Decision Making as a Window on Cognition. *Neuron* 80, 791–806 (2013).

62. Rao, R. P. N. Decision Making Under Uncertainty: A Neural Model Based on Partially Observable Markov Decision Processes. *Front. Comput. Neurosci.* 4, 146 (2010).

63. Beck, J. M. *et al.* Probabilistic Population Codes for Bayesian Decision Making. *Neuron* 60, 1142–1152 (2008).

Supplementary figures

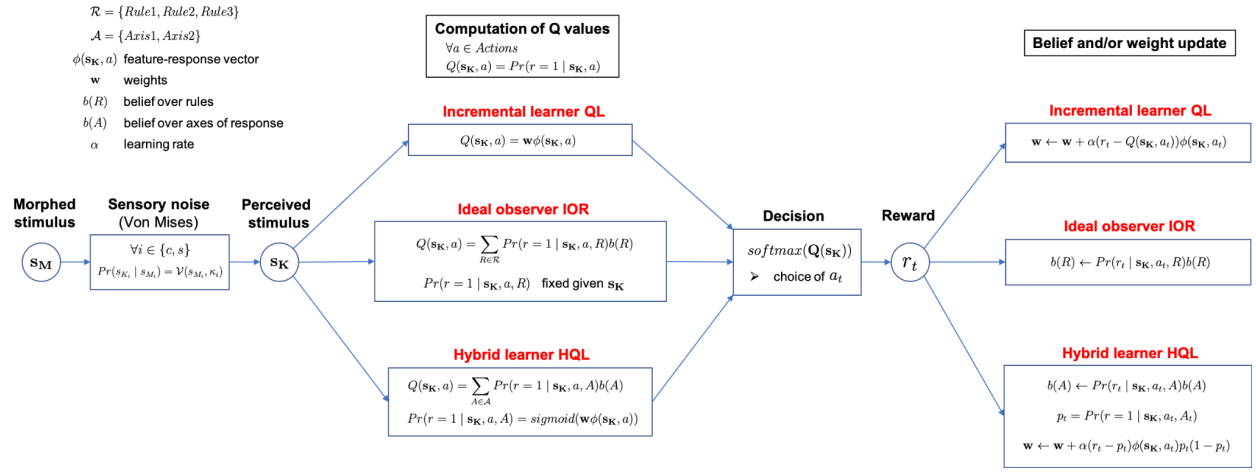


Figure S1: the 3 models.

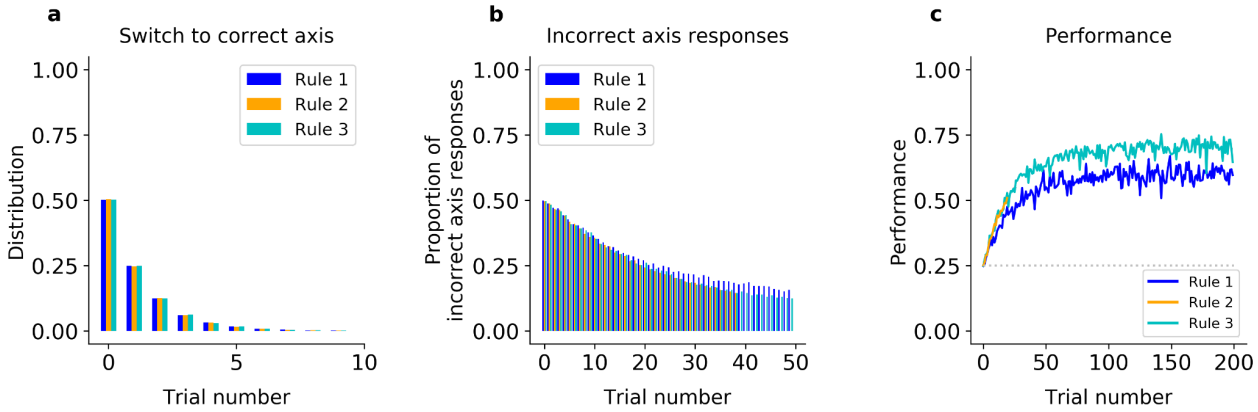


Figure S2: Incremental learner (QL) model fitted on Monkey C behavior (see Fig. 2 for Monkey S). (a) Trial number of the first response of the model on the correct axis after a block switch (compare to Fig. 1f, inset). (b) Proportion of responses of the model on the incorrect axis for the first 50 trials of each block (compare to Fig. 1f). (c) Model performance for each rule (averaged over blocks, compare to Fig. 1h). Statistics of QL model fitted on Monkey C: First, the model made a response on the correct axis on the first trial with a probability of only 50% in Rule 1, Rule 2 and Rule 3. The model performed 28% of off-axis responses after 20 trials in Rule 1, and 25% in Rule 2 and Rule 3. Second, the model performed correctly on the first trial in only 25% of Rule 2 blocks, and reached only 51% after 20 trials. As a result, on the first 20 trials, the difference in average percent performances was only Δ=3.3 between Rule 2 and Rule 1, and was only Δ=-0.39 between Rule 2 and Rule 3.

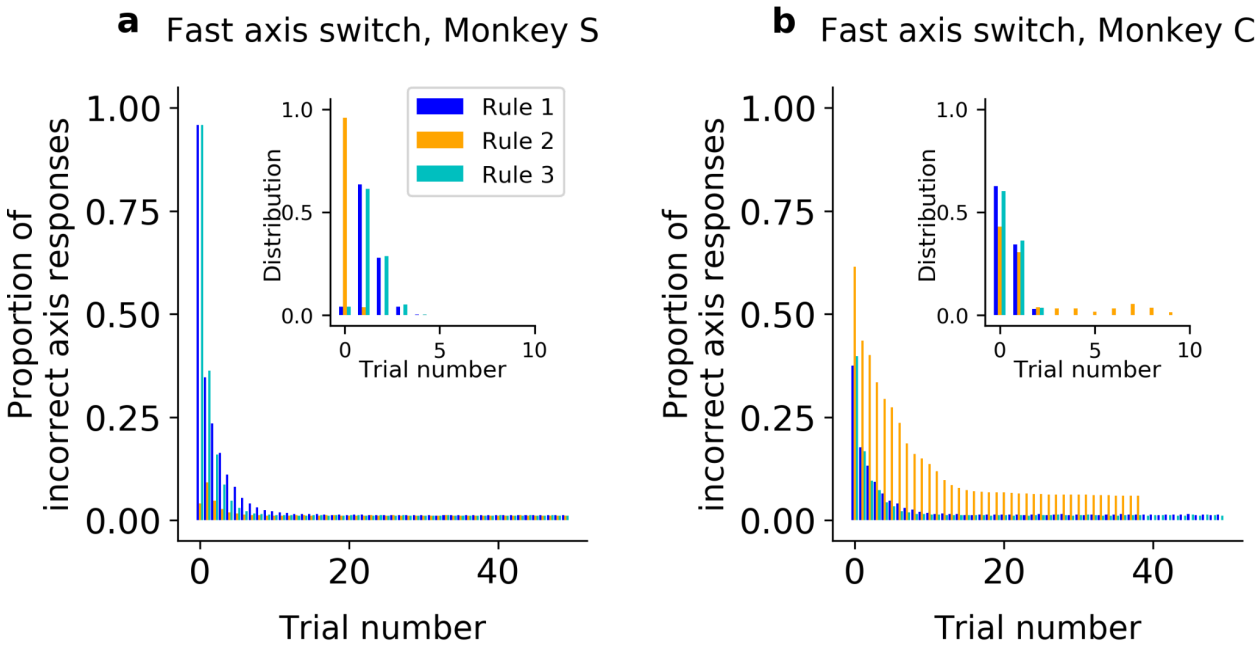


Figure S3: Proportion of responses on the incorrect axis for the first 50 trials of each block for (a) IO model fitted on Monkey S and (b) IO model fitted on Monkey C. Insets: Trial number of the first response on the correct axis after a block switch. Compare to Fig. 1e-h. Statistics of IO model fitted on Monkey C: The model made a response on the correct axis on the first trial with a probability of 63% in Rule 1, 38% in Rule 2 and 60% in Rule 3. The model maintained the correct axis with very few off-axis responses throughout the block (after trial 20, 1.3% in Rule 1 ; 6.8%, in Rule 2 ; 1.3% in Rule 3, Fisher test against monkey behavior: $p>0.5$ in all rules).

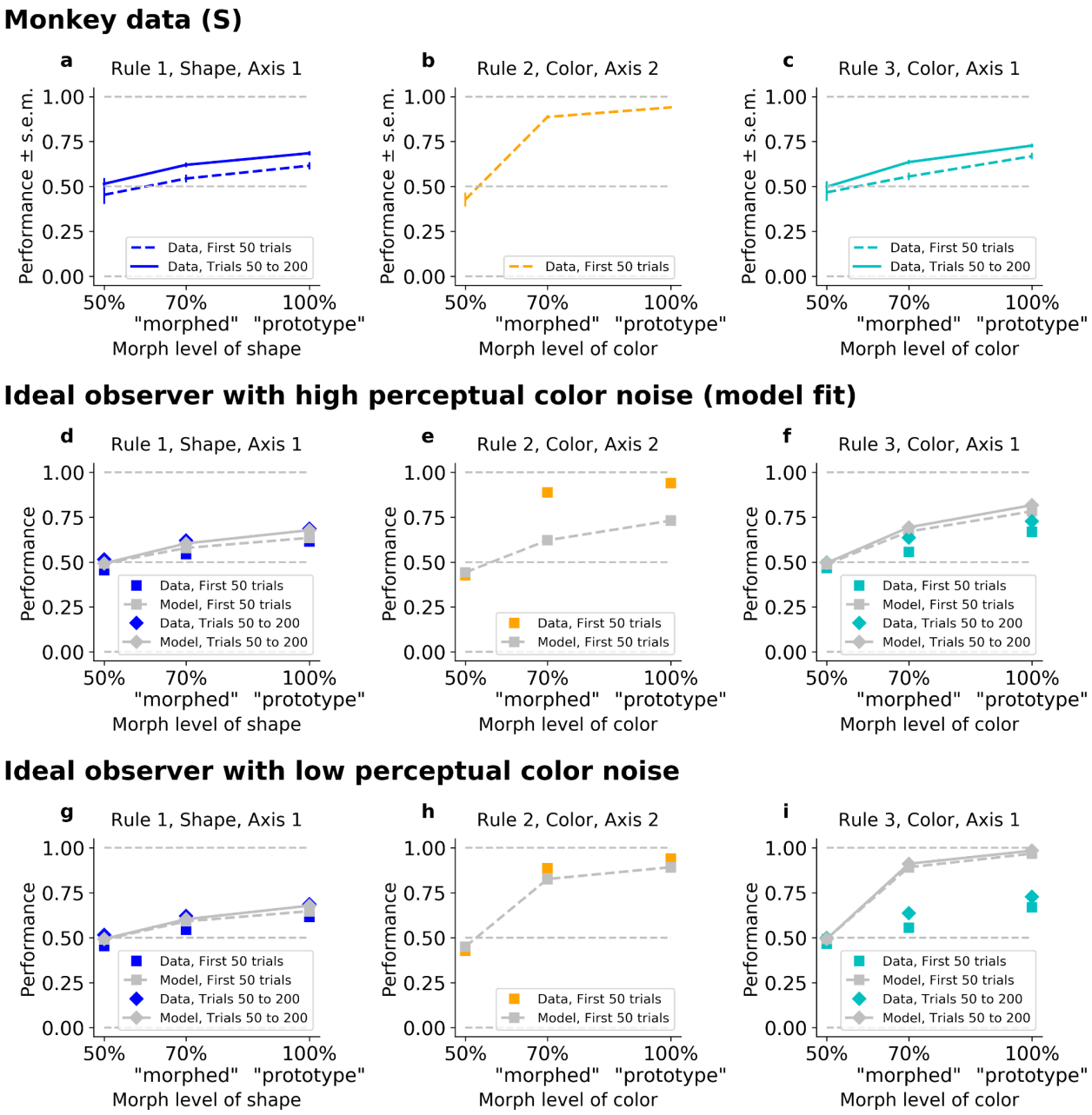


Figure S4: The ideal observer (IO), slow or fast, but not both. Fitted on Monkey C behavior (see Fig. 3 for Monkey C). (a-b-c) Performance for Rule 1, 2, and 3, as a function of the morphed version of the relevant feature. (d,e,f) Performance for Rule 1, 2, and 3, for IO model with high color noise. This parameter regime corresponds to the case where the model is fitted to the monkey's behavior (see Methods). (j,k,l) Performance for Rule 1, 2, and 3, for IO model with low color noise. Here, we fixed KC=6. Statistics on Monkey C: There was a discrepancy between the performance of 'morphed' stimuli in Rule 2 versus Rule 3, with a difference in

average percent performances of Δ=33 for the first 50 trials in both rules (p<10(-4)), and still Δ=24 if we considered Rule 2 against the last trials of Rule 3 (p<10(-4)). The same discrepancy was observed between the performance of 'prototype' stimuli in Rule 2 versus Rule 3, with a difference in average percent performances of Δ=27 for the first 50 trials in both rules (p<10(-4)), and still Δ=21 if we considered the last trials of Rule 3 (p<10(-4)). Statistics of IO model fitted on Monkey C: While the IO model, using best-fit parameters, reproduced poor asymptotic performance in Rule 3 by increasing color noise (low concentration), it then failed to capture the high performance on Rule 2 early on. The resulting difference in performance for 'morphed' stimuli was only Δ=4.8 for the first 50 trials and Δ=-7.2 if we considered the last trials of Rule 3 (respectively Δ=5.1 and Δ=-8.6 for 'prototype').

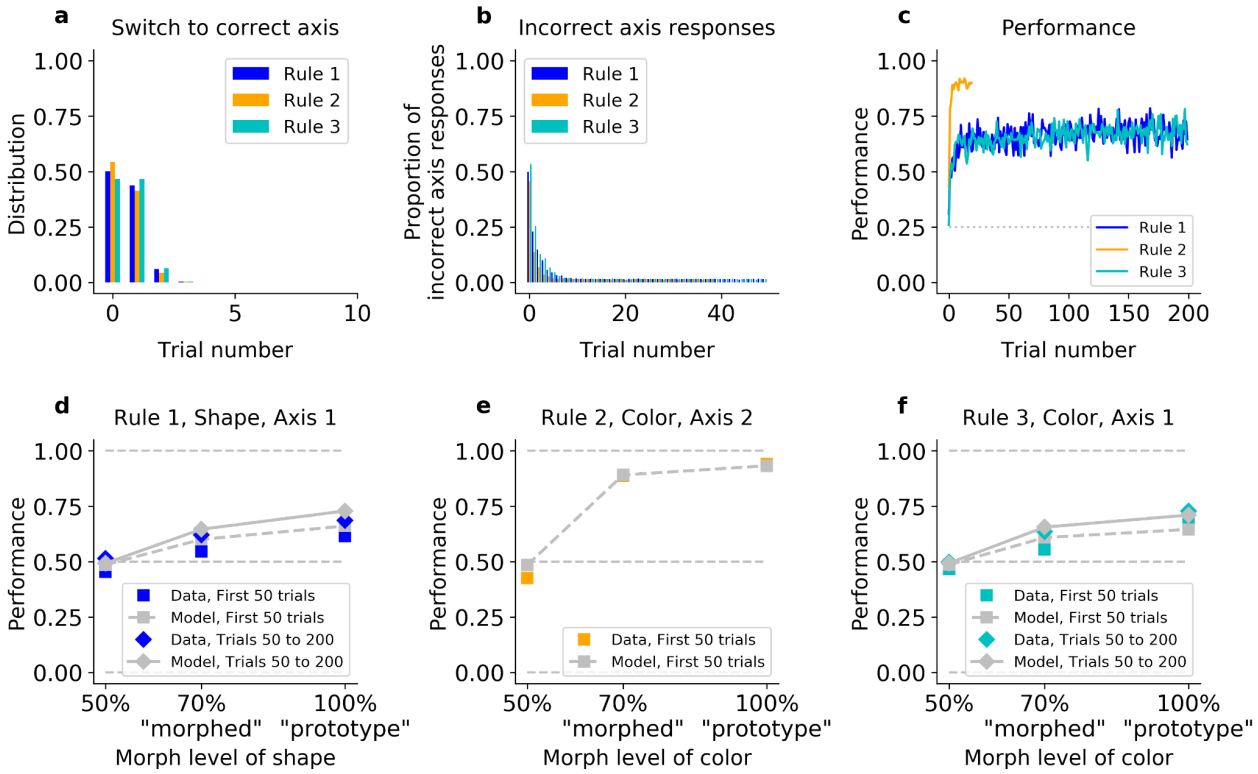


Figure S5: The hybrid learner (HQL) accounts both for fast switching to the correct axis, and slow relearning of Rule 1 and Rule 3. Model fit on Monkey C, see Fig. 4 for Monkey S. (a) Trial number for the first response on the correct axis after a block switch, for the model (compare to

Fig. 1f inset). (b) Proportion of responses on the incorrect axis for the first 50 trials of each block, for the model (compare to Fig. 1f). (c) Performance of the model for the three rules (compare to Fig. 1h). (d,e,f) Performance for Rule 1, Rule 2 and Rule 3, as a function of the morphed version of the relevant feature. Statistics of HQL model fitted on Monkey C: First, The model made a response on the correct axis on the first trial with a probability of 50% in Rule 1, 54% in Rule 2 and 47% in Rule 3. The model maintained the correct axis with very few off-axis responses throughout the block (after trial 20, 1.5% in Rule 1 ; 1.7%, in Rule 2 ; 1.5% in Rule 3, Fisher test against monkey's behavior: $p>0.5$ in all rules). Second, the HQL model could capture the animal's fast performance on Rule 2 and slower performance on Rules 1 and 3: the difference in average percent performances on the first 20 trials was $\Delta=28$ both between Rule 2 and Rule 1 and between Rule 2 and Rule 3. Third, not only the HQL model captured the performance ordering on morphed and prototype stimuli for each rule separately, but the model was able to trade-off between initial and asymptotic behavioral performance in Rule 2 and Rule 3, for both 'morphed' and 'prototype' stimuli. The resulting difference in performance for 'morphed' stimuli was $\Delta=29$ for the first 50 trials and $\Delta=24$ if we considered the last trials of Rule 3 (respectively $\Delta=29$ and $\Delta=22$ for 'prototype').

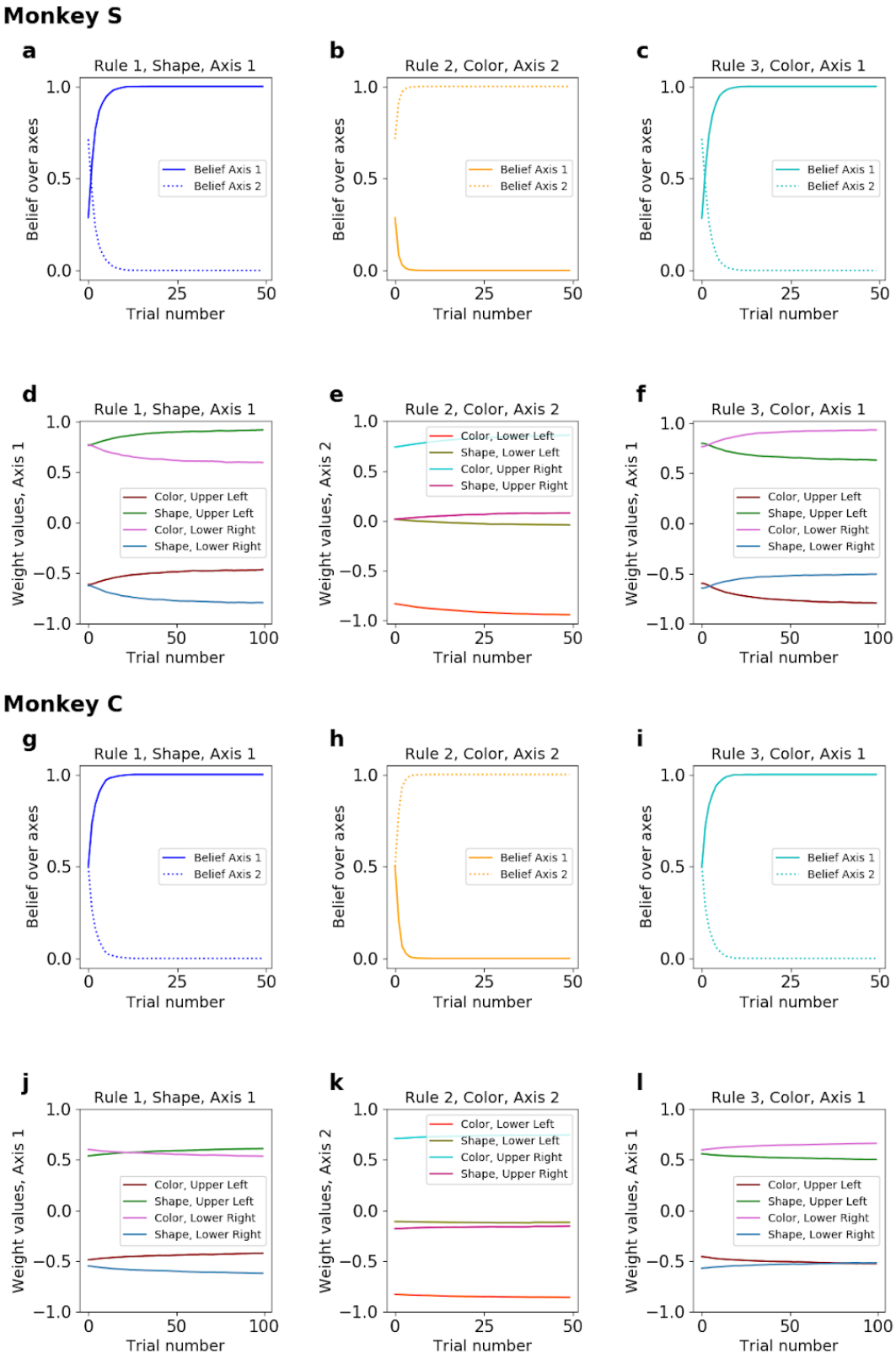
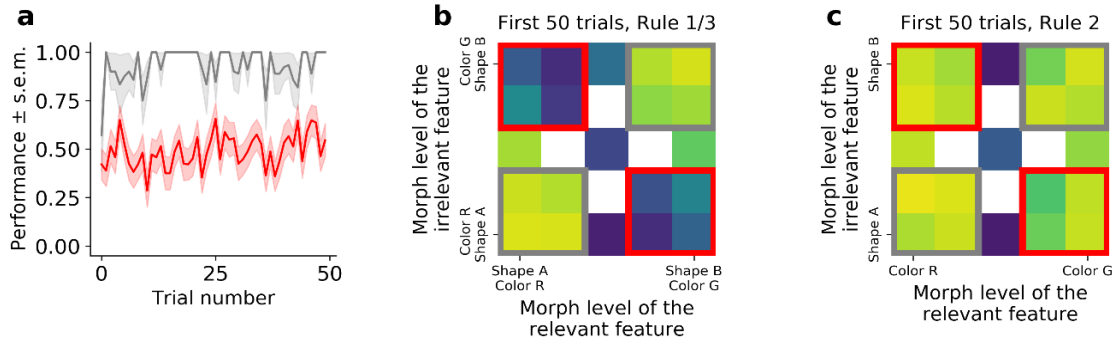
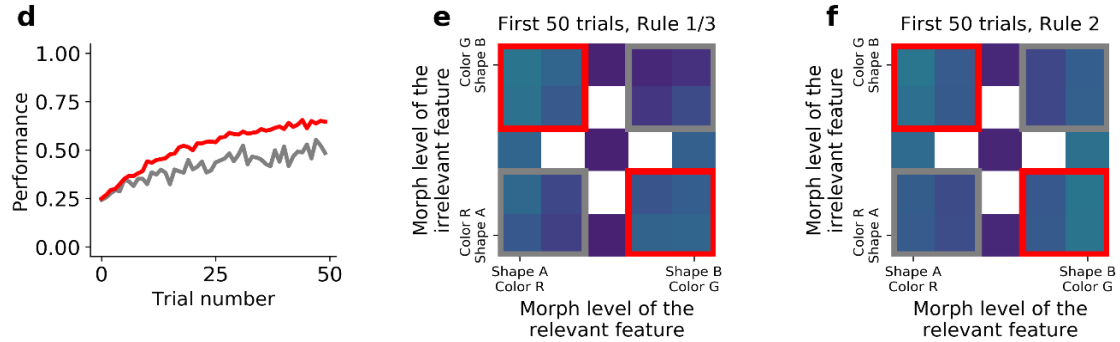


Figure S6: The hybrid learner fitted on Monkey S: belief over axes and feature weights values.

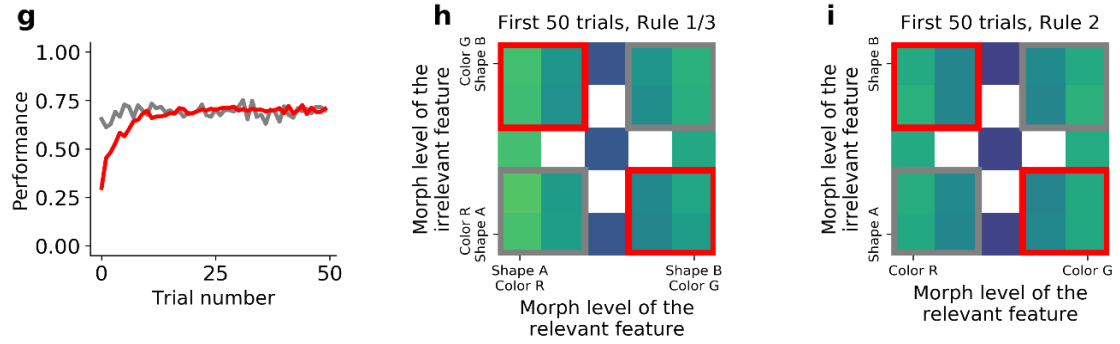
Monkey data (S)



QI learner



Ideal observer



Hybrid learner

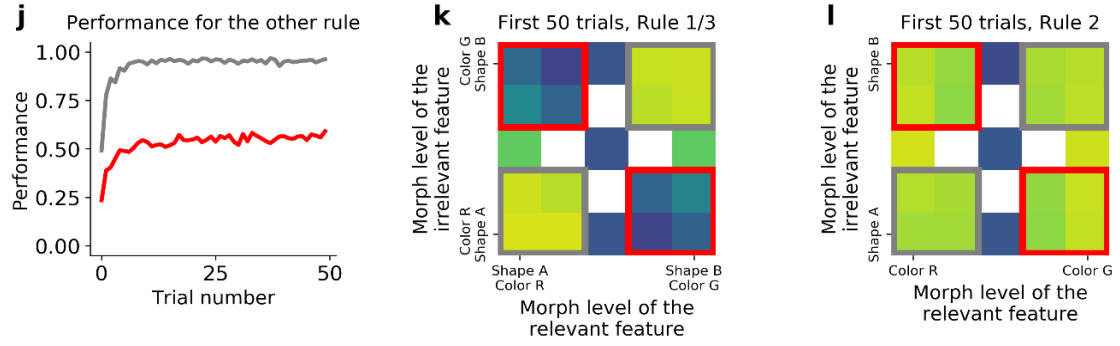


Figure S7: Comparison of incongruency effects in Monkey C and behavioral models (QL, IO and HQL models). (a) Performance as a function of trial number for Rule 1 and Rule 3 (combined), for congruent and incongruent trials. (b) Performance for Rule 1 and 3 (combined, first 50 trials), as a function of the morph level for both color (relevant) and shape (irrelevant) features. Grey boxes highlight congruent stimuli, red boxes highlight incongruent stimuli. (c) Performance for Rule 2, as a function of the morph level for both color (relevant) and shape (irrelevant) features. Note the lack of an incongruency effect. (d,e,f) Same as a-c but for the QL model. (g,h,i) Same as a-c for the IO model. (j,k,l) Same as a-c but for the HQL model. Statistics on Monkey C: During early trials of Rules 1 and 3, the monkeys' performance was significantly higher for congruent trials than for incongruent trials (gray vs. red squares ; 93%, CI=[0.91,0.95] versus 49%, CI=[0.47,0.51] respectively ; with Δ=44 ; Fisher test p<10(-4)). There was no difference in performance between congruent and incongruent stimuli during Rule 2 (grey vs. red squares ; performance was 94%, CI=[0.91,0.95], and 91%, CI=[0.89,0.92], respectively; with Δ=2.7 ; Fisher test p=0.07). Statistics of QL model fitted on Monkey C: The model performed worse on congruent than incongruent trials in Rule 1 and Rule 3 (41% and 52%, respectively; Δ=-10 ; Fisher test p<10(-4)), against our behavioral observations. Furthermore, the model produced a difference in performance during Rule 2 (48% for congruent versus 54% for incongruent ; Δ=-6.1 ; Fisher test p<10(-4)). Statistics of IO model fitted on Monkey C: Learning quickly reached a low asymptotic performance in Rule 1 and Rule 3, for both congruent and incongruent trials (69% and 67% respectively ; Δ=2.5 only). Statistics of HQL model fitted on Monkey C: The model reproduced the greater performance on congruent than incongruent stimuli in Rule 1 and Rule 3 (94% and 53%, respectively; Δ=41). It also captured the absence of incongruency effect in Rule 2 (green versus red squares ; 91% and 91%, respectively; Δ=0.081).

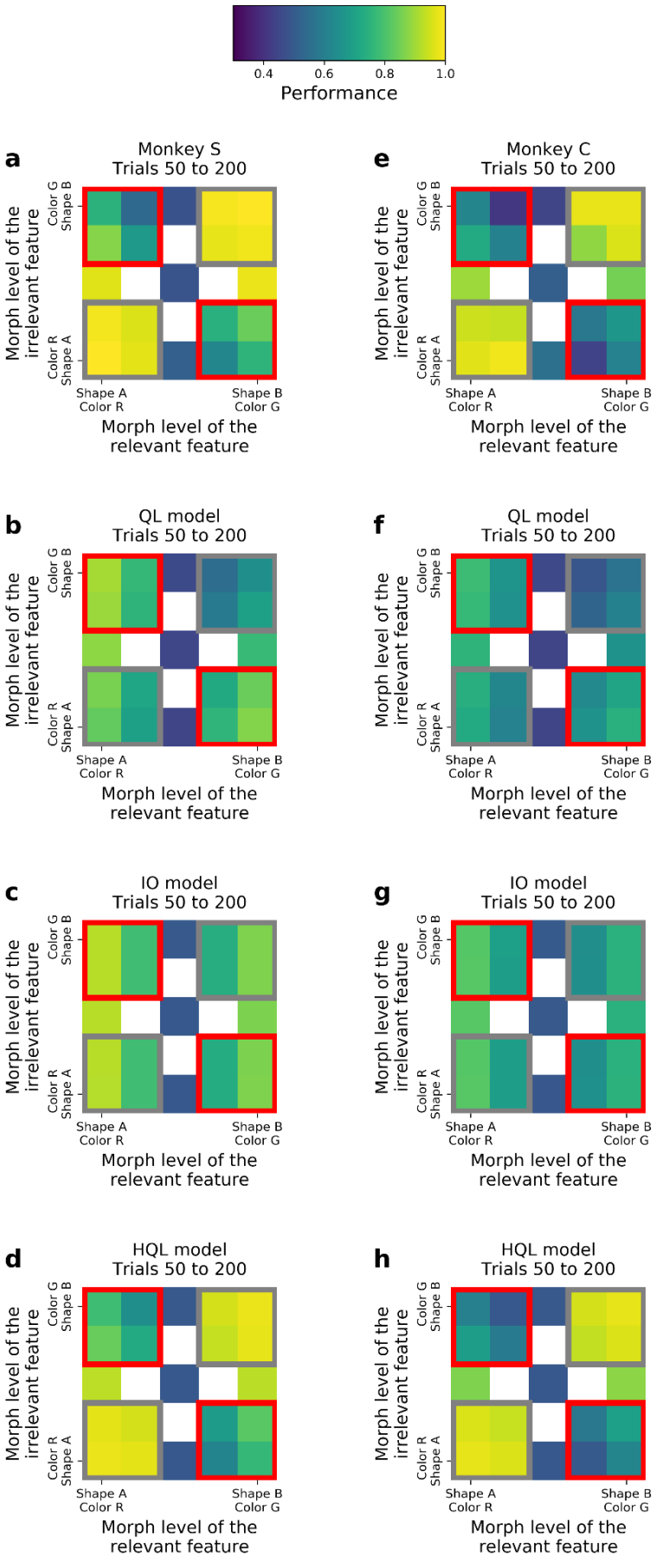


Figure S8: Incongruency effect for Monkey S (a), Monkey C (e) and models (respectively fitted on Monkey S: b-d ; and on Monkey C: f-h), for trials 50 to 200. Each plot represent the performance for Rule 1 and Rule 3 (combined), as a function of morphs for both relevant and irrelevant features. Grey corners for congruent stimuli, red corners for incongruent stimuli.

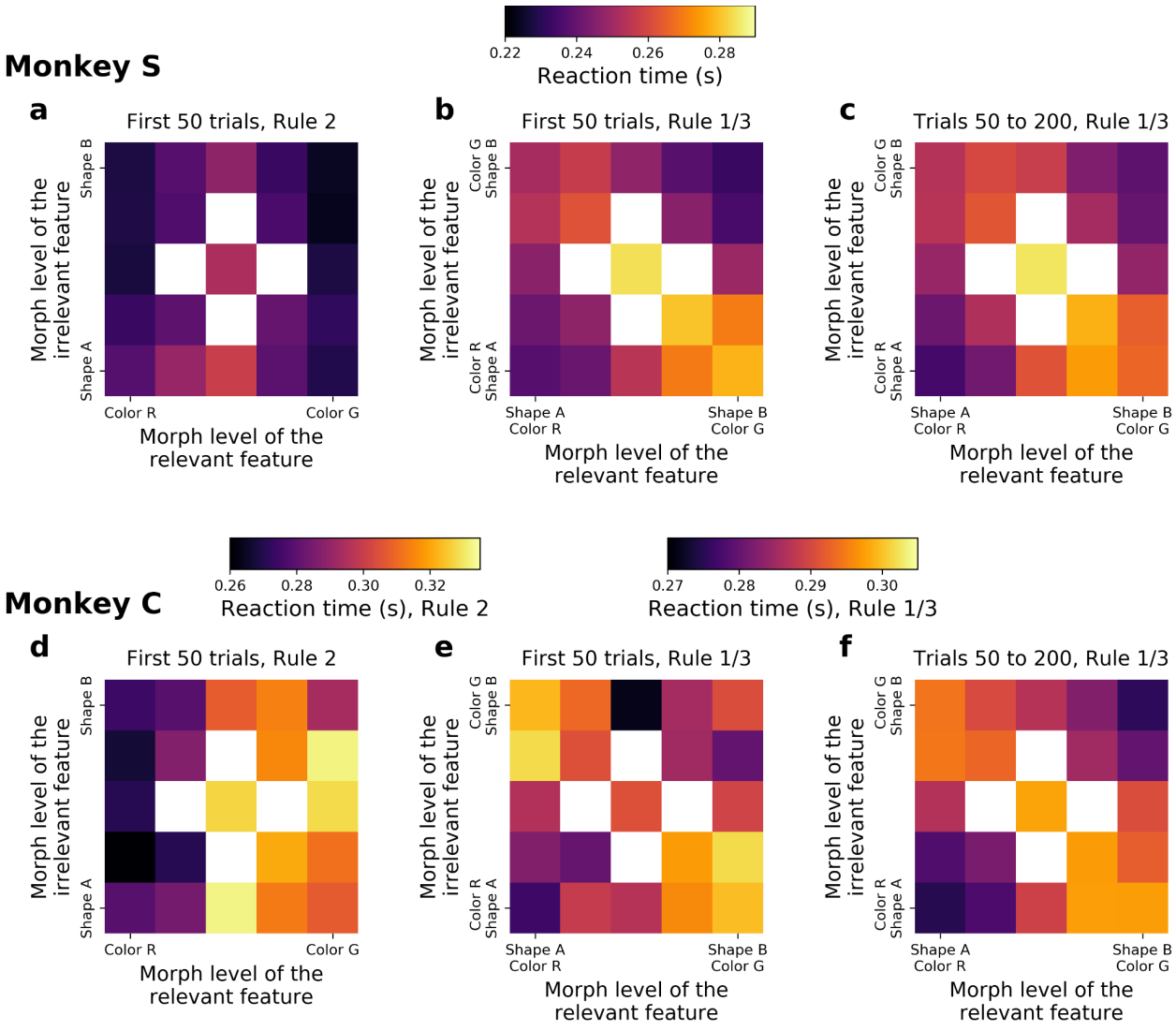


Figure S9: Reaction times for Rule 2 blocks, the first 50 trials of Rule 1/3 blocks, and the trials 50 to 200 of Rule 1/3 blocks as a function of the relevant and irrelevant features of the morphed stimulus presented. Top row: Monkey S, bottom row: Monkey C. Statistics on Monkey C: Δ(ms)=14 between incongruent and congruent, t-test p<10(-4)).

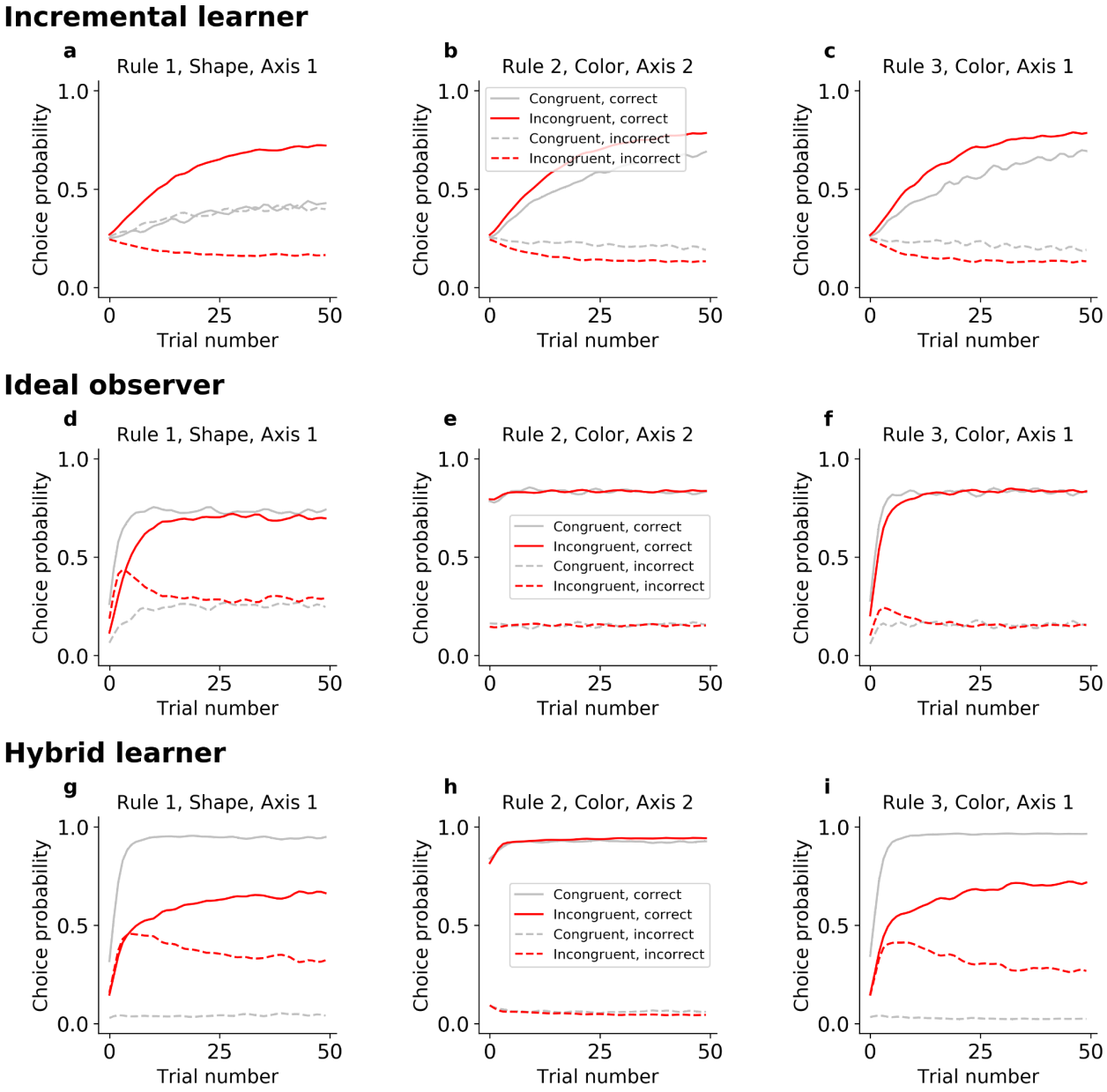


Figure S10: Choice probabilities for the models fitted on Monkey S.

Monkey S	Noise perception		Learning rate	Initial belief R1	Initial belief R3	Initial belief Axis 1	Weight decay	Initial weights
	κ(color)	κ(shape)	α	b1	b3	bax	η	**w0**
QL model	mean=2.2 std=0.44	mean=1.3 std=0.27	mean=0.23 std=0.039					
IO model	mean=2.4 std=0.46	mean=1.3 std=0.31		mean=0.091 std=0.089	mean=0.14 std=0.10			
HQL model	mean=11 std=1.2	mean=5.1 std=2.2	mean=0.23 std=0.10			mean=0.29 std=0.076	mean=0.046 std=0.022	mean=[-0.61,0.79,0.77,-0.64,-0.83,0.035,0.74,0.040] std=[0.23,0.089,0.13,0.17,0.053,0.59,0.19,0.57]
Monkey C	Noise perception		Learning rate	Initial belief R1	Initial belief R3	Initial belief Axis 1	Weight decay	Initial weights
	κ(color)	κ(shape)	α	b1	b3	bax	η	**w0**
QL model	mean=1.2 std=0.13	mean=0.71 std=0.090	mean=0.18 std=0.010					
IO model	mean=1.3 std=0.21	mean=0.71 std=0.18		mean=0.35 std=0.16	mean=0.32 std=0.16			
HQL model	mean=12 std=2.4	mean=7.0 std=3.4	mean=0.12 std=0.10			Fixed to 0.5	mean=0.067 std=0.060	mean=[-0.45,0.56,0.60,-0.56,-0.83,-0.12,0.70,-0.19] std=[0.16,0.12,0.093,0.16,0.051,0.45,0.25,0.41]

Figure S11: Models parameters.

Methods

1 Experimental design

Two rhesus macaques were faced with a compositional category-response task. In each trial, the monkeys made a saccade $a \in Actions$ with $Actions = \{Upper-Left, Upper-Right, Lower-Left, Lower-Right\}$ (Fig. 1a) in response to a two-dimensional stimulus combining a color C and a shape S (Fig. 1b). They received a deterministic reward $r \in \{0, 1\}$. The correct response depended on the rule in effect in blocks of 50-300 trials for Monkey S, and 40-435 trials for Monkey C. Each stimulus dimension was divided into two categories defined by two prototypes (red or green, bunny or tee). Creating a morph continuum between the prototypes allowed us to manipulate stimulus difficulty for each dimension independently, and we varied the morph levels across trials. Switches between blocks of trials were cued but the correct rule in each block was hidden. Three rules were used $\mathcal{R} = \{R_1, R_2, R_3\}$ (Fig. 1c,d). Each rule required to attend to only one feature of the stimulus, and to respond only on one axis ($\mathcal{A} = \{Axis1, Axis2\}$). Rule 1 ($R_1$) required a response on a diagonal, *Axis* 1, to the shape of the stimulus. Rule 2 (R_2) required a response on the other diagonal, *Axis 2*, to the color of the stimulus. Rule 3 (R_3) required a response on *Axis 1* to the color of the stimulus. Rule 2 and Rule 3 used the same feature of the stimulus, different from Rule 1. Importantly, Rule 1 and Rule 3 shared the same axis of response (*Axis 1*), different from Rule 2 (*Axis 2*). Rule 1 and Rule 3 blocks were randomly selected and interleaved by a Rule 2 block, such that the axis of response always changed following a block switch. Congruent trials used stimuli predicting the same correct response across Rule 1 and Rule 3 (e.g. a green bunny). Incongruent trials used stimuli predicting the opposite correct responses across Rule 1 and Rule 3 (e.g. a red bunny). There were four times more incongruent trials than congruent trials in the experiment.

A performance criterion of 70% on the last 100 trials (on "morphed" and "prototype" separately) was chosen to trigger a block switch from a Rule 1 or Rule 3 block to a Rule 2 block. Rule 1 and Rule 3 blocks were on average 199 trials long for Monkey S and 222 trials long for Monkey C. As the monkeys were performing very well in Rule 2 blocks, these were shorter (on average, 56 trials long for Monkey S and 52 trials long for Monkey C). The behavioral data for Monkey S corresponds to 20 days, with an average of 14 blocks per day. The behavioral data for Monkey C corresponds to 15 days for Monkey C, with an average of 6.5 blocks per day.

2 Modeling noisy perception of color and shape, independently

All the models studied below model stimulus perception in the same way (Fig. S1). The color and shape of each stimulus presented to the animals are either the prototype features $s_{Tc} \in \{red, green\}$ and $s_{Ts} \in \{bunny, tee\}$, or a morphed version of them. The presented stimulus

is noted $\mathbf{s_M} = (s_{Mc}, s_{Ms})$. We hypothesize that the monkeys perceive a noisy version of it, noted $\mathbf{s_K} = (s_{Kc}, s_{Ks})$. We model it by drawing two samples from two Von Mises distributions, centered around each feature (color and shape), parameterized by the concentrations κ_c and κ_s respectively. The models estimate each initial feature presented by computing its posterior distribution, given the perceived stimulus, i.e. by Von Mises distributions centered on s_{Kc} and s_{Ks}, with same concentrations κ_c and κ_s.

$\forall i \in \{c, s\}$

$$Pr(s_{K_i} \mid s_{M_i}) = \mathcal{V}(s_{M_i}, \kappa_i) \tag{1}$$

and so :

$$Pr(s_{M_i} \mid s_{K_i}) = VonMises(s_{K_i}, \kappa_i) \tag{2}$$

with, I_o being the modified Bessel function of order zero:

$$VonMises(\mu, \kappa) = \frac{\exp(\kappa \cos(x - \mu))}{2\pi I_o(\kappa)} \tag{3}$$

3 Modeling action-selection

All the models studied below use the same action-selection stage. Given the perceived stimulus at each trial $\mathbf{s_K} = (s_{Kc}, s_{Ks})$, an action is chosen so as to maximize the expected reward $\mathbb{E}(r \mid \mathbf{s_K})$ by computing $\max_a Pr(r = 1 \mid \mathbf{s_K}, a)$ which corresponds to maximizing the probability of getting a reward, given the perceived stimulus. We use the notation $Q(\mathbf{s_K}, a) = \mathbb{E}(r(a) \mid \mathbf{s_K})$ as in [1; 2] and refer to these values as *Q values*. Two fixed parameters implement an epsilon-greedy softmax action-selection rule: the lapse rate ϵ (for random exploration) and the inverse temperature β (for directed exploration depending on the actions' relative expected values [2]).

The action-selection rule is:

$\forall a \in Actions$

$$Pr(a \mid \mathbf{s_K}) = \frac{\epsilon}{4} + (1 - \epsilon) \cdot softmax[Q(\mathbf{s_K}, a)] \tag{4}$$

The lapse rate ϵ is directly estimated from the data, by computing the proportion of trials where the incorrect axis of response is chosen, asymptotically. It is evaluated to 0.02 for both Monkey S and Monkey C. The inverse temperature of the softmax β is also fixed ($\beta = 10$), and allows the algorithm to be differentiable (cf. use of Stan below).

4 Fit with Stan

All our models shared common noisy perceptual input and action selection stages (Fig. S1, and Methods). The models however differed in the intervening mechanism for dynamically mapping stimulus to action value (see also Fig. S1). Because of noise perception at each trial (Eq. 1), and because the cumulative distribution function of a Von Mises is not analytic, the models are fitted with Monte Carlo Markov chains (MCMC) using Stan [3]. Each day of recording is fitted separately, and the mean and standard deviation reported in Fig. S11 are between days. Fitting scripts are available on github at [link]. We validated convergence (all R-hat< 1.05) and efficiency diagnostics (all effective sample size > 100) of the models' fits .

Models' plots correspond to an average of 1000 simulations of each day of the dataset (with the same order of stimuli presentation). Statistics reported in the article were done with Fisher's exact test (except a t-test for reaction times, Fig. S9).

5 Incremental learner: QL model

This model corresponds to Fig. 2 (Monkey S) and S2 (Monkey C). It also appears in Fig. 5, S7, S8, S10.

In this model the agent is relearning each rule after a block as a mapping between stimuli and actions, by computing a stimulus-action value function as a linear combination of binary feature-response functions $\phi(\mathbf{s_K}, a)$ with feature-response weights $\mathbf{w}$. This implements incremental learning while allowing for some generalization across actions. The weights are updated through gradient descent (see [4], chapter 9). The weights are reset from one block to the other, and the initial values for each reset are set to the null. Fitting them does not change the results (see paragraph 5.4 below).

5.1 Computation of the feature-response matrix

Given a morph perception at trial t, $\mathbf{s_K} = (s_{Kc}, s_{Ks})$, a feature-response matrix is defined as:

$$\phi(\mathbf{s_K}) = \begin{pmatrix} x_C & 0 & 0 & 0 \\ x_S & 0 & 0 & 0 \\ 0 & x_C & 0 & 0 \\ 0 & x_S & 0 & 0 \\ 0 & 0 & x_C & 0 \\ 0 & 0 & x_S & 0 \\ 0 & 0 & 0 & x_C \\ 0 & 0 & 0 & x_S \end{pmatrix} \tag{5}$$

where $x_C \in \{-1, 1\}$ depends on whether the perceived morph for color s_{Kc} is classified as green or red (Eqs. 1 and 2), and $x_S \in \{-1, 1\}$ whether the perceived morph for shape s_{Ks} is classified as tee or bunny (Eqs. 1 and 2). In order for the algorithm to remain differentiable, we approximate $\{-1, 1\}$ with a sum of sigmoids (see scripts at [link]). Each column of the matrix $\phi(\mathbf{s_K})$ is written $\phi(\mathbf{s_K}, a)$ below and corresponds to an action $a \in Actions$.

5.2 Linear computation of Q values and action selection

In order to compute Q values, the feature-response functions $\phi(\mathbf{s_K}, a)$ are weighted by the feature-response weight vector $\mathbf{w} = (w_1, .., w_8)$ (see [4], equation (9.8)):

$\forall a \in Actions$

$$Q(\mathbf{s_K}, a) = \mathbf{w} \cdot \phi(\mathbf{s_K}, a) \tag{6}$$

Action selection is done through the epsilon-greedy softmax rule (Eq. 4).

Thus asymptotic learning of Rule 1 would require $\mathbf{w} = [0, 1, 0, -1, 0, 0, 0, 0]$. Learning Rule 2 would require $\mathbf{w} = [0, 0, 0, 0, -1, 0, 1, 0]$. Learning Rule 3 would require $\mathbf{w} = [-1, 0, 1, 0, 0, 0, 0, 0]$.

5.3 Weight vector update

Once an action a_t is chosen and a reward r_t is received at trial *t*, the weights are updated through gradient descent with learning rate α (see [4], equation (9.7)).

$$\mathbf{w} \leftarrow \mathbf{w} + \alpha(r_t - Q(\mathbf{s_K}, a_t))\phi(\mathbf{s_K}, a_t) \tag{7}$$

5.4 Parameter values

β and ϵ are fixed to respectively 10 and 0.02. See Fig. S11 for parameter values. As predicted from the behavior, noise perception is higher for shape than for color ($\kappa_C > \kappa_S$). The initial weight vector $\mathbf{w_0}$ is set to the null at the beginning of each block day. Fitting these weights instead gives the same results (as then $\mathbf{w_0}$ has a mean=[-0.070,0.031,0.093,-0.054,-0.081,-0.022,0.062,-0.0048] for Monkey S and $\mathbf{w_0}$ has a mean=[-0.099,0.070,0.069,-0.089,-0.053,-0.011,0.030,-0.0098] for Monkey C).

6 Optimal Bayesian inference over rules: IO model

This model corresponds to Fig. 3, and S3 ; also appears in Fig. 5, and S4, S7, S8, S10.

In this model, we assume a perfect knowledge of combination mappings between prototype stimuli and actions as *rules*. Learning is discovering which rule is in effect by Bayesian inference. This is done through learning, over the trials, the probability for each rule to be in effect in a

block (or *belief*) from the history of stimuli, actions and rewards. At each trial, this belief is linearly combined to the likelihood of a positive reward given the stimulus to compute a value for each action. This likelihood encapsulates knowledge of the three experimental rules. An action is chosen as per described above in Eq. 4. The beliefs over rules are then updated through Bayes rule using the likelihood of the reward received, given the chosen action and the stimulus perception. Once the rule is discovered, potential errors thus only depends on possible the miscategorization of the stimulus features (Eqs. 1 and 2), or eventually on exploration (Eq. 4).

6.1 Belief over rules

The posterior probability of rule $R \in \mathcal{R}$ to be in effect in the block is called the belief over the rule $b(R) = Pr(R \mid \mathbf{s_K}, a, r)$, given the perceived stimulus $\mathbf{s_K}$, the action a and the reward r. The beliefs $\mathbf{b}(R)$ at the beginning of each block are initialized to $\mathbf{b_0} = [b_1, 1 - b_1 - b_3, b_3]$ where b_1 and b_3 are fitted, to test for a systematic initial bias towards one rule.

6.2 Computation of values

The beliefs are used to compute the Q values for the trial:

$\forall a \in Actions$

$$\begin{aligned} Pr(r = 1 \mid \mathbf{s_K}, a) &= \sum_{R \in \mathcal{R}} Pr(r = 1, R \mid \mathbf{s_K}, a) \\ &= \sum_{R \in \mathcal{R}} Pr(r = 1 \mid \mathbf{s_K}, a, R) \cdot b(R) \end{aligned} \tag{8}$$

with (marginalization over the possible morph stimuli presented):

$\forall R \in \mathcal{R}$

$$\begin{aligned} Pr(r \mid \mathbf{s_K}, a, R) &= \sum_{\mathbf{s_M}} Pr(r, \mathbf{s_M} \mid \mathbf{s_K}, a, R) \\ &= \sum_{\mathbf{s_M}} Pr(r \mid \mathbf{s_M}, a, R) Pr(\mathbf{s_M} \mid \mathbf{s_K}) \end{aligned} \tag{9}$$

Noting $p_C = p(s_{Mc} = red \mid s_{Kc})$; and $p_S = p(s_{Ms} = bunny \mid s_{Ks})$, gives:

$$Q(\mathbf{s_K}, a = Upper - Left) = p_S \cdot b(R_1) + (1 - p_C) \cdot b(R_3) \tag{10}$$

$$Q(\mathbf{s_K}, a = Upper - Right) = (1 - p_S) \cdot b(R_1) + p_C \cdot b(R_3) \tag{11}$$

$$Q(\mathbf{s_K}, a = Lower - Left) = (1 - p_C) \cdot b(R_2) \tag{12}$$

$$Q(\mathbf{s_K}, a = Lower - Right) = p_C \cdot b(R_2) \tag{13}$$

6.3 Belief update

From making an action $a_t \in Actions$, the agent receives a reward $r_t \in \{0, 1\}$, and the beliefs over rules are updated:

$\forall R \in \mathcal{R}$

$$b(R) \leftarrow Pr(r_t \mid \mathbf{s_K}, a_t, R) \cdot b(R) \tag{14}$$

with $Pr(r_t \mid \mathbf{s_K}, a_t, R)$ the likelihood of observing reward r_t for the chosen action a_t.

Note that because of the symmetry of the task, $Pr(\neg r_t \mid \mathbf{s_K}, a_t, R) = 1 - Pr(r_t \mid \mathbf{s_K}, a_t, R)$.

6.4 Parameter values

β and ϵ are respectively fixed to 10 and 0.02. See Fig. S11 for parameter values. As predicted from the behavior, there is an initial bias for Rule 2 for the model fitted on Monkey S behavior ($b2 > b_3 > b_1$). Also, noise perception is higher for shape than for color for both monkeys ($\kappa_C > \kappa_S$). In the version of the model with low perceptual color noise (Fig. 3 and S3), all the parameters remain the same, except that we fix $\kappa_C = 6$ for all simulated days.

7 Hybrid incremental learner: HQL model

The hybrid incremental learner combines inference over axes with incremental learning, using a Q-learning with function approximation to relearn the likelihood of rewards given stimuli per axis of response. This model corresponds to Fig. 4 and S5 ; also appears in Fig. 5, S6, S7, S8, S10.

7.1 Belief over axes

The posterior probability of an axis $A \in \mathcal{A}$ to be the correct axis of response in a block is called the belief over axis $b(A) = Pr(A \mid \mathbf{s_K}, a, r)$, given the perceived stimulus $\mathbf{s_K}$, the action a and the reward r. The beliefs over axes are initialized at the beginning of each block to $\mathbf{b_0} = (b_{ax}, 1 - b_{ax})$.

7.2 Computation of the feature-response matrix

As for the incremental learner above, given a morph perception at trial t, $\mathbf{s_K} = (s_{Kc}, s_{Ks})$, a feature-response matrix is defined as:

$$\phi(\mathbf{s_K}) = \begin{pmatrix} x_C & 0 & 0 & 0 \\ x_S & 0 & 0 & 0 \\ 0 & x_C & 0 & 0 \\ 0 & x_S & 0 & 0 \\ 0 & 0 & x_C & 0 \\ 0 & 0 & x_S & 0 \\ 0 & 0 & 0 & x_C \\ 0 & 0 & 0 & x_S \end{pmatrix} \tag{15}$$

where $x_C \in \{-1, 1\}$ depends on whether the perceived morph for color s_{Kc} is classified as green or red (Eqs. 1 and 2), and $x_S \in \{-1, 1\}$ whether the perceived morph for shape s_{Ks} is classified as tee or bunny (Eqs. 1 and 2). Each column of the matrix $\phi(\mathbf{s_K})$ is written $\phi(\mathbf{s_K}, a)$ below and corresponds to an action $a \in Actions$.

7.3 Computation of values

The beliefs are used to compute the Q values for the trial:

$\forall a \in Actions$

$$\begin{aligned} Q(\mathbf{s_K}, a) = Pr(r = 1 \mid \mathbf{s_K}, a) &= \sum_{A \in \mathcal{A}} Pr(r = 1, A \mid \mathbf{s_K}, a) \\ &= \sum_{A \in \mathcal{A}} Pr(r = 1 \mid \mathbf{s_K}, a, A) \cdot b(A) \end{aligned} \tag{16}$$

Contrary to the ideal observer, here the likelihood of reward per action $Pr(r = 1 \mid \mathbf{s_K}, a, A)$ is learned through function approximation.

$$Pr(r = 1 \mid \mathbf{s_K}, a, A) = sigmoid(\mathbf{w} \cdot \phi(\mathbf{s_K}, a)) \tag{17}$$

Action selection is done through the epsilon-greedy softmax rule (Eq. 4).

7.4 Weight vector update

Once an action a_t is chosen and a reward r_t is received at trial *t*, the weights are updated through gradient descent with learning rate α (see [4], equation (9.7)).

$$p_t = Pr(r = 1 \mid \mathbf{s_K}, a_t, A_t)) \tag{18}$$

$$\mathbf{w} \leftarrow \mathbf{w} + \alpha(r_t - p_t)\phi(\mathbf{s_K}, a_t)p_t(1 - p_t) \tag{19}$$

As learning improved steadily in this model contrary to the asymptotic behavior of monkeys, we implemented a weight decay to asymptotic values $\mathbf{w_0}$:

$$\mathbf{w} \leftarrow (1 - \eta) \cdot \mathbf{w} + \eta \cdot \mathbf{w_0} \tag{20}$$

Note that resetting the weights at the beginning of each block and adding a weight decay (or a learning rate decay) provide similar fits to the dataset. Also, this decay can be included in the previous two models without any change of our results and conclusions.

7.5 Belief update

From making an action $a_t \in Actions$, the agent receives a reward $r_t \in \{0, 1\}$, and the beliefs over axes are updated:

$\forall A \in \mathcal{A}$

$$b(A) \leftarrow Pr(r_t \mid \mathbf{s_K}, a_t, A) \cdot b(A) \tag{21}$$

with $Pr(r_t \mid \mathbf{s_K}, a_t, A)$ the likelihood of observing reward r_t for the chosen action a_t.

7.6 Parameter values

β and ϵ are fixed to respectively 10 and 0.02. As predicted from the behavior, noise perception is higher for shape than for color for both monkeys ($\kappa_C > \kappa_S$). Also, the model fitted on Monkey S behavior has an initial bias for $Axis2$ ($b_{ax} < 0.5$). For fitting the model on Monkey C behavior, we fix $b_{ax} = 0.5$. Finally, the fitted values of $\mathbf{w_0}$ correspond to an encoding of an average between Rule 1 and Rule 3 on $Axis1$, and an encoding of Rule 2 on $Axis2$, for both monkeys.

References

[1] Dayan, P. & Daw, N. D. Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience* **8**, 429–453 (2008).

[2] Daw, N. D., O'doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. Cortical substrates for exploratory decisions in humans. *Nature* **441**, 876–879 (2006).

[3] Carpenter, B. *et al.* Stan: A probabilistic programming language. *Journal of statistical software* **76**, 1–32 (2017).

[4] Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction* (MIT press, 2018).