Pre-training RNNs on ecologically relevant tasks explains sub-optimal behavioral reset

AUTHOR NAMES AND AFFILIATIONS

Manuel Molano-Mazon[1], Daniel Duque[1], Guangyu Robert Yang[2] and Jaime de la Rocha[1]

[1] IDIBAPS, Rosselló 149, Barcelona, 08036, Spain

[2] Center for Theoretical Neuroscience, Columbia University

CORRESPONDING AUTHOR

Manuel Molano-Mazón

Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS)

Rosselló 149-153, Barcelona, 08036, Spain

molano@clinic.cat

Abstract

When faced with a new task, animals' cognitive capabilities are determined both by individual experience and by structural priors evolved to leverage the statistics of natural environments. Rats can quickly learn to capitalize on the trial sequence correlations of two-alternative forced choice (2AFC) tasks after correct trials, but consistently deviate from optimal behavior after error trials, when they waive the accumulated evidence. To understand this outcome-dependent gating, we first show that Recurrent Neural Networks (RNNs) trained in the same 2AFC task outperform animals as they can readily learn to use previous trials' information both after correct and error trials. We hypothesize that, while RNNs can optimize their behavior in the 2AFC task without a priori restrictions, rats' strategy is constrained by a structural prior adapted to a natural environment in which rewarded and non-rewarded actions provide largely asymmetric information. When pre-training RNNs in a more ecological task with more than two possible choices, networks develop a strategy by which they gate off the across-trial evidence after errors, mimicking rats' behavior. Our results suggest that the observed suboptimal behavior reflects the influence of a structural prior that, adaptive in a natural multi-choice environment, constrains performance in a 2AFC laboratory task.

Introduction

In order to make good decisions in real life scenarios, animals are equipped with extensive implicit knowledge about the world. For instance, in foraging tasks, animals balance exploitation of the best alternatives with exploration of the less promising ones (Addicott et al. 2017), a behavioral pattern that seems imprinted in the brain (Daw et al. 2006; Blanchard and Gershman 2018; Chakroun et al. 2020) and that is present even when the difference in the alternatives values is large and stable (Vulkan 2000). These innate behaviors, that we called structural priors, shape the entire landscape of behavioral solutions available to the animal. However it is still unknown how to identify them and to which extent they influence the behavioral strategy adopted by animals in a given task.

Structural priors are thought to impact the behavior of animals in tasks organized sequentially into trials. For instance, animals' responses are influenced by the history of sensory stimuli (Akaishi et al. 2014; Fischer and Whitney 2014; Akrami et al. 2018) and of previous responses and outcomes (Corrado et al. 2005; Lau and Glimcher 2005; Busse et al. 2011; Donahue, Seo, and Lee 2013; Abrahamyan et al. 2016; Urai et al. 2019). These sequential effects are suboptimal in most laboratory tasks with trials presented independently. However, they are highly robust and prevalent, which may indicate the existence of a hardwired circuitry that prevents animals from learning and implementing a more optimal strategy.

Many sequential effects are directly related to the outcome of the decisions made by the animal. Feedback has been shown to impact the bias towards the different options (Donahue, Seo, and Lee 2013; Abrahamyan et al. 2016; Urai et al. 2019), the speed of the subsequent responses (Rabbitt and Rodgers 1977) and the strategy used by subjects (Fusi et al. 2007; McDougle et al. 2016; Purcell and Kiani 2016b; Sarafyazd and Jazayeri 2019). We have recently shown that rats are able to use the recent history of transitions after correct trials but consistently ignore it after error trials, following a sub-optimal reset strategy (Hermoso-Mendizabal et al. 2020). These studies demonstrate that animals do not process success and failure merely as mirroring outcomes and that the neural mechanisms they trigger are qualitatively different (Lyon and Kuchling 2021). However, the structural prior underlying this asymmetry is still a matter of much debate (Baumeister et al. 2001; Alves, Koch, and Unkelbach 2017).

Recurrent Neural Networks (RNNs) constitute a useful tool to study the neural circuit mechanisms implementing the computations required to solve sequential tasks (Sussillo 2014; Barak 2017; Yang and Wang 2021). When trained, RNNs are typically allowed to adjust their connections to adopt the best strategy for the particular task, without any constraint imposed by preexisting structural priors. However, this approach can potentially produce networks that use fundamentally different solutions from the ones

used by animals, which are influenced by a plethora of structural priors. To overcome this mismatch between animals and RNNs, it has been recently suggested that networks should be pre-trained in more naturalistic tasks that induce the necessary priors in the networks before they face the laboratory task (Ma and Peters 2020; Yang and Molano-Mazon, n.d.). However, to our knowledge, no study has shown a successful example of the need of pre-training RNNs in order to replicate a sup-optimal behavior observed experimentally. Besides, RNNs are usually trained with supervised learning (SL) algorithms (Werbos 1990), which is at odds with the way animals primarily learn novel tasks (Niv 2009). Nevertheless, it is still unknown to what extent SL algorithms produce networks that use fundamentally distinct strategies to those used by subjects performing the same task.

Here, we conduct rat behavioral experiments to show that the sub-optimal reset strategy adopted in a 2AFC task with serial correlations (Hermoso-Mendizabal et al. 2020) is a prevalent behavior across individuals and task variants. RNNs trained directly on the 2AFC task failed to replicate the rats behavior, outperforming them by fully exploiting the binary structure of the task. On the other hand, RNNs pre-trained in more naturalistic environments containing more than two alternatives replicated the reset strategy used by rats. Finally, as rats do, pre-trained networks maintained the trial history information after making a mistake, and used it again as soon as they made a correct choice. Our results suggest that the suboptimal strategy exhibited by rats in the 2AFC task is the result of a hardwired structural prior that both guides and constrains their learning towards solutions consistent with more natural environments. Furthermore, our work demonstrates that comparing animals' behavior with that exhibited by RNNs may benefit from pre-training the networks in more ecologically relevant environments before testing them on the task of interest.

Results

Rats develop a robust suboptimal behavior in a 2AFC task with serial correlations

To investigate the extent to which rats utilize the information present in the statistical structure of the trial-to-trial sequence, we trained them in an auditory two-alternative forced choice (2AFC) task that included serial correlations (Hermoso-Mendizabal et al. 2020). In particular, the task was structured into trial blocks in which the probability that the previous stimulus category was repeated in the current trial, P_{rep}, varied between high and low values (P_{rep}=0.8 in repeating blocks and P_{rep}=0.2 in alternating blocks; block size 80-200 trials; Fig. 1a). Rats learned to categorize the current stimulus, but also to infer P_{rep} using the trial history of previous repetitions and alternations, what we called the transition history: within the repeating block, rats showed a tendency towards repeating their previous choice, whereas in the alternating block they tended to alternate (Fig. 1b). However, rats only displayed this transition bias after correct trials. After error trials, they consistently ignored such information and followed a strategy only based on the current stimulus (Hermoso-Mendizabal et al. 2020) (Fig. 1c). Importantly, this strategy, which we called reset strategy, is suboptimal in an environment presenting only two alternatives and in which contexts are very stable (i.e. repeating/alternating blocks are long). A more optimal agent would, after an error trial, perform the simple counterfactual inference: “had I chosen the opposite side, I would have received a reward”; and use such inference to anchor the transition bias from the non-chosen side. For example, if the current estimate of the transition probability was to alternate, after an incorrect Right choice the agent should infer that the reward was present in the Left port and generate an alternation from there (i.e. Left→ Right), resulting in a rightward bias in the next choice (Fig. 1d). Because such a strategy implies a momentary reversal of the transition bias after an error trial, we call it the reverse strategy (Fig. 1e, gray area). Note that because the probability of an incongruent transition is relatively high (e.g. repetitions in the alternating block occur with probability P_{rep}=0.2), a single error should not be interpreted as a context change which occurs with a much lower probability (P<1/80).

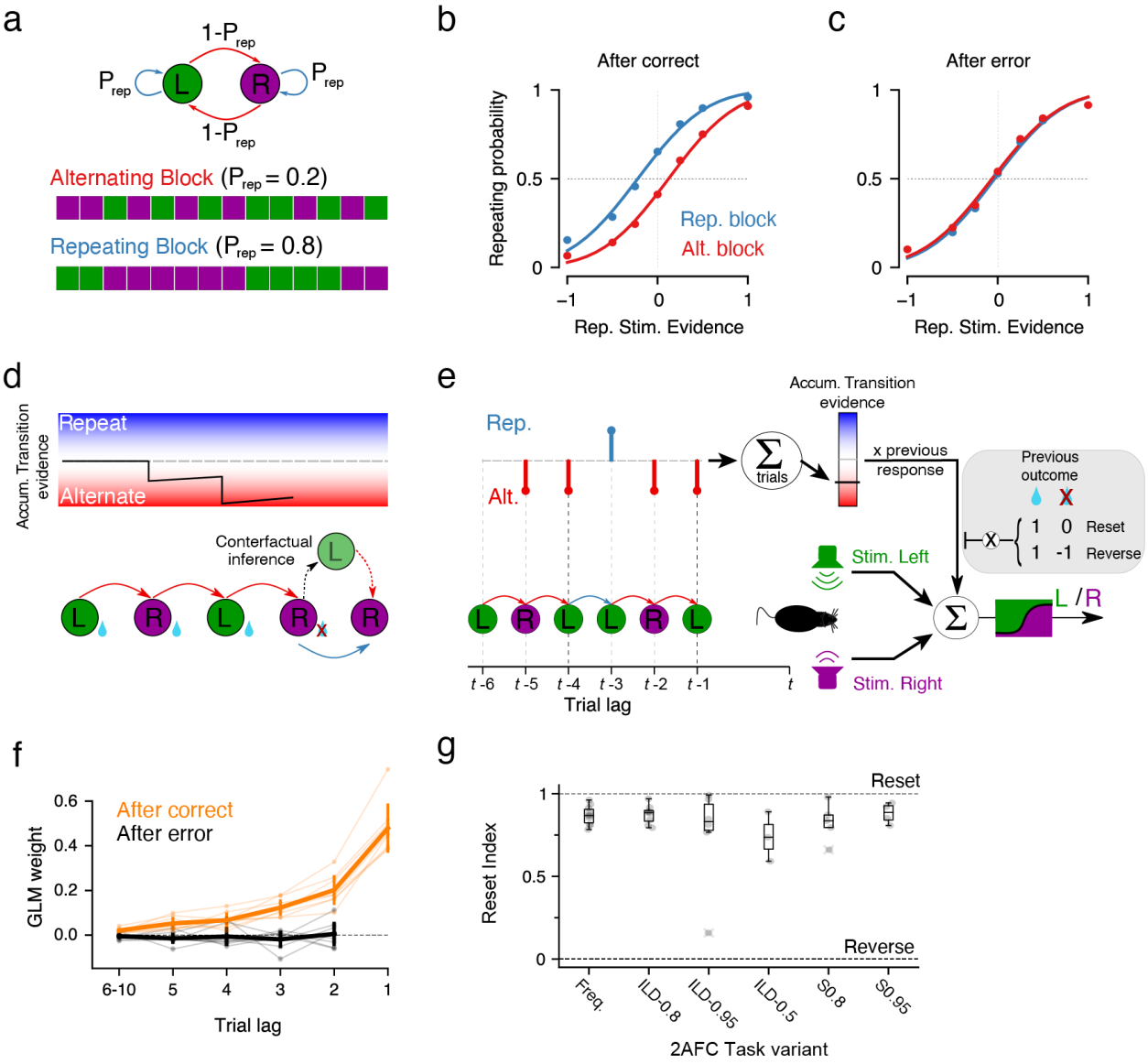


Figure 1. **Rats develop a history bias only after correct trials**. **a)** The stimulus trial sequence in the categorization 2AFC task is generated by a two-state Markov chain parametrized by the repeating probability P_{rep} (top) which is varied in Repeating (P_{rep}=0.8) and Alternating trial blocks (P_{rep}=0.2) (bottom). Blocks length was set between 80 and 200 trials. **b-c)** Average Psychometric curves showing the probability of repeating the previous choice as a function of the stimulus evidence supporting the repetition (rat Group ILD-0.8, n=8 rats). Curves were computed separately for trials after correct. **(b)** and after-error **(c)** choices, and for the repeating (blue) and alternating (red) blocks. Dots represent the data and lines are logistic fits. **d)** Example sequence of choices made by an optimal agent (bottom), together with the trace of the accumulated transition evidence (top). After an error the agent infers the rewarded side (black arrow) and uses the accumulated alternating evidence to bias its next response towards repeating the error choice (effectively reversing the alternating bias). **e)** Rats' responses are modeled using a GLM that combines the current stimulus evidence (bottom) and the history of transitions, i.e. previous repetitions (blue) and alternations (red). The weighted sum of transition history provides the accumulated transition evidence (color bar) which captures the current tendency to repeat or alternate the previous rewarded side. Shaded area shows the effect that the previous outcome has on the contribution of the transition evidence, both in

the reset strategy exhibited by rats and in the reverse strategy expected from an ideal observer. **f)** GLM weights of previous correct transitions (i.e. formed by two correct choices) computed after correct (orange) and after error (black) trials; mean and std. dev. from rat Group ILD-0.8 (n=8, thick traces) and individual animals (light traces) **g)** Reset Index (see methods) for different rat groups performing different task variants. Group Freq.: frequency discrimination task, n= 10 (Hermoso-Mendizabal et al. 2020); Group ILD-0.8: intensity level discrimination task with P_{REP} = 0.8, n=8; Group ILD-0.95: ILD task with P_{REP} = 0.95, n=6; Group ILD-Uncorr: ILD task with uncorrelated sequences, (i.e. P_{REP} = 0.5), n=6 (rats from Group ILD-0.8); Group S0.8: silent task without any stimuli and P_{REP} = 0.8, n=5; Group S0.95: silent task without any stimuli and P_{REP} = 0.95, n=4. The reset strategy implies RI~1 whereas in the reverse strategy RI~0.

To quantify the extent to which the behavior of individual rats followed the reset or the reversal strategies, we trained rats in different variants of the task and modeled their choices using a Generalized Linear Model (GLM) (Lau and Glimcher 2005; Abrahamyan et al. 2016; Braun, Urai, and Donner 2018). The GLM assumes that the rats' choices are the result of a linear combination of various task variables such as the current stimulus evidence or the previous choices and transitions together with their outcomes (Fig. 1e) (see Methods). This analysis allows us to quantify the contribution of the past correct transitions (those formed by two consecutive correct choices) to the rat's current choice. To uncover the potentially different strategies followed by rats depending on the outcome in the previous trial, we separately fitted their choices after correct and error trials with two different models. Thus, the model fitting after-error choices should yield vanishing weights associated to the previous transitions for the reset strategy, while showing negative ones for the reverse strategy (Fig. 1e, inset). We found that rats consistently utilized the transition history after a correct trial and ignored it after an error (Fig. 1f). We used the fitted transition weights from after correct and after error trials to define the Reset Index (RI), which was zero for perfectly symmetric kernels (i.e. reverse) and approached one as the after-error kernel vanished (i.e. reset; see Methods). The RI was close to one (mean±SD 0.83 ±0.06) for almost every animal tested in different variants of the 2AFC task, in which we changed the stimulus feature to be categorized (frequency or intensity) and the repeating probability (P_{rep} = 0.8-0.2, 0.95-0.05 and 0.5-0.5, i.e. no serial correlations) (Fig. 1g). Furthermore, we trained rats in a variant of the task with no stimuli (Fig. 1g, Group S0.8, n=5), and with extremely predictable repeating and alternating sequences (P_{rep} = 0.95-0.05, Group S0.95, n=4) with the aim of making the task conceptually simpler and thus the pattern of transitions more evident. Despite this major simplification, rats still displayed a robust reset strategy (Fig. 1g). Moreover, rats also discarded any information provided by transitions containing incorrect choices (i.e. correct-error, error-error or error-correct transitions) (Supp. Fig. 1) (Hermoso-Mendizabal et al. 2020). Importantly, as with the reset strategy, this disregard of any information involving error trials is not optimal, since in the 2AFC task all types of transitions are equally informative (Supp. Fig. 2). This robust behavior further supports the hypothesis that there exists a fundamental difference between the strategies followed by rats when dealing with correct and error outcomes.

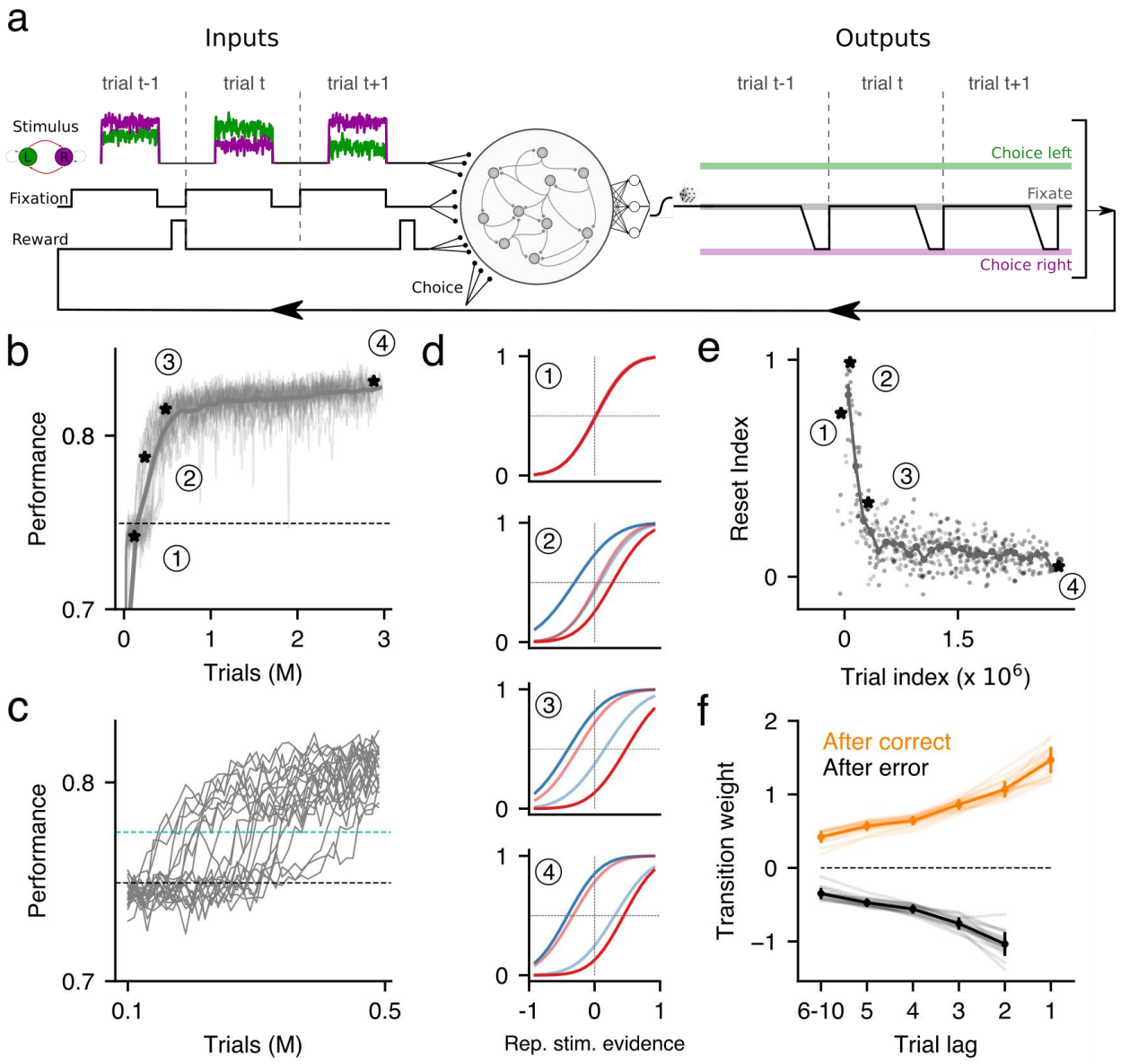


Figure 2. Behavior of RNNs trained directly on the 2AFC task. a) Networks training: 1024-unit LSTM networks were trained using Reinforcement Learning (Z. Wang et al. 2016). At each timestep, the networks received as input a fixation cue, two inputs corresponding to the two stimuli (left/right), and the reward and action at the previous timestep; in turn, at each timestep networks had to fixate, go left or go right (see methods). The example shows a hypothetical sequence in which the network (middle) experienced 3 alternating trials (left) and chooses the right side at the end of each of them (right). **b)** Performance across training of 16 RNNs (thin traces). The thick, dark line corresponds to the median performance. Dashed line corresponds to the performance of an agent that only uses the current stimulus and not the transition history to make a decision. Black asterisks and numbers correspond to different periods of training for which we plotted the psychometric curves in panel d. **c)** Expanded view of the period during which RNNs learn to use the transition history. Aha moments were detected by setting a threshold on the performance (dashed cyan line). **d)** Psychometric curves as the ones shown in Fig. 1b-c, for the different training periods indicated in

panel a. **e)** Reset Index values for the same networks shown in panel a, when tested in the 2AFC task at different stages of training (light dots). Values are aligned to each network's aha moment (see c). Dark dots correspond to the median reset index across individual networks. Error-bars correspond to standard error. **f)** Transition kernels as the ones shown in Fig. 1e corresponding to trained networks (point 4 in panels a, d and e). Error-bars correspond to standard deviation.

RNNs learn to fully leverage trial-history information

What causes the reset strategy observed in rats? One possibility is that the counterfactual inference needed after error trials constitutes a complex computation that is difficult to implement in a neural circuit. To test this hypothesis, we trained Recurrent Neural Networks (RNNs) in the 2AFC task. RNNs were presented with noisy stimuli following a sequence with serial correlations which were structured on Repeating and Alternating blocks (Fig. 2a). In order to test the ability of RNNs to perform counterfactual inference, we trained them using Reinforcement Learning (RL) (Z. Wang et al. 2016; Sutton and Barto 2018), by which networks only receive feedback in the form of a scalar (the reward), without being explicitly told the correct answer at each timestep, as it is done in Supervised Learning techniques (Werbos 1990). Thus, at each timestep, networks received the following inputs: two stimuli providing evidence for each of the choices and a fixation cue. Further, we provided the networks with the reward and action from the previous timestep, which allows them to be more adaptive to the different contexts of a task (J. X. Wang et al. 2018). In turn, at each timestep the output of the networks could be to fixate, to choose left or to choose right (Fig. 2a; see Methods).

RNNs quickly learned to integrate the stimuli to inform their decision, reaching a performance comparable to that of an agent that integrates the stimulus perfectly but acts independently of previous history (Fig. 2b, dashed line). After exploiting the perfect-integrator strategy for a variable period of the training, networks underwent an aha moment and started exploiting the information provided by the transition history, hence further improving their performance (Fig. 2b, c). After the aha moment, the shift in strategy occurred relatively fast and in a few thousand trials the performance of the networks reached a relative plateau in which they already used both the stimulus and the transition history (Fig. 2b-d). The behavior at the beginning of this plateau (Fig. 2b, point 3) was already close to the reverse strategy, with the transition bias being present both after correct and after error trials (Fig. 2d, points 3 and 4). Along this long, final phase of training, the reverse strategy was refined with the transition bias after correct and error trials becoming more symmetric, causing the Reset Index to reach lower values (Fig. 2e). Thus, RNNs trained directly on the 2AFC task learned to leverage both the stimulus and the transition history information, as rats do. However, the RNNs strategy after error trials

fully exploited the structure of the task, thus departing from the reset strategy displayed by rats (Fig. 2f).

We next investigated whether the networks adopted the reset strategy at any point during the training. To do that we tested them with their weights frozen at different time-points during the training. The RI values peaked at the aha moment (Fig. 2e and point 2 in panels a and d) and then dropped, reaching a reverse strategy at the end of the training (Fig. 2e and point 4 in panels a and d). However, close inspection of the factors influencing the RNNs responses revealed that the RI peak reflected a highly transient behavior caused by a short lag in the development of after-correct and after-error transition bias, ruling out the idea that the reset could constitute a stable solution (Supp. Fig. 3).

That a simple RNN can exploit the symmetry of the 2AFC task and adopt a reversal strategy (Fig. 1d) suggests that rats' reset strategy was not caused by a limit of their computational capacity but by some more fundamental factors we aim to explore next.

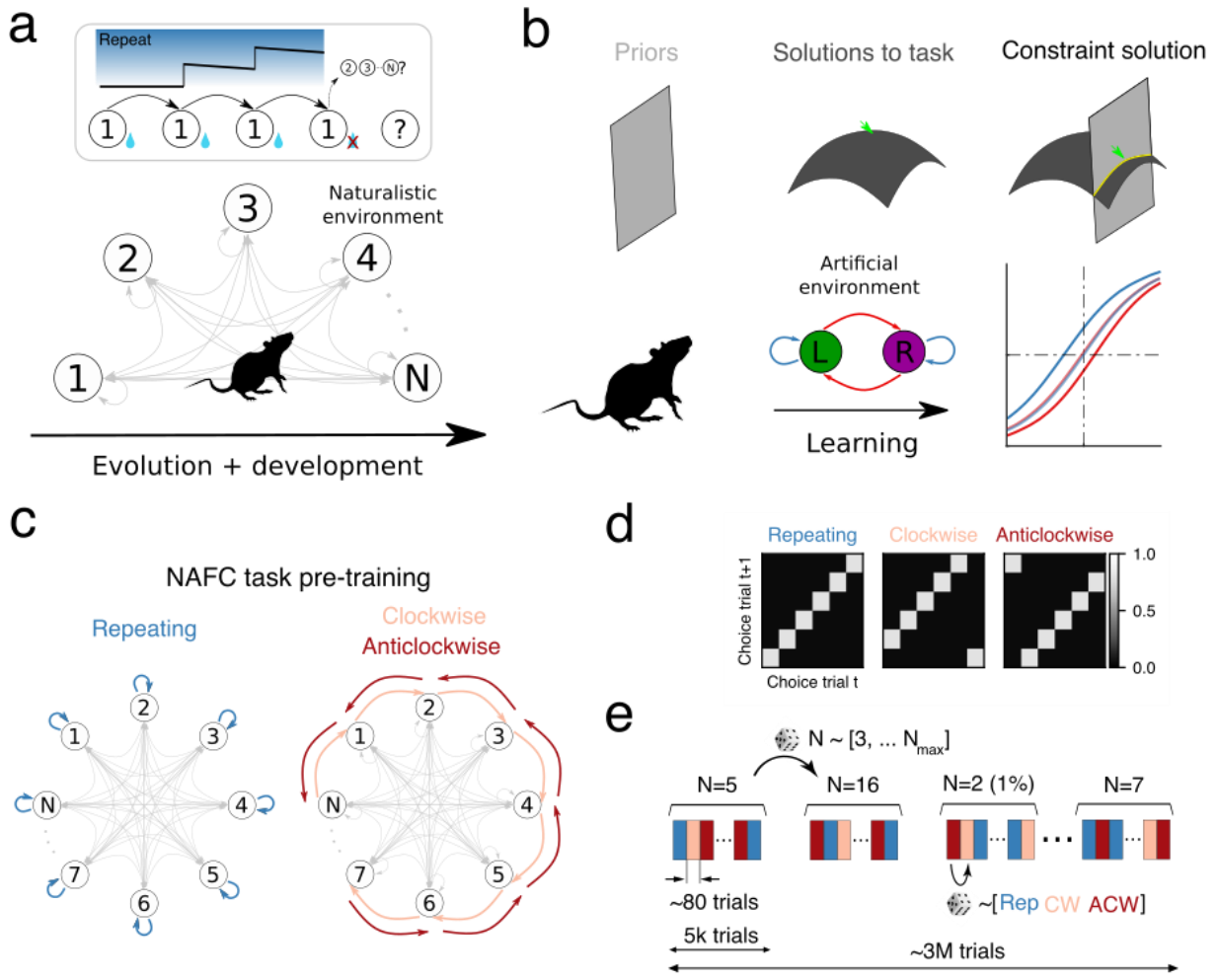


Figure 3. The reset strategy naturally emerges in an environment presenting more than two alternatives. a) Rats have evolved in environments with more than two alternatives (bottom) in which a

single error prevents subjects from predicting the next state of a sequence despite knowing the transition probabilities (top). **b)** Cartoon describing how the priors imposed by evolution and development in naturalistic environments (left column) affect the animal's capacity to find the optimal solution in the 2AFC task (middle columns, green arrow); the final solution (right column) constitutes a compromise between the priors and the task's optimal solution. **c-d)** In the N-AFC task the across-trial stimulus sequence is organized into contexts (i.e. trial blocks) with distinct serial correlations, each parametrized by a transition matrix (d). We pre-trained the networks using repeating, clockwise and anticlockwise contexts, each one having a most likely sequence (colored arrows) and less frequent transitions (gray arrows). **e)** During pretraining in the NAFC task, context length was random (avg. block of 80 trials) and the number of choices N was changed every 5k trials. The percentage of blocks with N = 2, i.e. the 2AFC task, was set to 1%.

The reset strategy is adaptive in environments with more than two alternatives

If the reverse strategy can be implemented in a small neural circuit, why do rats reset after making a mistake? We reasoned that a fundamental difference between the training of our rats and our RNNs is that, while the former face the learning of the task with an extensive background knowledge about the statistical structure of the world, RNNs start as a blank slate whose wiring can be fully optimized to perform a 2AFC task. Perhaps the reset strategy originates from evolution as an adaptation to more naturalistic environments, giving rise to a structural prior that hinders the rats from fully exploiting the structure of the 2AFC task. We hypothesize that one key difference between the 2AFC task and naturalistic environments is the number of available alternatives. As explained above (Fig. 1d), in any 2AFC task subjects have always implicit access to the identity of the correct alternative, independently of the outcome of their decision. Thus, correct and incorrect answers provide the same exact amount of information about the environment. This is no longer true when we increase the number of alternatives to more than two: in a task with N>2 alternatives, information about the reward location is ambiguous after making an incorrect choice, because it could be any of the other N-1 alternatives. This does not occur after a rewarded choice. Furthermore, this asymmetry grows with the number of alternatives, with the certainty after error trials quickly decreasing as 1/N, while remaining unaltered after correct trials. In such a scenario, any strategy exploiting serial correlations becomes infeasible: knowing that you should repeat is useless if you do not know what to repeat (Fig. 3a). We thus propose that the reset strategy is the result of a structural prior which, having adapted to an environment with multiple alternatives, waives the accumulated transitions evidence after error trials (Fig. 3b).

Pre-training RNNs in a more ecological task recovers the reset strategy

To test this hypothesis we embedded the 2AFC task in an N-AFC environment presenting a variable number of alternatives that was larger than two (Fig. 3c). With this simplified setup, we seeked to emulate the existing tension between the structural priors that guide and constrain the learning of the rat and the rat's necessity to adapt to a new, artificial laboratory task. RNNs trained in these conditions, called pre-trained RNNs, are pushed to find the best solution for the 2AFC task within the realm of solutions set by the NAFC environment. The NAFC task modeled the spatiotemporal correlations commonly found in natural environments, by generating a sequence of rewarded states using non-homogeneous transition probability matrices (Fig. 3c-d). In particular, trials were blocked into a repeating, a clockwise and an anticlockwise context, each defined by a different transition matrix and presented randomly (Fig. 3). The maximum number of alternatives was fixed per network (N_{max}), and the number of available alternatives, N, was randomly selected between 3 and N_{max} every 5k trials, setting N=2 in a small fraction of the blocks (1%) (Fig. 3e). To characterize the development of the structural prior, we computed the performance in the NAFC of pre-trained networks on trials with zero stimulus evidence. From early stages of the training, the average performance was above chance for all values of N (Fig. 4a), demonstrating that networks quickly developed and used a transition bias regardless of the number of alternatives. Moreover, although initially networks did better in the repeating context, they eventually achieved a very similar performance in all contexts (Fig. 4b).

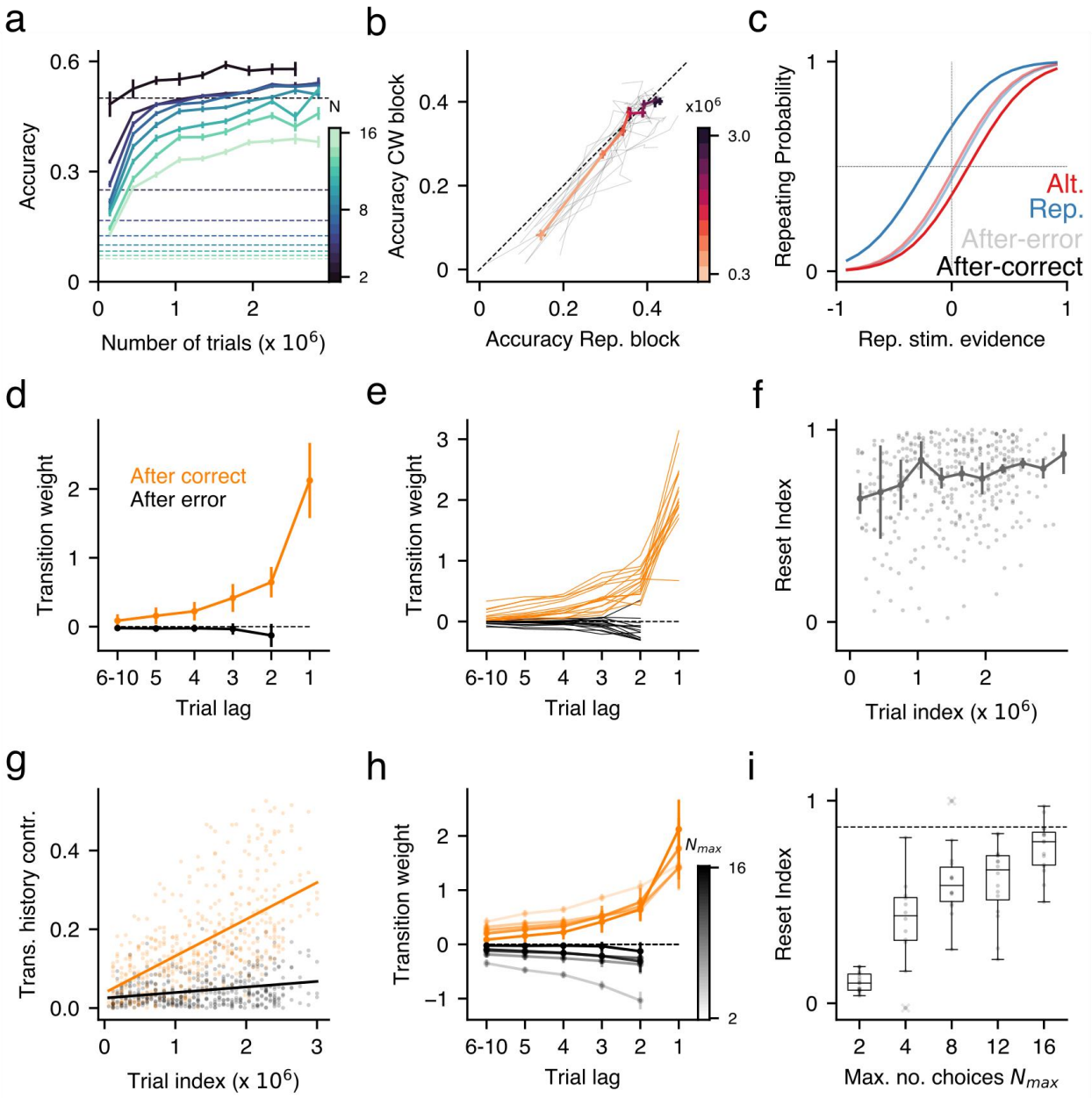


Figure 4. Behavior of networks (n=16) pretrained in a 16-AFC task. a) Median accuracy for trials with zero stimulus evidence conditioned on different number of choices N=2, 4, 8, 16 (see colorbar). Dashed lines indicate the chance level (1/N) corresponding to each value of N. **b)** Median accuracy in the repeating versus the clockwise blocks as the RNNs learn along pre-training (color represents number of trials; see colorbar). Gray lines show individual RNNs. **c)** Psychometric curves for each block (see color code) computed in after-correct and after-error trials at the end of the training for one example RNN. **d)** Average after-correct and after-error transition kernels for pre-trained RNNs tested in the original 2AFC task. **e)** Same as d for individual networks. **f-g)** Median Reset Index (f) and contribution of transition history (g) in the 2AFC task across pre-training. Points show individual RNNs. Transition history contribution is computed after correct (orange) and after error (black) trials as the absolute value of the sum of the individual weights at different lags (see Methods). Lines correspond to linear fits. **h)** Average transition kernels obtained from networks pre-trained with a different maximum number of choices N_{max} (see colorbar). **i)** Reset Index versus N_{max}. Dots correspond to individual networks. Dashed line indicates the mean Reset Index obtained across

all rats and task variants. Error-bars correspond to standard error in panels a, b and f and standard deviation in panels d and h.

We next investigated the extent to which the structural prior induced by the NAFC environment conditioned the RNNs strategy in the 2AFC task. We characterized the behavior of the networks by freezing their connectivity and testing them in the 2AFC task, thus preserving the wiring sculptured by the pre-training. The pre-trained networks followed the reset strategy observed in the rats. This can be seen in the difference between the psychometric curves obtained after correct and after error trials (Fig. 4c) and was confirmed by the vanishing transition weights obtained from the GLM fit after error trials (Fig. 4d, e). The reset strategy was steadily developed as the pre-training proceeded (Fig. 4f) due to an increase of the transition weights in after-correct trials with only a marginal change in after-error trials (Fig. 4g). As expected, the asymmetry between after-correct and after-error transition weights increased as the maximum number of alternatives N_{max} increased (Fig. 4h), which made the Reset Index rise accordingly reaching values similar to those found in rats for N_{max}=16 (Fig. 4i). Furthermore, pre-trained networks not only ignored transition history after error trials, but for large N_{max} they mostly disregarded transitions containing at least one error, just as rats do (Supp. Fig. 4).

The reset behavior stemmed from two features of the pre-training: the information asymmetry between after-correct and after-error trials in the NAFC environment and the appropriate balance between the learning of the 2AFC and the contextual influence of the NAFC. To show the key role of these factors, we first pre-trained RNNs in a variant of the NAFC task in which networks received the identity of the correct alternative, independently of the outcome of their response. Having access to this information, networks learned to make use of the transition history both after correct and error trials, showing a clear reverse strategy (Supp. Fig. 5a). Second, we manipulated the trade-off between the reset and the reversal strategy by varying the proportion of 2AFC trials that the RNNs experienced during the pre-training (with N_{max}=16). As expected, the RI decreased continuously and rapidly as this proportion increased from 0.1 to 100% (Supp. Fig. 5b). Moreover, although the NAFC environment could prevent the adoption of the optimal reverse strategy in the 2AFC task, it also facilitated a faster learning to exploit, at least partially, the across-trial serial correlations as can be revealed by comparing pre-trained RNNs with networks trained directly in the 2AFC (Supp. Fig. 5c): pre-trained RNNs needed many fewer 2AFC trials (e.g. 25k trials or 1% of the total) to reach the accuracy that naive networks reached only after a much larger number of trials (~250k). By contrast, removing the serial correlations from the NAFC (for N>2), promoted a memoryless trial-history prior that burdened pre-trained RNNs from learning transition biases (Supp. Fig. 5c). Thus the pre-training in the NAFC task both constrained and helped the quick learning of the 2AFC task, guiding it towards a more consensual solution.

In summary, the above results demonstrate that by pre-training the networks in a more naturalistic task in which the number of alternatives is larger than two, we are able to recover the reset behavior displayed by rats in the 2AFC task.

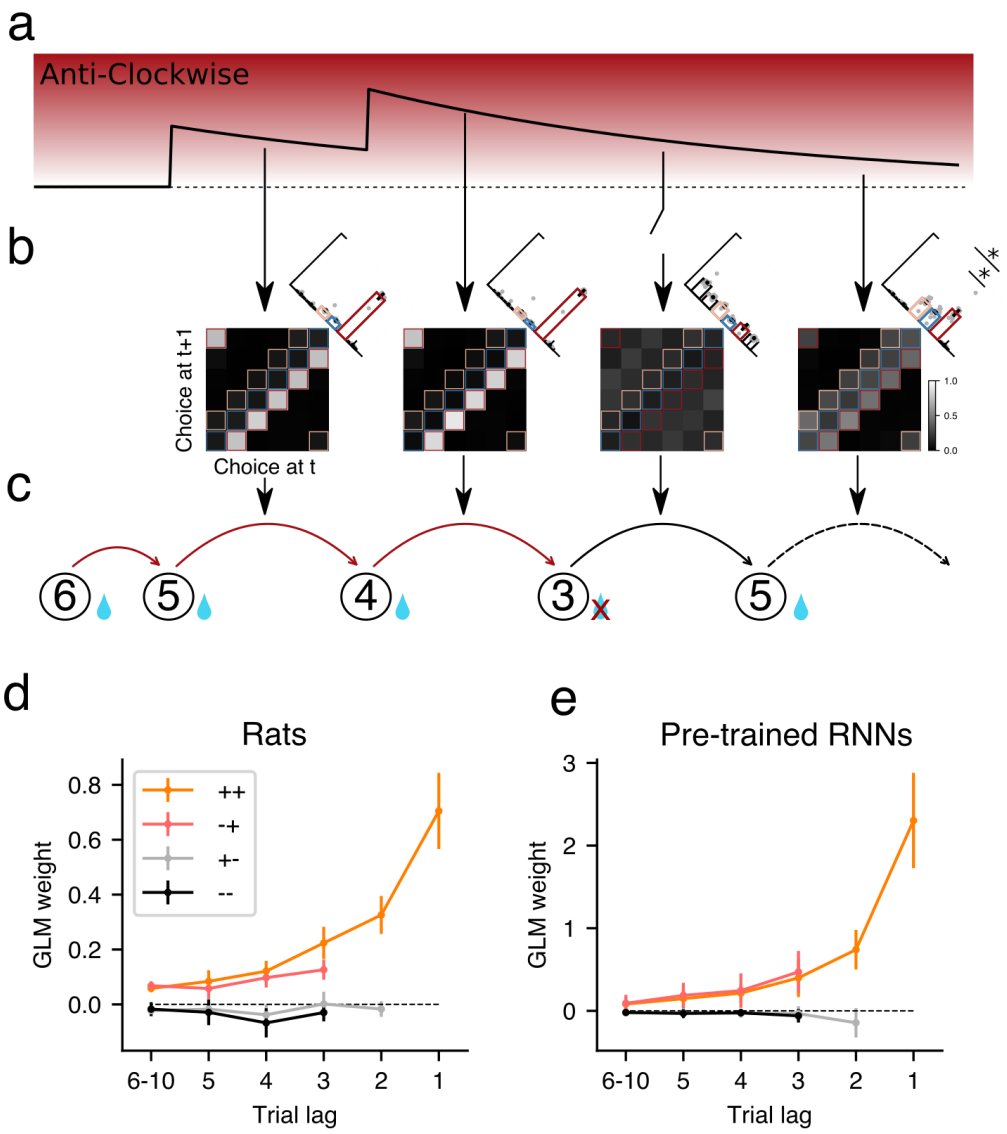


Figure 5. Pre-trained networks show gate-and-recovery dynamics after an error followed by a correct response. a-c) Schematic illustrating the across-trial dynamics of a latent variable encoding the probability of an anticlockwise (ACW) transition throughout a sequence containing three correct responses following an ACW pattern, followed by an error and then a correct response (panel c shows an example of such sequence). Notice that the occurrence of an error does not reset this latent probability but it gates off its impact onto the next choice (see broken arrow). The evolution of the transition bias matrix throughout the sequence (b) was numerically computed using pre-trained RNNs (Nmax= 16, RNNs are tested in N=6). Inset: histograms showing the probability of making a Repeating (blue), Clockwise (pink) and Anticlockwise (red) choice. In the last trial, the histogram shows a higher probability ($p<0.002$, paired t-test) to make an ACW than a Rep or CW choice. **d-e)** Average T^{++} transition kernel obtained in the 2AFC task when the

GLM is fitted independently depending on the outcome of the last two trials for rats **(d;** n= 10**)**, and pre-trained RNNs **(e**; n= 16**) (**see inset; e.g. −+ represents that *t-2* was incorrect and *t-1* was correct).

Pre-trained RNNs maintain the transition history information after an error

Having shown that pre-trained networks display the reset strategy after error choices, we next aimed to understand the dynamics of the transition bias beyond this reset. Do pre-trained networks recover part of the accumulated transition evidence after an error trial or does the inference of the transition probabilities start up *de novo*? Because the Markov transition matrix generating state sequences in the NAFC environment is relatively stable (i.e. blocks are ~80 trials long), the adaptive behavior in the NAFC is to transiently gate off the use of the accumulated transition evidence after an error but recover its use after the next correct choice. For example, if the network is in the NAFC repeating context and makes an error, then it should go unbiased until making a correct choice, after which it should resume repeating because, most likely, it remains in the repeating context. To test whether pre-trained RNNs exhibited such a gate-and-recovery behavior, we investigated the across-trial dynamics of the latent transition bias matrix $T_{ij}(t)$ (Fig. 5b). This matrix quantifies the average probability that, having selected choice *i* in trial *t*-1*,* the networks select choice *j* in trial *t* in response to a stimulus with zero evidence. We computed $T_{ij}(t)$ throughout a sequence composed of three correct responses following a pattern congruent with one of the contexts, e.g. two anticlockwise transitions (ACW), followed by an error and then a correct response (Fig. 5c). After the first correct ACW transition, the networks showed a clear tendency to reproduce the ACW transition in the next trial, a tendency that got more pronounced after the second correct ACW transition (Fig. 5b, left panels). An error response afterwards completely flattened the transition matrix, reflecting the reset strategy. However, a single correct response after the error, recovered the networks' ACW tendency, evidencing that the transition history accumulated previous to the error had not been erased (Fig. 5b, inset in rightmost panel). The networks were therefore able to update and maintain an internal representation of the accumulated transition evidence, transiently gating it off after errors and recovering it after the next correct choice (Fig. 5a-b). We finally checked that the gating of the transition evidence was specific to after-error trials as it did not occur after correct unexpected responses that broke the context's transition pattern (Supp. Fig. 6).

The adaptive gate-and-recovery behavior developed by pre-trained networks in the NAFC percolated into the 2AFC task mimicking the behavior found in rats (Hermoso-Mendizabal et al., 2020). To show this, we computed the transition kernel separately from trials conditioned on the outcome of the last two trials, giving rise to four conditions: after correct-correct (++), after error-error (--), after correct-error (+-) and after error-correct (-+). After an error (+- or --), the kernel weights were close to zero in both rats and pre-

trained RNNs (Fig. 5d-e) consistent with previous analysis (Fig. 1f and 4d-e, respectively). However, after a correct response following an error (-+), the weight of previously accumulated transitions (trial lags -3, -4, …) was recovered, indicating that those past transitions had again the same impact on choice as if there have not been any error (i.e. their weights were the same as after two correct responses ++). Thus, networks pre-trained in the NAFC reproduced the complex gate-and-recovery strategy displayed by rats in the 2AFC task, supporting the idea that their behavior is constrained by a structural prior deeply founded in the asymmetry between correct and error responses characteristic of natural environments.

Discussion

We have investigated the causes for the suboptimal strategy displayed by rats in a 2AFC task containing serial correlations: rats ignore the transition evidence accumulated across the trial history after error trials. This so-called reset strategy was highly robust and pervasive across many animals performing different task variants. RNNs trained directly on the task did not replicate the reset strategy because they were able to adopt the more optimal reverse strategy that leverages the binary symmetry of the 2AFC task. On the other hand, RNNs pre-trained in an environment containing more than two alternatives exhibited the reset strategy displayed by the rats. Furthermore, although the pre-trained RNNs did not show a transition bias after error trials, they maintained, as rats do, the transition history information to use it as soon as they made a correct choice (Fig. 5).

Pre-training RNNs in ecologically relevant environments

RNNs have become a widely used tool to investigate the neural mechanisms underlying the behavior of animals in laboratory tasks (Mante et al. 2013; Sussillo 2014; Sussillo et al. 2015; Carnevale et al. 2015; Barak 2017; J. X. Wang et al. 2018; Remington et al. 2018; Mastrogiuseppe and Ostojic 2018; Yang and Wang 2021; Feulner and Clopath 2021; Saxena et al. 2021). The usual approach is to train the RNN directly on the task of interest and then investigate the underlying circuit mechanisms. By contrast, animals arrive at the laboratory with a plethora of priors that both guide and constrain the solutions learned by the animal to solve different tasks. By ignoring these priors RNNs are able to explore a much larger realm of strategies, frequently outperforming the animals but departing from the solutions they use. A quantitative discrepancy in performance has been frequently solved by adding noise to the units in the network (e.g. Mante et al., (Mante et al. 2013)) or to the stimulus (e.g. Sohn et al., (Sohn et al., n.d.)). However, although noise may be a mechanism limiting the performance of neural circuits (Faisal, Selen, and Wolpert 2008), it cannot account for all factors that constrain the space of solutions available to the animal. A complementary approach is to train the RNNs in several cognitive tasks, hence forcing them to develop strategies that are general enough to assure a good performance in all tasks (J. X. Wang et al. 2018; Yang et al. 2019; Hadsell et al. 2020). Here we followed a more specific approach (Ma and Peters 2020; Yang and Molano-Mazon, n.d.): pre-training the RNNs in an environment that presents more than two alternatives to address a qualitative discrepancy between the behavior of networks and rats in a 2AFC task with serial correlations.

This pre-training approach has been rarely used in neuroscience and, to the best of our knowledge, only for feedforward networks (Stoianov and Zorzi 2012; Roseboom et al. 2019). Nevertheless, within the machine learning (ML) community, pre-training is a

routine procedure aiming to speed up the learning of a given task by assuring a good performance in a more general one (Dahl et al. 2012; Devlin et al. 2018; Tan and Le 2019). The structural priors induced by such pre-training help in naturalistic tasks like the ones used in ML, but can be a handicap in the tasks used in the laboratory that present oversimplified statistics or artificial symmetries. For instance, the visual system of humans possess extraordinary capabilities that are highly adapted to the statistics of natural scenes (Geisler 2008) but fail to correctly interpret seemingly simple stimuli, causing a visual percept that consistently differs from reality (i.e. a visual illusion) (e.g. (Weiss, Simoncelli, and Adelson 2002)). Similar phenomena have also been studied at a cognitive level (Kahneman 2011). For example, humans show a tendency to avoid losses at the expense of disregarding potential gains (Kahneman and Tversky 2012) a behavior that may have been useful in the past, helping to make fast decisions in dangerous environments (Kahneman 2011). Our results illustrate both sides of structural priors, with the NAFC pre-training helping the networks to leverage the transition history after correct responses but impeding them to do so after an error (Supp. Fig. 5c).

Modelling the behavior of rats in the 2AFC task using RNNs

We chose RNNs to model the impact of structural priors on rats' behavior because modern training methods allowed us to explore a vast realm of solutions in a relatively hypothesis-free fashion (i.e. without the choice of model introducing any obvious priors). Moreover, in contrast with more phenomenological methods to model behavior, RNNs offer the possibility to study the underlying neural mechanisms and can be used as hypotheses generators that can guide the analysis and interpretation of neural recordings from the rats as they implement the reset strategy.

The pre-training protocol we used with the pre-trained RNNs cannot be but a simplification of the real environment in which the brain of rats evolved. First, we reduced the statistical structure of the NAFC environment to the basic features needed to induce the transition bias, presenting only three contexts with a simplified structure in their transition probability matrices (Fig. 3d). Introducing more contexts or more complex transition matrices would have required longer training and possibly larger networks but we do not foresee an *a priori* limitation in this regard. A possible alternative to the statistics we used in the NAFC task is to generate the state sequences independently (i.e. no serial correlations), with each context being defined by a different vector of heterogeneous alternative probabilities. This pre-training would promote a structural prior adapted to estimate the first order rates of each alternative and a history bias which would tend to repeat rewarded options and stay away from unrewarded options. But it would not generate any gatting of the accumulated evidence after errors. Interestingly, this type of history bias is commonly

observed in the form of win-stay lose-switch bias in 2AFC tasks using uncorrelated sequences (Lau and Glimcher 2005; Busse et al. 2011; Donahue, Seo, and Lee 2013; Abrahamyan et al. 2016; Urai et al. 2019) and in our animals in the 2AFC task with serial correlations (Hermoso-Mendizabal et al. 2020), rising the question of whether RNNs pre-trained in such statistics could indeed reproduced this ubiquitous behavior. Moreover, in our simplified approach, the training of the RNN in the laboratory task was embedded in the pre-training in the NAFC environment (Fig. 3e). A more realistic approach to approximate the multi-timescale learning of biological neural circuits, would be to build the connections of the RNNs using two different processes working at different time scales (Yang and Molano-Mazon, n.d.). For instance, a slow learning process similar to the one used here would simulate the effect of evolution, whereas a faster training, limited to certain connections and based on reward-modulated Hebbian plasticity (Miconi 2017), would simulate the learning of the laboratory task.

Reinforcement versus supervised learning

RNNs are commonly trained with Supervised Learning (SL) protocols (Werbos 1990), which provide them with the correct answer at each timestep. Thus, networks not only know when they are wrong but also what they should have done. Here we have trained the networks using Reinforcement Learning techniques (RL) (Sutton and Barto 2018) which are primarily based on sparse feedback that reinforces the correct choices and discourages the mistakes, akin to the way animals learn in the laboratory (Niv 2009). This approach is a key feature of our training protocol, since the (lack of) access to the identity of the previous correct choice is at the core of the asymmetry between correct and error responses. Accordingly, when pre-trained networks had access to the previous correct answer they were able to adopt the more optimal reverse strategy (Supp. Fig. 5a). To the best of our knowledge, the extent to which SL and RL techniques promote different solutions for a given cognitive task has not been systematically explored. However, the differences at the behavioral (Lyon and Kuchling 2021) and neural (Donahue, Seo, and Lee 2013; Purcell and Kiani 2016a) level between after-error and after-correct responses should serve as a warning when comparing SL-trained networks with the activity of biological neural circuits.

Processing feedback in correct versus error responses

Previous studies have also reported reset behaviors after erroneous choices. In a sensorimotor mapping task, monkeys trained to associate two visual stimuli with two different actions reverted to chance performance after making a single mistake (Fusi et

al. 2007). This after-error reset of the stimulus-response association contrasts with the reverse behavior an optimal agent would develop in this task. Rats also showed a reset in their strategy, abandoning an exploitation strategy and adopting an exploratory behavior upon detection of a contingency change in a two-alternative task (Karlsson, Tervo, and Karpova 2012). In light of our findings, these two seemingly suboptimal behaviors may follow from the same principle: animals have adapted to environments with multiple alternatives where identifying the best exploitation strategy upon a contingency change may require some exploration.

The main hypothesis of the current work relies on a simple observation: in real life, rewarded actions reinforce the existing action policy (Sutton and Barto 2018), while errors are only informative about what not to do. The uncertainty after erroneous choices is magnified in scenarios that require decisions at different hierarchical levels beyond the perceptual categorization process: from the strategy to follow (Purcell and Kiani 2016b; Sarafyazd and Jazayeri 2019), to the exact motor trajectory (McDougle et al. 2016) or the timing to execute the selected action (Hernández-Navarro et al. 2020). This multi-level process makes the asymmetry between positive and negative outcomes larger. Such an asymmetry can be viewed as a direct consequence of the so-called Anna Karenina Principle, which states that there are many ways in which things can go wrong but only one in which they will go right (Diamond 2017). Hence, inferences about the environment after erroneous decisions are difficult because there are countless explanations about what went wrong. The extent to which all happy families are alike is an open question; our findings suggest, however, that our cognitive abilities can indeed be quite alike as a consequence of our shared evolutionary history in a highly structured world.

Methods

All experimental procedures were approved by the local ethics committee (Comité d'Experimentació Animal, Universitat de Barcelona, Spain, Ref 390/14).

Animal Subjects

Animals were male Long-Evans rats (Charles River), pair-housed during behavioral training and kept in stable conditions of temperature (23 °C) and humidity (60%) with a constant light-dark cycle (12h:12h, experiments conducted during light phase). Rats had *ad libitum* food, and *ad libitum* water on days with no experimental sessions, but water was restricted during behavioral sessions.

Groups. **Group Freq.**: frequency discrimination task with P_{rep} = 0.7, n= 10 (Hermoso-Mendizabal et al. 2020); Intensity level discrimination (ILD) tasks: **Group ILD-0.8** with P_{rep} = 0.8, n=8; **Group ILD-0.95** with P_{rep} = 0.95, n=6; **Group ILD-Uncorr** with uncorrelated sequences (i.e. P_{rep}= 0.5), n=6 (rats from Group ILD-0.8); Silent tasks without any stimuli: **Group S0.8** with P_{rep} = 0.8:, n=5; **Group S0.95** with P_{rep} = 0.95: n=4;

	# rats	# sessions (mean, range)	# trials (mean, range)
FD task 0.7	10	59	56,506
ILD task 0.8	8	64 [21-87]	39299 [11532-55792]
ILD task 0.95	6	16 [15-17]	7666 [6523,8786]
ILD task Unc	6	30 [28-34]	18903 [14840-26516]
Silent task 0.8	5	19 [18-21]	14380 [8940-19412]
Silent task 0.95	4	52 [44-60]	31190 [18512-38307]

Behavioral tasks

Frequency discrimination task: rats performed an auditory reaction-time two-alternative forced choice task (Hermoso-Mendizabal et al. 2020). Briefly, at each trial, an LED on the center port indicated that the rat could start the trial by poking in. After a fixation period of 300 ms, the LED switched off and an acoustic stimulus consisting in a superposition of two amplitude-modulated frequencies was presented (see details below). Each frequency was associated with a specific side and reward was available at one of the two lateral ports, depending on the dominant frequency. Animals could respond any time after the stimulus onset. Correct responses were rewarded with a 24 µL drop of water and incorrect

responses were punished with a bright light and a 5 s time-out. Trials in which the rat did not make a side poke response within 4 seconds after leaving the center port were considered invalid trials and were excluded from the analysis (average of 0.4% invalid trials per animal). Withdrawal from the center port before stimulus onset (fixation break) cancelled stimulus presentation. After a fixation break, rats were allowed to initiate fixation again.

Intensity level discrimination task (Pardo-Vazquez et al. 2019; Hermoso-Mendizabal et al. 2020). Two speakers positioned at both sides of the box played simultaneously an amplitude-modulated white noise (see below for details). Rats had to discriminate the side with the loudest sound (right or left) and seek reward in the associated port. The rest of the details of the task are the same as in the frequency discrimination task.

Silent task. At each trial, an LED on the center port acted as a go cue indicating the rat could start the trial by poking in. After a fixation period of 300 ms, the LED switched off and rats had to guess at which of the two lateral ports the reward would appear. The rest of the details of the task are the same as in the frequency discrimination task. All the experiments were conducted in custom-made operant conditioning cages, the behavioral set-up was controlled by an Arduino-powered device (BPod v0.5, Sanworks, LLC, Stony Brook, NY, USA) and the task was run using the open-source software PyBPod (pybpod.com).

Acoustic stimulus

In the two acoustic tasks used, the stimulus $S_k(t)$ was created by simultaneously playing two amplitude-modulated (AM) sounds $T_R(t)$ and $T_L(t)$:

$$S_k(t) = [1 + sin(f_{AM} \; t + \varphi)] \, [{a_k}^L(t) \, T_L(t) + {a_k}^R(t) \, T_R(t)]$$

Where the frequency was f_{AM} = 20 Hz. The phase delay φ = 3π/2 made the envelope zero at t = 0. In the frequency discrimination task, $T_L(t)$ and $T_R(t)$ were pure tones with frequencies 6.5 kHz and 31 kHz, respectively, played simultaneously in the two speakers. In the interaural level discrimination task (ILD), they were broadband noise bursts played on the left and on the right speaker, respectively. The amplitudes of the sounds $T_L(t)$ and $T_R(t)$ were calibrated at 65 dB SPL using a free-field microphone (Med Associates Inc, ANL-940-1). Sounds were delivered through generic electromagnetic dynamic speakers (ZT-026 YuXi) located on each side of the chamber.

Stimulus Sequence

A two-state Markov chain parametrized by the conditioned probabilities P_{REP} =P(L|L) and P(R|R) generated a sequence of stimulus category c_k = {-1,1}, which determined the side

of the reward (left/right). In each trial the stimulus strength s_k was randomly drawn from a fixed set of values = [0, 0.25, 0.5, 1]. s_k defined the relative weights of the rewarded and non-rewarded sounds. The stimulus evidence was defined in each trial as the combination $e_k = c_k * s_k$, thus generating seven different options (0, ±0.25, ±0.5, ±1). The value of e_k determined the distribution from which the instantaneous evidence $S_{k,f}$ was drawn every 50 ms (*f* refers to the frame index) (Hermoso-Mendizabal et al. 2020). Finally, the amplitudes $a_k^L(t)$ and $a_k^R(t)$ of the two envelopes were defined as $a_k{}^L(t) = (1 + S_{k,f}) / 2$ and $a_k{}^R(t) = (1 - S_{k,f}) / 2$. With this choice the sum of the two envelopes was constant in all frames: $a_k^L(t) + a_k^R(t) = 1$.

Psychometric curve analysis

The *repeating* psychometric curves were computed by computing the proportion of repeated responses as a function of the repeating stimulus evidence ($\hat{e}$) defined for the *t*-th trial as $\hat{e}_t = r_{t-1} e_t$, with $r_{t-1} = \{-1,1\}$, representing the response in the previous trial (left or right, respectively). Thus, positive (negative) values of $\hat{e}$ corresponded to trials in which the stimulus evidence pointed to repeating (alternating) the previous choice, respectively. Psychometric curves were fitted to a 2-parameter probit function:

$$P_{Repeat}(\hat{e}) = \frac{1}{2}(1 + erf(\frac{\beta \hat{e} + B}{\sqrt{2}}))$$

The sensitivity β quantified the stimulus discrimination ability, while the fixed side bias B captured the animal side preference for the left (B < 0) or right port (B > 0).

Generalized linear model (GLM)

For both rats and RNNs we fitted the same GLM model to quantify the weight that the various factors of the task had on the choices (Busse et al. 2011; Frund, Wichmann, and Macke 2014; Abrahamyan et al. 2016; Braun, Urai, and Donner 2018; Hermoso-Mendizabal et al. 2020). The probability $p(r_t=+1|y_t)$ that the response r_t in the t-th trial was Rightwards was modeled as a linear combination of the current stimulus and trial history passed through a logistic function:

$$p(r_t=1|y_t) = \frac{1}{(1+ e^{-y_t})}$$

Where the argument of the function was

$$y_t = \sum_f \omega_f^S S_{t,f} + \sum_o \sum_{k=1}^{6} \omega_{k,o}^L r_{t-k}^o + (\sum_{o,q} \sum_{k=1}^{6} \omega_{k,o,q}^T T_{t-k}^{o,q}) r_{t-1} + \beta$$

The current stimulus was given by $S_{t,f}$ defined as the intensity difference between the two tone sounds in frame *f* with *f* = 1,2…N_t and N_t being the number of frames listened in trial

t . The trials history contributions included the impact of the previous ten trials (t-1, t-2, t-3...; grouping the impact of trials t−6 to trial t−10 in one term). The terms r_{t-k}^{+} represented the previous rewarded responses being −1 (correct left), +1 (correct right), or 0 (error response). Similarly, r_{t-k}^{-} represented previous unrewarded responses being −1 (incorrect left), +1 (incorrect right), or 0 (correct response). Previous transitions were given by:

$$T_k^{o,q} = r_{k-1}^{o} r_k^{q}$$

Reset Index

The Reset Index (RI) quantifies the extent to which an agent follows the reset strategy. It is computed as

$$RI = 1 - \frac{Tr_{after-error}}{Tr_{after-correct}}$$

Where $Tr_{after-error}$ and $Tr_{after-correct}$ (Fig. 4g) are computed as the absolute value of the sum of transition weights $\omega_{k,+,+}^{T}$ with k = 2, …, 6-10 (excluding the most recent transition), obtained from separately fitting the GLM in after error and after correct trials, respectively. Note that the Reset Index is only meaningful for agents that present a transition bias in the first place. For this reason, we set a threshold, th_{contr}, for the sum of the transition history contributions ($Tr_{after-error}$+$Tr_{after-correct}$) and discarded any agent with a contribution below that threshold. This only affected experiments with uncorrelated trial sequences (Group ILD-Uncorr, th_{contr} = 0.05) (Fig. 1g) and RNNs early in the (pre)training (Figs. 2e and 4f-g, th_{contr}=0.1). We obtained similar results for different values of th_{contr}.

Recurrent Neural Networks

All networks are fully connected, recurrent neural networks composed of 1024 long short-term memory (LSTM) units (Hochreiter and Schmidhuber 1997). In all experiments except when specified, the input included a fixation binary cue indicating when to respond and the stimuli, which corresponds to 1 or 2 samples (mean=1.3) drawn from a multivariate normal distribution with dimension equal to the number of choices. Further, networks received as input two scalars indicating the reward received and the index of the choice made on the preceding time-step, respectively. The output of the network consisted of a real vector with length equal to the number of possible choices. The final choice was obtained by applying a softmax function to this vector and sampling from the resulting probabilities.

Tasks

All tasks were implemented using the NeuroGym toolbox (https://neurogym.github.io/). Tasks were organized into trials. Trials had a fixed duration and were composed of 3-4 time steps: fixation, stimulus (1-2 samples, mean=1.3) and decision period. There was no inter-trial interval. At the decision time-step, the network had to choose the action associated with the stimulus that was larger on average. The reward given to the network was 0 throughout the trial, except at the decision time-step, when a reward of 1 was given to the network if it made the correct choice and 0 otherwise.

Two-Alternative Forced Choice (2AFC) task. As in the experiments with rats, the stimulus trial sequence in the categorization 2AFC task was generated by a two-state Markov chain parametrized by the repeating probability P_{rep} which is varied in Repeating (P_{rep}=0.8) and Alternating trial blocks (P_{rep}=0.2). At the end of every trial context was changed with probability p = 0.0125, yielding an average block length of 80 trials, which is the block length used in most of the rat experiments (Groups ILD-0.8, ILD-0.95, S0.8, S0.95).

N-Alternative Forced Choice (NAFC) task pre-training. As in the 2AFC task, the stimulus trial sequence is organized into blocks with distinct serial correlations. Correlations are generated by a N-state Markov chain parametrized by the corresponding transition matrices. We pre-trained the networks in 3 different contexts: repeating, clockwise and anticlockwise, wherein the most likely transition (with probability = 0.8) was a repetition, a clockwise transition and an anticlockwise transition, respectively. As in the 2AFC task, at the end of every trial, the context was changed with probability p = 0.0125. The number of choices was also randomly chosen between 3 and the maximum number of choices (N_{max}) allowed every 5000 trials (except where specified, the percentage of trials with N = 2 was fixed to 1%). This number was set to be large enough so networks had time to habituate to different numbers of choices. In a given block with the number of choices set to N_0, stimuli associated with choices $N > N_0$ were set to 0.

Training of the RNNs

Training was done using an actor-critic deep reinforcement learning algorithm with experience replay (ACER) (Z. Wang et al. 2016) (Stable-Baselines toolbox (https://github.com/hill-a/stable-baselines). We initially tested other algorithms (A2C (Mnih et al. 2016), ACKTR (Wu et al. 2017) and PPO2 (Schulman et al. 2017)) and chose ACER because it was the one that learned the NAFC task faster. However, preliminary analyses of the networks trained with these different algorithms did not show any clear difference in the strategy they used to solve the task.

Actor-critic (Sutton and Barto 2018). Actor-critic methods use a separate function (the actor) to explicitly represent the policy independently of the function (the critic, or value function) that represents the value of each environment state. In ACER, a single deep neural network infers both the policy and the value function.

Experience replay (Lin 1992). The experience replay approach allows the network to remember past experiences and use them to learn from them off-line.

Training hyperparameters. The discount factor was set to 0.99 and the learning rate to 7×10^{-4}. Weights were optimized using RMSProp and backpropagation through time, with a rollout of 15 time-steps (~6 trials). The number of threads was set to 20. The ACER algorithm estimates the gradient of the Kullback–Leibler (KL) divergence between the old and updated policy and uses it to determine step size. Other hyperparameters were kept as in the original paper (weight for the loss on the Q value = 0.5, weight for the entropy loss = $1x10^{-2}$, clipping value for the maximum gradient = 10, scheduler for the learning rate update = 'linear', RMSProp decay parameter = 0.99, RMSProp epsilon = 1×10^{-5}, buffer size in number of steps = 5×10^{3}, number of replay learning per on policy learning on average, using a poisson distribution = 4, minimum number of steps in the buffer, before learning replay =1×10^{3}, importance weight clipping factor = 10.0, decay rate for the Exponential moving average of the parameters = 0.99, max KL divergence between the old policy and updated policy = 1).

Testing the RNNs

After the training, the networks were tested on the standard 2AFC task, with their weights frozen.

Data code availability

Data and code will be made available on a public repository once the manuscript is published.

Acknowledgements

We thank Jorge del Pozo for preliminary analyses. We thank Ainhoa Hermoso-Mendizabal for useful discussion and Lorena Jiménez for help with training of the animals. This research was supported by the Beatriu de Pinós fellowship, Generalitat de Catalunya (2017-BP-00305 to M.M.M), the Spanish Ministry of Economy and Competitiveness together with the European Regional Development Fund (IJCI-2016-29358 to D.D.; RTI2018-099750-B-I00 to J.R.), the European Research Council (ERC-2015-CoG - 683209 Priors to J.R.) and the Simons Foundation Junior Fellowship, NSF NeuroNex Award 589 DBI-1707398, and the Gatsby Charitable Foundation which supported G.R.Y..

Author contributions

M.M.M. and J.R. conceived the project; M.M.M., G.R.Y. and J.R. designed the model; D.D. and J.R. designed the experiments; D.D. carried the experiments; M.M.M. and D.D., analyzed the experimental data; M.M.M. trained and analyzed the networks; M.M.M., D.D., G.R.Y. and J.R. interpreted the data; M.M.M. and J.R. wrote the manuscript with contributions from D.D. and G.R.Y.

Competing interests

The authors declare no competing interests.

References

Abrahamyan, Arman, Laura Luz Silva, Steven C. Dakin, Matteo Carandini, and Justin L. Gardner. 2016. "Adaptable History Biases in Human Perceptual Decisions." *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1518786113.

Addicott, M. A., J. M. Pearson, M. M. Sweitzer, D. L. Barack, and M. L. Platt. 2017. "A Primer on Foraging and the Explore/Exploit Trade-Off for Psychiatry Research." *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology* 42 (10): 1931–39.

Akaishi, Rei, Kazumasa Umeda, Asako Nagase, and Katsuyuki Sakai. 2014. "Autonomous Mechanism of Internal Choice Estimate Underlies Decision Inertia." *Neuron* 81 (1): 195–206.

Akrami, Athena, Charles D. Kopec, Mathew E. Diamond, and Carlos D. Brody. 2018. "Posterior Parietal Cortex Represents Sensory History and Mediates Its Effects on Behaviour." *Nature* 554 (7692): 368–72.

Alves, Hans, Alex Koch, and Christian Unkelbach. 2017. "Why Good Is More Alike Than Bad: Processing Implications." *Trends in Cognitive Sciences* 21 (2): 69–79.

Barak, Omri. 2017. "Recurrent Neural Networks as Versatile Tools of Neuroscience Research." *Current Opinion in Neurobiology* 46 (October): 1–6.

Baumeister, Roy F., Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D. Vohs. 2001. "Bad Is Stronger than Good." *Review of General Psychology: Journal of Division 1, of the American Psychological Association* 5 (4): 323–70.

Blanchard, Tommy C., and Samuel J. Gershman. 2018. "Pure Correlates of Exploration and Exploitation in the Human Brain." *Cognitive, Affective & Behavioral Neuroscience* 18 (1): 117–26.

Braun, Anke, Anne E. Urai, and Tobias H. Donner. 2018. "Adaptive History Biases Result from Confidence-Weighted Accumulation of Past Choices." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 38 (10): 2418–29.

Busse, Laura, Asli Ayaz, Neel T. Dhruv, Steffen Katzner, Aman B. Saleem, Marieke L. Schölvinck, Andrew D. Zaharia, and Matteo Carandini. 2011. "The Detection of Visual Contrast in the Behaving Mouse." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 31 (31): 11351–61.

Carnevale, Federico, Victor de Lafuente, Ranulfo Romo, Omri Barak, and Néstor Parga. 2015. "Dynamic Control of Response Criterion in Premotor Cortex during Perceptual Detection under Temporal Uncertainty." *Neuron* 86 (4): 1067–77.

Chakroun, Karima, David Mathar, Antonius Wiehler, Florian Ganzer, and Jan Peters. 2020. "Dopaminergic Modulation of the Exploration/exploitation Trade-off in Human Decision-Making." *eLife* 9 (June). https://doi.org/10.7554/eLife.51260.

Corrado, Greg S., Leo P. Sugrue, H. Sebastian Seung, and William T. Newsome. 2005. "LINEAR-NONLINEAR-POISSON MODELS OF PRIMATE CHOICE DYNAMICS." *Journal of the Experimental Analysis of Behavior*. https://doi.org/10.1901/jeab.2005.23-05.

Dahl, George E., Dong Yu, Li Deng, and Alex Acero. 2012. "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition." *IEEE*

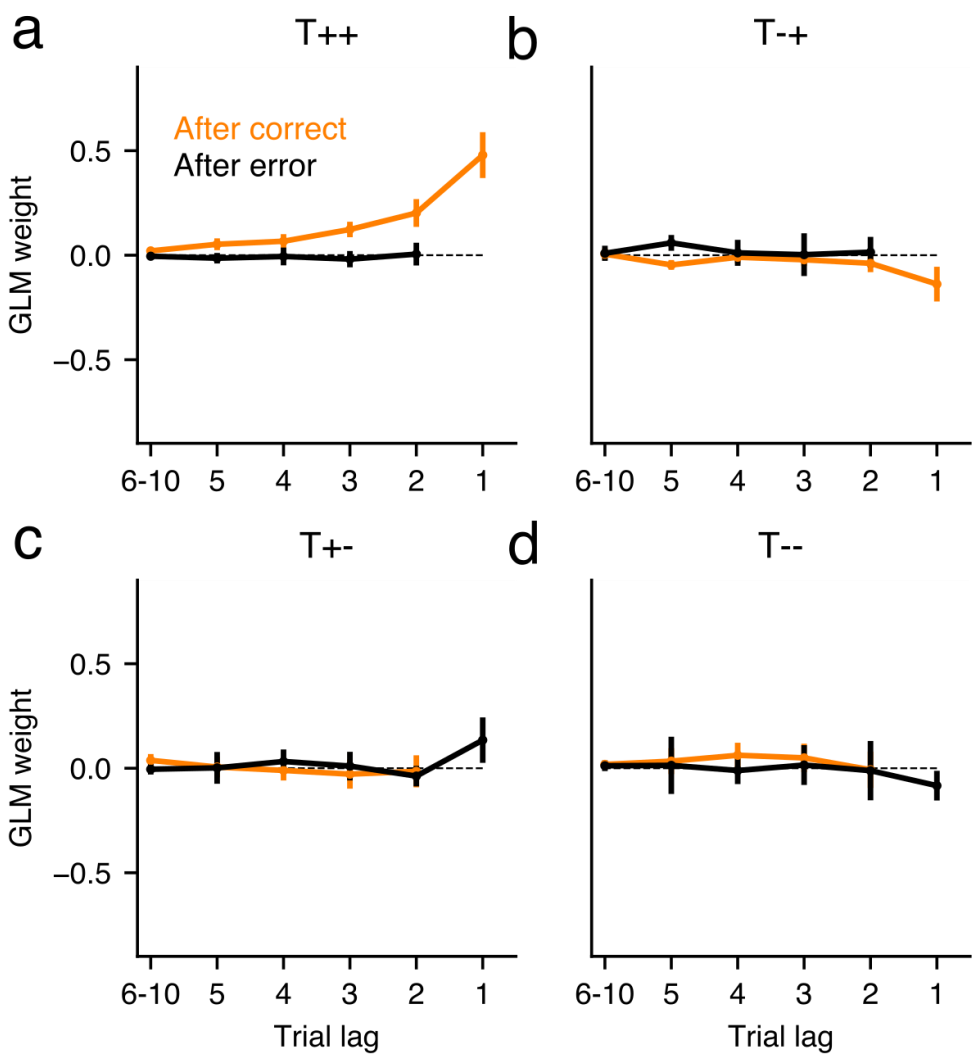
Transactions on Audio, Speech, and Language Processing 20 (1): 30–42.
Daw, Nathaniel D., John P. O'Doherty, Peter Dayan, Ben Seymour, and Raymond J. Dolan. 2006. "Cortical Substrates for Exploratory Decisions in Humans." *Nature* 441 (7095): 876–79.
Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1810.04805.
Diamond, Jared. 2017. *Guns, Germs, and Steel: The Fates of Human Societies*. W. W. Norton & Company.
Donahue, Christopher H., Hyojung Seo, and Daeyeol Lee. 2013. "Cortical Signals for Rewarded Actions and Strategic Exploration." *Neuron* 80 (1): 223–34.
Faisal, A. Aldo, Luc P. J. Selen, and Daniel M. Wolpert. 2008. "Noise in the Nervous System." *Nature Reviews. Neuroscience* 9 (4): 292–303.
Feulner, Barbara, and Claudia Clopath. 2021. "Neural Manifold under Plasticity in a Goal Driven Learning Behaviour." *PLoS Computational Biology* 17 (2): e1008621.
Fischer, Jason, and David Whitney. 2014. "Serial Dependence in Visual Perception." *Nature Neuroscience*. https://doi.org/10.1038/nn.3689.
Frund, I., F. A. Wichmann, and J. H. Macke. 2014. "Quantifying the Effect of Intertrial Dependence on Perceptual Decisions." *Journal of Vision*. https://doi.org/10.1167/14.7.9.
Fusi, Stefano, Wael F. Asaad, Earl K. Miller, and Xiao-Jing Wang. 2007. "A Neural Circuit Model of Flexible Sensorimotor Mapping: Learning and Forgetting on Multiple Timescales." *Neuron* 54 (2): 319–33.
Geisler, Wilson S. 2008. "Visual Perception and the Statistical Properties of Natural Scenes." *Annual Review of Psychology* 59: 167–92.
Hadsell, Raia, Dushyant Rao, Andrei A. Rusu, and Razvan Pascanu. 2020. "Embracing Change: Continual Learning in Deep Neural Networks." *Trends in Cognitive Sciences* 24 (12): 1028–40.
Hermoso-Mendizabal, Ainhoa, Alexandre Hyafil, Pavel E. Rueda-Orozco, Santiago Jaramillo, David Robbe, and Jaime de la Rocha. 2020. "Response Outcomes Gate the Impact of Expectations on Perceptual Decisions." *Nature Communications* 11 (1): 1057.
Hernández-Navarro, Lluís, Ainhoa Hermoso-Mendizabal, Daniel Duque, Jaime de la Rocha, and Alexandre Hyafil. 2020. "A Race between Proactive and Reactive Processes during Perceptual Decisions." https://doi.org/10.31234/osf.io/9k84v.
Hilt, Donald E., and Donald W. Seegrist. 1977. *Ridge, a Computer Program for Calculating Ridge Regression Estimates*. Department of Agriculture, Forest Service, Northeastern Forest Experiment Station.
Hochreiter, S., and J. Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9 (8): 1735–80.
Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.
Kahneman, Daniel, and Amos Tversky. 2012. "Choices, Values, and Frames." In *Handbook of the Fundamentals of Financial Decision Making*, 4:269–78. World Scientific Handbook in Financial Economics Series. WORLD SCIENTIFIC.
Karlsson, Mattias P., Dougal G. R. Tervo, and Alla Y. Karpova. 2012. "Network Resets in Medial Prefrontal Cortex Mark the Onset of Behavioral Uncertainty."

Science 338 (6103): 135–39.
Lau, Brian, and Paul W. Glimcher. 2005. “Dynamic Response-by-Response Models of Matching Behavior in Rhesus Monkeys.” *Journal of the Experimental Analysis of Behavior* 84 (3): 555–79.
Lin, Long-Ji. 1992. “Self-Improving Reactive Agents Based on Reinforcement Learning, Planning and Teaching.” *Machine Learning* 8 (3): 293–321.
Lyon, Pamela, and Franz Kuchling. 2021. “Valuing What Happens: A Biogenic Approach to Valence and (potentially) Affect.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 376 (1820): 20190752.
Mante, Valerio, David Sussillo, Krishna V. Shenoy, and William T. Newsome. 2013. “Context-Dependent Computation by Recurrent Dynamics in Prefrontal Cortex.” *Nature* 503 (7474): 78–84.
Mastrogiuseppe, Francesca, and Srdjan Ostojic. 2018. “Linking Connectivity, Dynamics, and Computations in Low-Rank Recurrent Neural Networks.” *Neuron* 99 (3): 609–23.e29.
Ma, Wei Ji, and Benjamin Peters. 2020. “A Neural Network Walks into a Lab: Towards Using Deep Nets as Models for Human Behavior.” *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2005.02181.
McDougle, Samuel D., Matthew J. Boggess, Matthew J. Crossley, Darius Parvin, Richard B. Ivry, and Jordan A. Taylor. 2016. “Credit Assignment in Movement-Dependent Reinforcement Learning.” *Proceedings of the National Academy of Sciences of the United States of America* 113 (24): 6797–6802.
Miconi, Thomas. 2017. “Biologically Plausible Learning in Recurrent Neural Networks Reproduces Neural Dynamics Observed during Cognitive Tasks.” *eLife* 6 (February). https://doi.org/10.7554/eLife.20899.
Mnih, Volodymyr, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. “Asynchronous Methods for Deep Reinforcement Learning.” In *Proceedings of The 33rd International Conference on Machine Learning*, edited by Maria Florina Balcan and Kilian Q. Weinberger, 48:1928–37. Proceedings of Machine Learning Research. New York, New York, USA: PMLR.
Niv, Yael. 2009. “Reinforcement Learning in the Brain.” *Journal of Mathematical Psychology* 53 (3): 139–54.
Pardo-Vazquez, Jose L., Juan R. Castiñeiras-de Saa, Mafalda Valente, Iris Damião, Tiago Costa, M. Inês Vicente, André G. Mendonça, Zachary F. Mainen, and Alfonso Renart. 2019. “The Mechanistic Foundation of Weber’s Law.” *Nature Neuroscience* 22 (9): 1493–1502.
Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. “Scikit-Learn: Machine Learning in Python.” *The Journal of Machine Learning Research* 12: 2825–30.
Purcell, Braden A., and Roozbeh Kiani. 2016a. “Neural Mechanisms of Post-Error Adjustments of Decision Policy in Parietal Cortex.” *Neuron* 89 (3): 658–71.
———. 2016b. “Hierarchical Decision Processes That Operate over Distinct Timescales Underlie Choice and Changes in Strategy.” *Proceedings of the National Academy of Sciences of the United States of America* 113 (31): E4531–40.
Rabbitt, Patrick, and Bryan Rodgers. 1977. “What Does a Man Do after He Makes

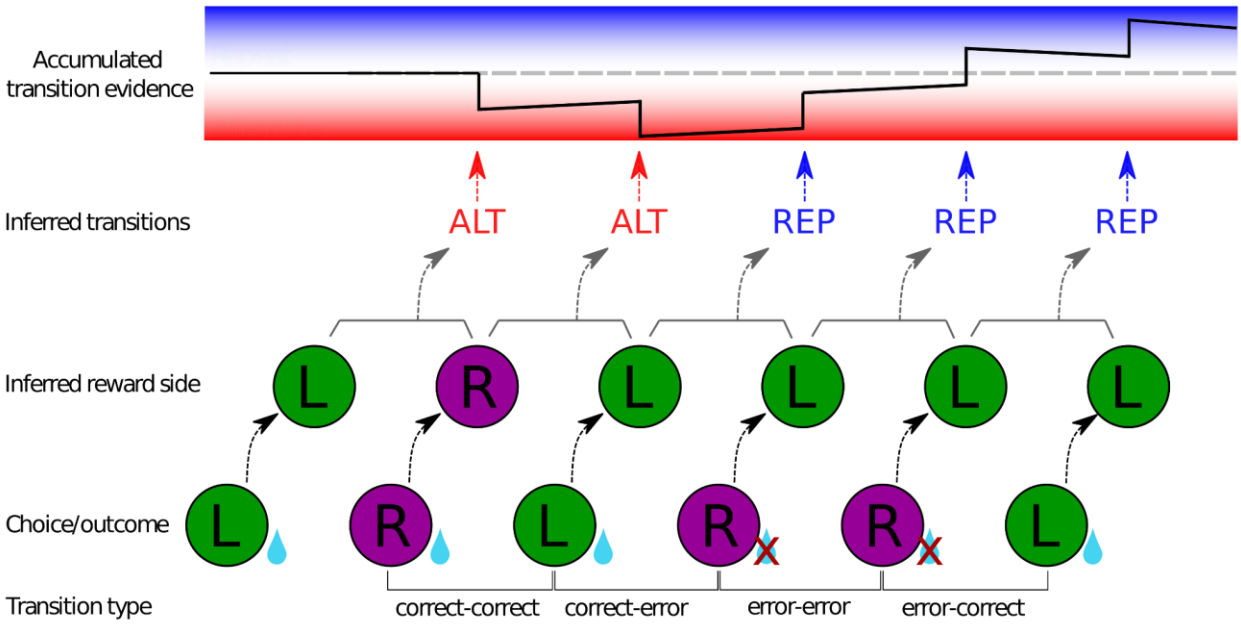
an Error? An Analysis of Response Programming." *The Quarterly Journal of Experimental Psychology* 29 (4): 727–43.
Remington, Evan D., Devika Narain, Eghbal A. Hosseini, and Mehrdad Jazayeri. 2018. "Flexible Sensorimotor Computations through Rapid Reconfiguration of Cortical Dynamics." *Neuron* 98 (5): 1005–19.e5.
Roseboom, Warrick, Zafeirios Fountas, Kyriacos Nikiforou, David Bhowmik, Murray Shanahan, and Anil K. Seth. 2019. "Activity in Perceptual Classification Networks as a Basis for Human Subjective Time Perception." *Nature Communications* 10 (1): 267.
Sarafyazd, Morteza, and Mehrdad Jazayeri. 2019. "Hierarchical Reasoning by Neural Circuits in the Frontal Cortex." *Science* 364 (6441). https://doi.org/10.1126/science.aav8911.
Saxena, S., A. Russo, J. Cunningham, and M. M. Churchland. 2021. "Motor Cortex Activity across Movement Speeds Is Predicted by Network-Level Strategies for Generating Muscle Activity." *bioRxiv*. https://www.biorxiv.org/content/10.1101/2021.02.01.429168v1.abstract.
Schulman, John, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. "Proximal Policy Optimization Algorithms." *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1707.06347.
Sohn, Hansem, Devika Narain, Nicolas Meirhaeghe, and Mehrdad Jazayeri. n.d. "Bayesian Computation through Cortical Latent Dynamics." https://doi.org/10.1101/465419.
Stoianov, Ivilin, and Marco Zorzi. 2012. "Emergence of a 'Visual Number Sense' in Hierarchical Generative Models." *Nature Neuroscience* 15 (2): 194–96.
Sussillo, David. 2014. "Neural Circuits as Computational Dynamical Systems." *Current Opinion in Neurobiology* 25 (April): 156–63.
Sussillo, David, Mark M. Churchland, Matthew T. Kaufman, and Krishna V. Shenoy. 2015. "A Neural Network That Finds a Naturalistic Solution for the Production of Muscle Activity." *Nature Neuroscience* 18 (7): 1025–33.
Sutton, Richard S., and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction*. A Bradford Book.
Tan, Mingxing, and Quoc V. Le. 2019. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks." *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1905.11946.
Urai, Anne E., Jan Willem de Gee, Konstantinos Tsetsos, and Tobias H. Donner. 2019. "Choice History Biases Subsequent Evidence Accumulation." *eLife* 8 (July). https://doi.org/10.7554/eLife.46331.
Vulkan, Nir. 2000. "An Economist's Perspective on Probability Matching." *Journal of Economic Surveys* 14 (1): 101–18.
Wang, Jane X., Zeb Kurth-Nelson, Dharshan Kumaran, Dhruva Tirumala, Hubert Soyer, Joel Z. Leibo, Demis Hassabis, and Matthew Botvinick. 2018. "Prefrontal Cortex as a Meta-Reinforcement Learning System." *Nature Neuroscience* 21 (6): 860–68.
Wang, Ziyu, Victor Bapst, Nicolas Heess, Volodymyr Mnih, Remi Munos, Koray Kavukcuoglu, and Nando de Freitas. 2016. "Sample Efficient Actor-Critic with Experience Replay." *arXiv Preprint arXiv:1611. 01224*.
Weiss, Yair, Eero P. Simoncelli, and Edward H. Adelson. 2002. "Motion Illusions as

Optimal Percepts." *Nature Neuroscience* 5 (6): 598–604.
Werbos, P. J. 1990. "Backpropagation through Time: What It Does and How to Do It." *Proceedings of the IEEE. Institute of Electrical and Electronics Engineers* 78 (10): 1550–60.
Wu, Yuhuai, Elman Mansimov, Shun Liao, Roger Grosse, and Jimmy Ba. 2017. "Scalable Trust-Region Method for Deep Reinforcement Learning Using Kronecker-Factored Approximation." *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1708.05144.
Yang, Guangyu Robert, Madhura R. Joglekar, H. Francis Song, William T. Newsome, and Xiao-Jing Wang. 2019. "Task Representations in Neural Networks Trained to Perform Many Cognitive Tasks." *Nature Neuroscience* 22 (2): 297–306.
Yang, Guangyu Robert, and Manuel Molano-Mazon. n.d. "Next-Generation of Recurrent Neural Network Models for Cognition." https://doi.org/10.31234/osf.io/w34n2.
Yang, Guangyu Robert, and Xiao-Jing Wang. 2021. "Artificial Neural Networks for Neuroscientists: A Primer." *Neuron* 109 (4): 739.

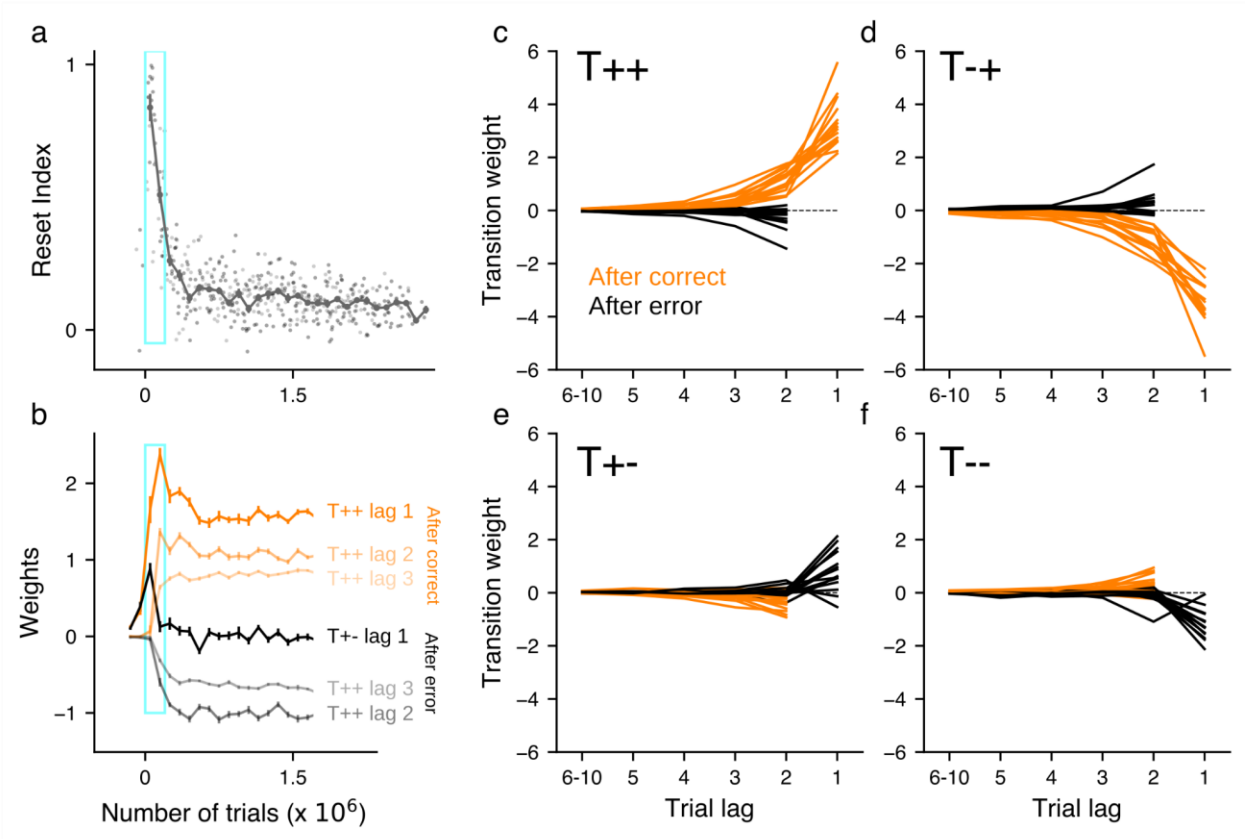
Supplementary figures



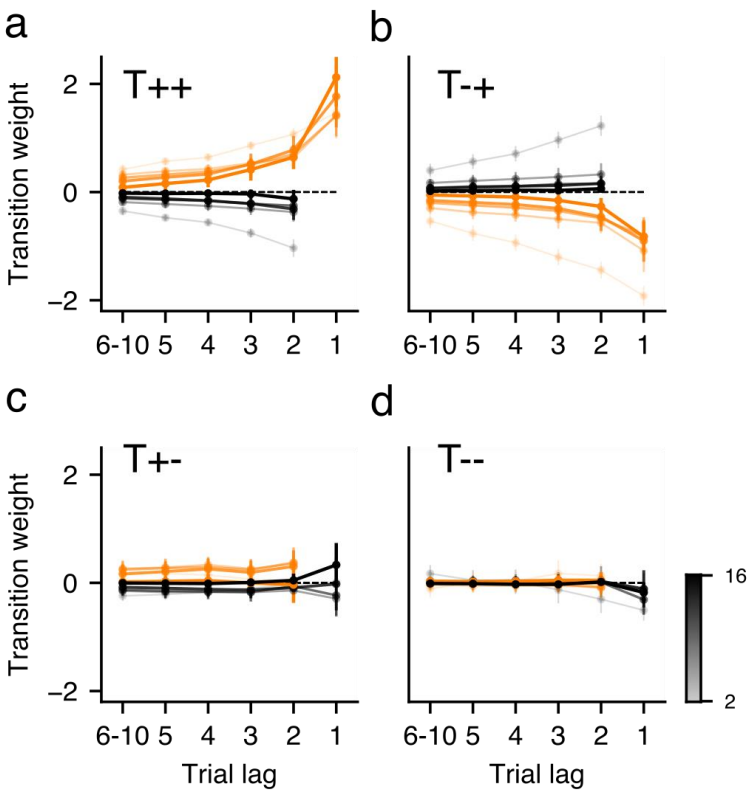
Supplementary figure 1 (related to Fig. 1). Only transitions composed by two correct choices influence the decision of rats. The figure shows all transition kernels obtained from the rats behavior separately fitted for after-correct and after-error trials (see color code in a). Each panel shows the contribution of the different types of transitions: T++: transitions made of two consecutive rewarded trials or correct-correct transitions **(a)**, T-+: error-correct transitions **(b)**, T+-: correct-error transitions **(c)**, T--: error-error transitions **(d)**. Dark lines show mean values (Group ILD-0.8, n=8). Error-bars correspond to standard deviation.



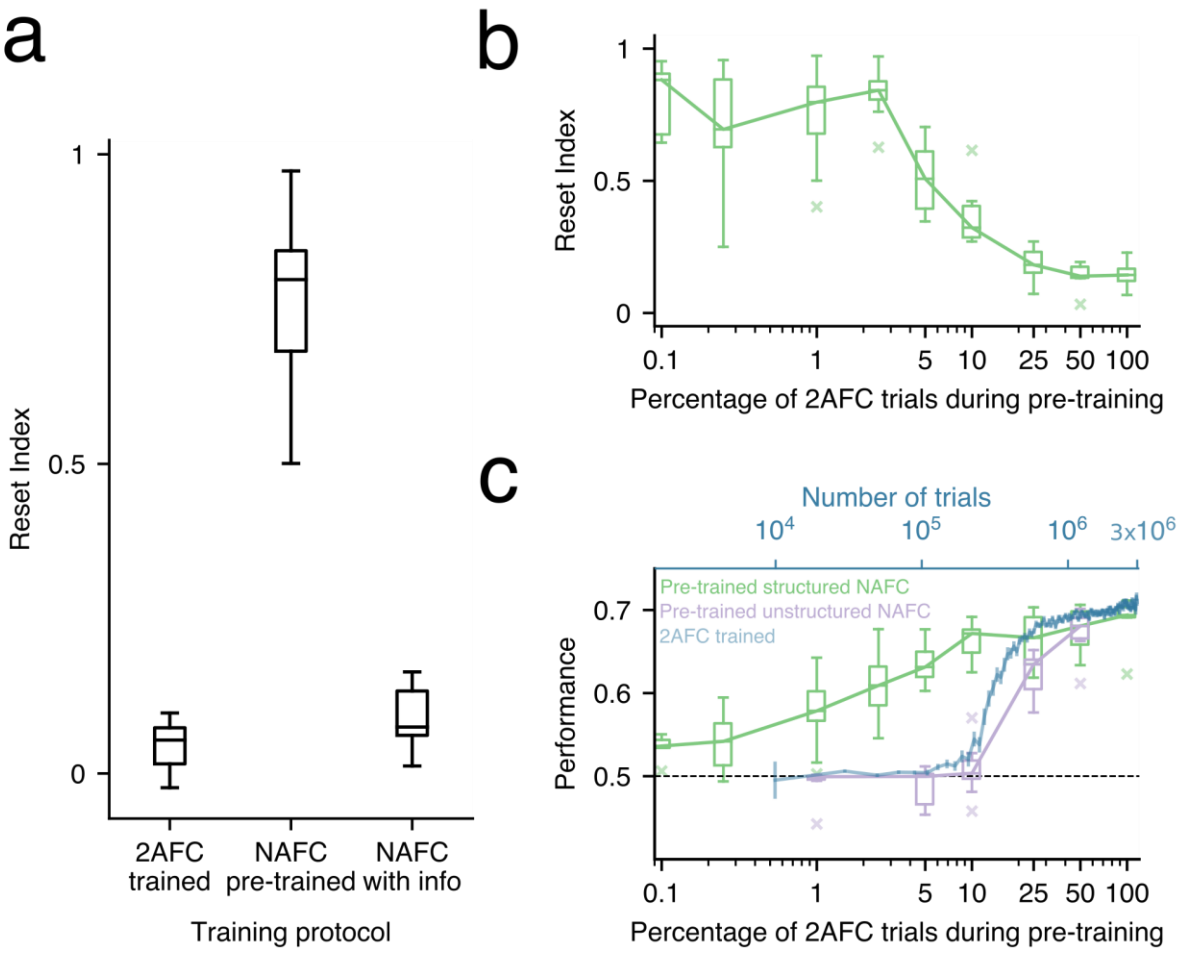
Supplementary figure 2 (related to Fig. 1). Inference of ground-truth reward sides and transitions, and accumulation of transition evidence made by an optimal agent for different transition types. Bottom: example sequence of choices and outcomes. The type of transition connecting each pair of consecutive choices is indicated in the bottom. **Middle:** after making a choice and learning the outcome, the agent infers the rewarded side (black dashed arrows). Having inferred two consecutive rewarded sides, the agent infers the last ground-truth transition (dark gray dashed arrows). **Top:** Each inferred ground-truth transition is used to update the accumulated transition evidence (blue/red dashed arrows). Notice that following this inference procedure, an optimal agent can use each transition to update the accumulated evidence independently of its type (i.e. ++, +-, -- and -+). The agent can then use this accumulated evidence to bias the upcoming choices (not shown for simplicity).



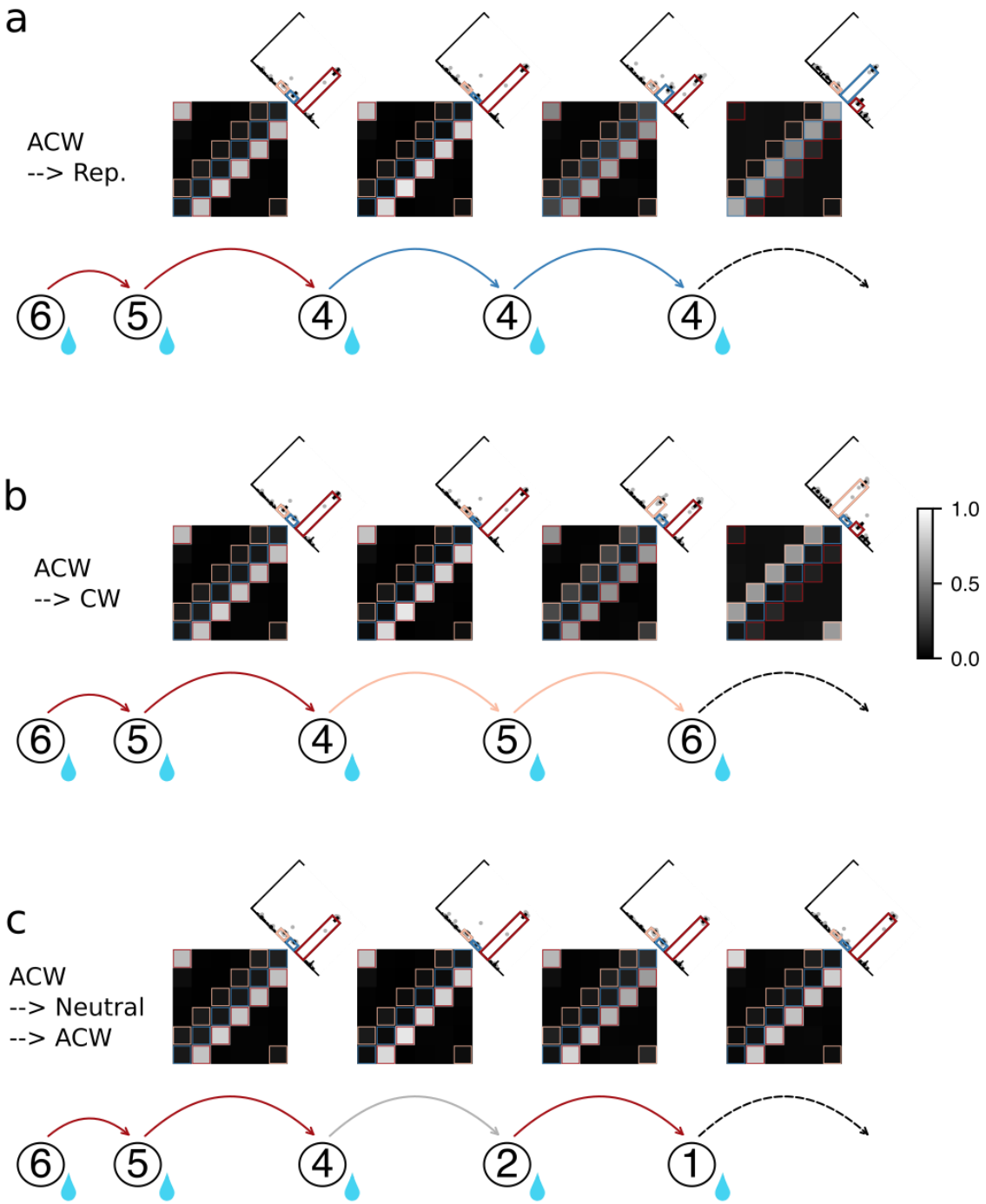
Supplementary figure 3 (related to Fig. 2). Networks trained directly in the 2AFC momentarily display the reset strategy before reaching the reverse strategy. a) Reset Index evolution shown in Fig. 2e. Cyan box highlights the period of training in which the reset index peaks and from which the kernels in c-f are obtained. **b)** Evolution of the transition weights at different lags across training. The maximum value reached by the contribution of the T+- regressor at lag 1 (black trace) coincides with the Reset Index peak, which suggests that the networks are quickly transitioning through a strategy that only uses the latest transition (dark orange and black traces) to infer the context. Therefore the peak in the Reset Index arises because of a short delay between the increase in the weight of the T++ regressors after correct (orange traces) and the decrease of the weights of those same regressors after error (gray traces). Both Reset Index and the transition weights traces are aligned to the aha moment for each network. **c-f)** Transition kernels for RNNs obtained from the period of training at which the Reset Index peaks for each network.



Supplementary figure 4 (related to Fig. 4). Average transition kernels obtained from pre-trained RNNs when testing in the 2AFC task, for pre-training NAFC environments with a different maximum number of alternatives, from N_{max}=2 (thick, light traces) to N_{max}=16 (thin, dark traces). Notice that as N_{max} increases, all kernels vanish , except the T++ after-correct choices.



Supplementary figure 5 (related to Fig. 4). Reset strategy in pre-trained networks depends on features of pre-training. a) Comparison of the Reset index for RNNs directly trained in the 2AFC task (as in Fig. 2), pre-trained in standard NAFC (as in Fig. 4) and RNNs pre-trained in a variant of the NAFC task in which we provided as an input to the network the correct alternative in the previous timestep ("NAFC with info"). **b-c)** Reset index (b) and accuracy in trials with no stimulus evidence (c) as a function of percentage of 2AFC trials embedded in the NAFC pre-training, for networks pre-trained in the standard NAFC (Structured NAFC, green) and networks pre-trained in a NAFC environment with no serial correlations when N>2 (i.e. transition matrices were uniform, Unstructured NAFC, purple). As a reference, the performance of networks trained directly on the 2AFC task for a comparable amount of trials is shown (top axis in c). Only RNNs that developed a transition bias were used to compute the Reset Index values. Both the structured and the unstructured NAFC pre-training were done with N_{max} = 16. x-axis in log-scale in b and c.



Supplementary figure 6 (related to Fig. 5). a, b) The evolution of the transition bias matrix computed using pre-trained RNNs (Nmax= 16, RNNs are tested in N=6) throughout choice sequences involving an unexpected correct response occurring in context change-point (switching from the ACW to the Rep context in a and from the ACW to the CW in b). Insets above each matrix show histograms quantifying the trial-by-trial estimates of being in a Repeating (blue bar), Clockwise (pink) and Anticlockwise (red) context (as in Fig. 5). The first unexpected correct transition at the change-point induces a decrease in the networks estimate to be in the old context and a slight increase in favor of the new context; the second unexpected correct transition makes the networks completely switch their internal estimate towards the new context. **c)** Evolution of the transition matrix throughout a sequence with correct unexpected neutral transition, i.e. a

transition which is incongruent with any of the contexts. Neutral transitions, which constitute a significant portion of incongruent transitions which are themselves 20% of the total, had only a slight effect on the transition matrices.