Humans adaptively select different computational strategies in different learning environments

Pieter Verbeke[1], Tom Verguts[1]

[1]Department of Experimental Psychology; Ghent University; Ghent; B9000

Abstract

The Rescorla-Wagner rule remains the most popular tool to describe human behavior in reinforcement learning tasks. Nevertheless, it cannot fit human learning in complex environments. Previous work proposed several hierarchical extensions of this learning rule. However, it remains unclear when a flat (non-hierarchical) versus a hierarchical strategy is optimal, or when it is implemented by humans. To address this question, current work evaluates multiple models in multiple reinforcement learning environments both computationally (which approach performs best) and empirically (which approach fits human data best). We consider ten empirical datasets (N = 410) divided over three reinforcement learning environments. Consistent with the idea of the human brain as a mixture of expert system, our results demonstrate that different environments are best solved with different learning strategies; and that humans seemed to adaptively select the optimal learning strategy. Specifically, while flat learning fitted best in less complex stable learning environments, humans employed more hierarchically complex models in more complex environments.

Keywords: Adaptive model selection, Hierarchical learning, Reinforcement learning.

Corresponding author:

Pieter Verbeke

Department of Experimental psychology at Ghent university

Henri Dunantlaan 2, Ghent, Belgium, B900

pjverbek.verbeke@ugent.be

Acknowledgements

We thank Pieter Huycke, Liyu Xia, Judith Goris, Meng Liu, Qi Chen, Dahlia Mukherjee, Michael X. Cohen and Thomas P. Hein for sharing their datasets either via open science framework or via direct communication with us. Additionally, we thank Juliet Davidow to share her stimuli with us to use for the collection of the online additional dataset. PV and TV were both supported by the Research Foundation Flanders (FWO) / Fonds Nationale de la Reserche Scientifique EOS grant G0F3818N. Additionally, PV was supported by grant 1102519N from Research Foundation Flanders. We thank Irene Cogliati-Dezza and Thomas Colin for their comments on an earlier version of the paper.

Code and data availability statement.

All code as well as the behavioral data that we collected is available on our GitHub repository: https://github.com/CogComNeuroSci/PieterV_public/tree/master/Model_Meta. For data from previous publications that we included in our analyses, we refer to the original manuscripts for links to repositories or contact information of the corresponding authors

Introduction

Humans and other primates are remarkably flexible in adapting to changing environments. For this purpose, they track of contingencies between stimuli, actions, and rewards. To date, such reward contingency learning is most often described by what is called flat (i.e., non-hierarchical) learning algorithms. Perhaps the most popular instance of flat learning is the Rescorla-Wagner (RW) model (1, 2). This RW learning rule allows to describe the behavioral learning curve by estimating a single learning rate parameter. Importantly, flat learning algorithms learn only one set of stimulus-action mappings (further referred to as rule set). Hence, each stimulus is associated to exactly one optimal action. This is limited as a model of human cognition since the most optimal action given a certain stimulus often depends on the current goal or context (i.e., the task). For example, while one needs to drive on the left side of the road in the UK, the European mainland requires driving on the right side of the road. A flat learning agent would need to constantly overwrite/ relearn this information when it switches contexts (countries).

Thus, although flat learning works well in stable and simple environments it has often proven to be insufficient to describe adaptive human behavior in more complex or volatile environments (3–7). An empirical example are probabilistic reversal learning tasks (3, 8–10). In these tasks, an agent faces two contradictory goals. On the one hand, the agent should be flexible to deal with sudden reversals of the stimulus-action-reward contingencies (plasticity). On the other hand, the agent should also be robust to the noise that is induced by the probabilistic feedback (stability). In contrast to humans (11), flat learning algorithms cannot achieve a balance between stability and plasticity. This is because high learning rates would lead the agent to “chase the noise” and adapt behavior when the contingencies did not change but low learning rates would significantly decrease the flexibility of the agent in adapting to changing contingencies. As a result, several hierarchical extensions of the RW rule have been proposed. Current work considers three of these previously proposed extensions and evaluates them in multiple reinforcement learning environments of different complexity. An in-depth model comparison over multiple environments allows to investigate when a flat versus a hierarchical strategy is optimal.

A first popular extension to the flat learning approach is to introduce a hierarchically higher parameter which allows to adaptively change lower-level parameters, such as the learning rate (3, 5, 12, 13). For instance, in reversal learning, it can be beneficial to increase the learning rate if uncertainty in the environment (measured e.g., by prediction errors) is high and decrease the learning rate if uncertainty is low (as in the Kalman filter; (14)). Such an adaptive learning rate provides robustness against noise during stable periods (by decreasing the learning rate), and flexibility when reversals occur (by increasing the learning rate). Current work considers a Hybrid parameter (see (3)) which adapts the learning rate in response to changing uncertainty in the environment.

A second extension consists of storing multiple rule sets and adaptively switch between these sets when there is a strong change in stimulus-action reward contingencies (6, 10, 15). Importantly, this extension implements a hierarchical architecture for learning. At the lower level, multiple rule sets are learned. At the higher level, the agent decides when to switch between different rule sets. Thus, if there are multiple sufficiently distinctive sets of contingencies that are frequently revisited over the time course of the task (as in reversal learning tasks), a hierarchical agent can learn multiple (robust)

rule sets and then on a hierarchically higher level decide when to switch between rule sets. Here, the switch decision is implemented as an accumulation-to-bound process in which negative feedback (i.e., absence of reward) is accumulated over time. When a bound is reached, the agent switches to another rule set. For this purpose, a Cumulation parameter is implemented which specifies the speed at which negative feedback is accumulated for the switch decision (6, 10).

Often, although not always, models with a hierarchical architecture also implement learning at the hierarchically higher level (10, 16–18). This allows to switch between different rule sets more adaptively. For instance, one could learn to map rule sets to task contexts. As a result, one could infer from context which rule set to switch to (16, 17, 19). Additionally, learning on the hierarchical level could allow to adapt to changing levels of noise in reward feedback and hence learning when to switch (10, 13, 20). Specifically, while with low levels of noise in feedback, an agent might decide to switch to another rule set every time it receives negative feedback, high levels of noise in feedback would require agents to be more conservative about switching since negative feedback might be caused by noise instead of an actual change in stimulus-action reward contingencies (21). Hence, the weight of negative feedback in the decision to switch might be decreased. Current work implements learning on the hierarchically higher level to adapt the weight of negative feedback in the accumulation-to-bound process (6, 10).

In sum, three extensions are evaluated: adaptive learning rate, hierarchical architecture, and hierarchical learning. Each extension is implemented by adding a single free parameter to the model (Hybrid, Cumulation and Higher learning rate respectively). A schematic representation of these extensions is provided in Fig. 1A. As demonstrated in the model list of Fig. 1A, six possible models can be constructed by making different combinations of the three possible extensions (see also Methods). Critically, current work evaluates the added value of each extension both computationally in terms of model performance via simulations as well as empirically in terms of model fit to human behavioral data.

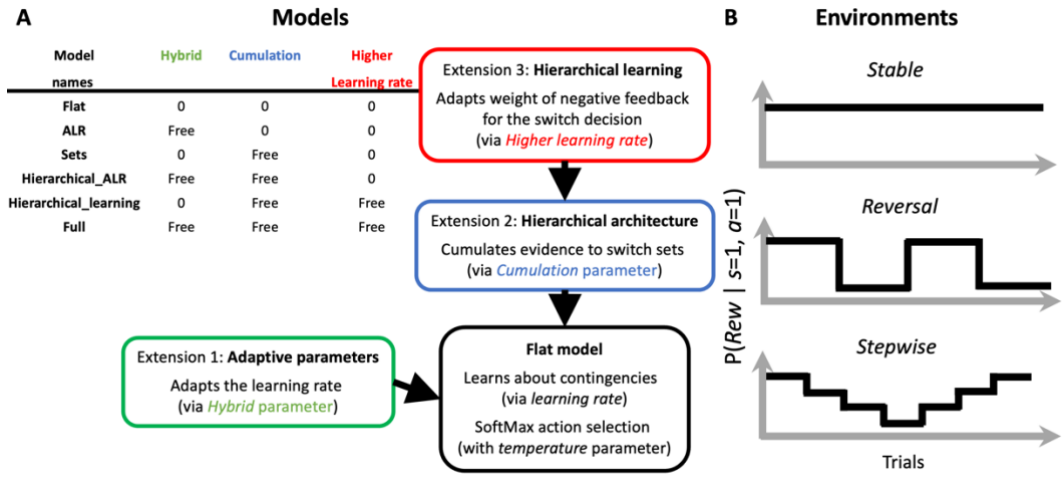

Figure 1. Study set-up. We evaluate multiple models in multiple environments. A: The models. A schematic illustration of the three extensions to the Flat model. Each extension is implemented by adding a single parameter (indicated in red, blue or green). The model list demonstrates how different combinations of these extensions result in six possible models. Here, parameters are freely estimated on the data or fixed to zero. Parameters that are fixed to zero do not influence model behavior. Each model is an extension of the Flat model, so they all have a learning rate and temperature parameter that is freely estimated (see Methods for details). B: The reinforcement learning environments. This figure provides a schematic illustration of how stimulus-action-reward contingencies evolve in each of the three environments.

Whether and which hierarchical extensions are needed, may depend on the environment. Therefore, current work considers three types of environments with different reinforcement dynamics (see Fig. 1B). One popular approach to study reinforcement learning is via two-armed bandit tasks (2, 22–24). Here, on each trial there is one optimal action ($a_{optimal}$) yielding the highest probability of reward and one suboptimal action yielding a lower probability of reward. Subjects thus need to learn what the optimal action is for a certain stimulus. Crucially, the complexity of two-armed bandit tasks can be increased by implementing more complex reinforcement dynamics. Current work evaluates multiple models in three task environments that differ in terms of the reinforcement dynamics.

In the Stable environment, one rule set needs to be learned which does not change. Hence, the identity of $a_{optimal}$ as well as the probability of reward given an optimal action (P(*Rew* | $a_{optimal}$)) remain stable over time (over trials). In the Reversal environment, two orthogonal rule sets are frequently revisited during the time course of the task. Thus, while P(*Rew* I $a_{optimal}$) remains stable over time, the identity of $a_{optimal}$ regularly reverses during the time course of the task. In the Stepwise environment, the identity of the optimal action also changed over time. However, in these datasets, also P(*Rew* I $a_{optimal}$) could change. An example of the reinforcement dynamics in each environment is illustrated in Fig. 1B.

In sum, we provide an extensive model comparison over multiple reinforcement learning environments and empirical datasets. This allows to investigate when and which hierarchical extensions to the flat learning RW model are optimal. Specifically, six models are created by combining three possible extensions: adaptive learning rate, hierarchical architecture, and hierarchical learning (see Fig. 1A). These models are tested on the Stable, Reversal and Stepwise environments (see Fig. 1B). By using model simulations, we illustrate that different models are optimal in different environments. By fitting all models on a wide range of human behavioral datasets, we also demonstrate that humans select different computational strategies for these different environments.

Results

Descriptive statistics

We collected ten empirical datasets across three types of environments: Three in the Stable environment (25–27), five in the Reversal environment (4, 10, 25, 28) and two in the Stepwise environment (29, 30). Specifics of each dataset are described in the Supplementary materials and summarized in Fig. 2A. In Fig. 2B, we present some descriptive analyses for each dataset. Accuracy is defined as the proportion of optimal actions (i.e., actions that would elicit reward if feedback was deterministic; $P(a_{optimal})$). In general, accuracy decreases in more complex environments. Additionally, accuracy decreases for lower values of $P(Rew|a_{optimal})$. This is observed as a decreasing accuracy trend from left to right in each grey box of Fig. 2B. Interestingly, the Goris/Volatile dataset, seems to be a strong outlier in this trend. While $P(Rew|a_{optimal})$ = 90% in this dataset, accuracy is lower than for Reversal datasets where $P(Rew|a_{optimal})$ is 85 or 80%.

A

Datasets	Environment	$P(Rew\|a_{optimal})$	Subjects	Trials
Huycke	Stable	100%	24	512
Xia	Stable	80%	104	120
Goris/Stable	Stable	70%	38	90
Online	Reversal	100%	49	220
Goris/Volatile	Reversal	90%	49	90
Liu	Reversal	85%	23	320
Verbeke	Reversal	80%	30	480
Mukherjee	Reversal	70%	58	90
Hein	Stepwise	90, 70 & 50% (Mean = 74%)	20	400
Cohen	Stepwise	75 & 50% (Mean = 66%)	15	1200

B

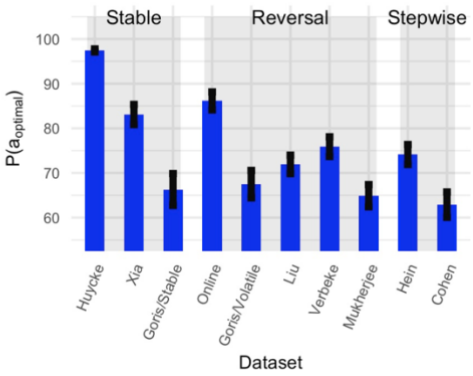

Figure 2. Datasets. A: Summarizing list. This provides some basic information about each dataset. B: Descriptive statistics. Blue bars indicate the average accuracy (defined as proportion of optimal actions; $P(a_{optimal})$) across all subjects for each dataset. Black lines represent the 95% confidence interval. Datasets are grouped for different environments (grey boxes) and sorted for $P(Rew|a_{optimal})$ in descending order.

Humans adaptively select the most optimal model

First, model performance was evaluated in each environment. Because we aimed to evaluate performance of each model under optimal parameter settings, we first estimated for each model the optimal set of parameters in each environment (optimal in the sense of maximizing reward). Then, each model was simulated in each environment with its respective optimal set of parameters. For comparison, we use a weighted accuracy measure (see Methods for details). Hence, each cell in Fig. 3A represents the relative accuracy for a specific model (columns) in a specific environment (rows) compared to the accuracy of the other five models in that environment (sum of each row is 1).

Second, model fit to empirical data was evaluated. For this purpose, each model was fitted (by maximizing log likelihood) on each behavioral dataset. Empirical support for each model is then evaluated by using weighted Akaike information criterion values (wAIC). This provides a relative (compared to other models) measure of fit (higher values means better fit) and incorporates a penalty for the number of freely estimated parameters. Hence, each cell in Fig. 3B represents the relative goodness of fit for a specific model (columns) in a specific dataset (rows).

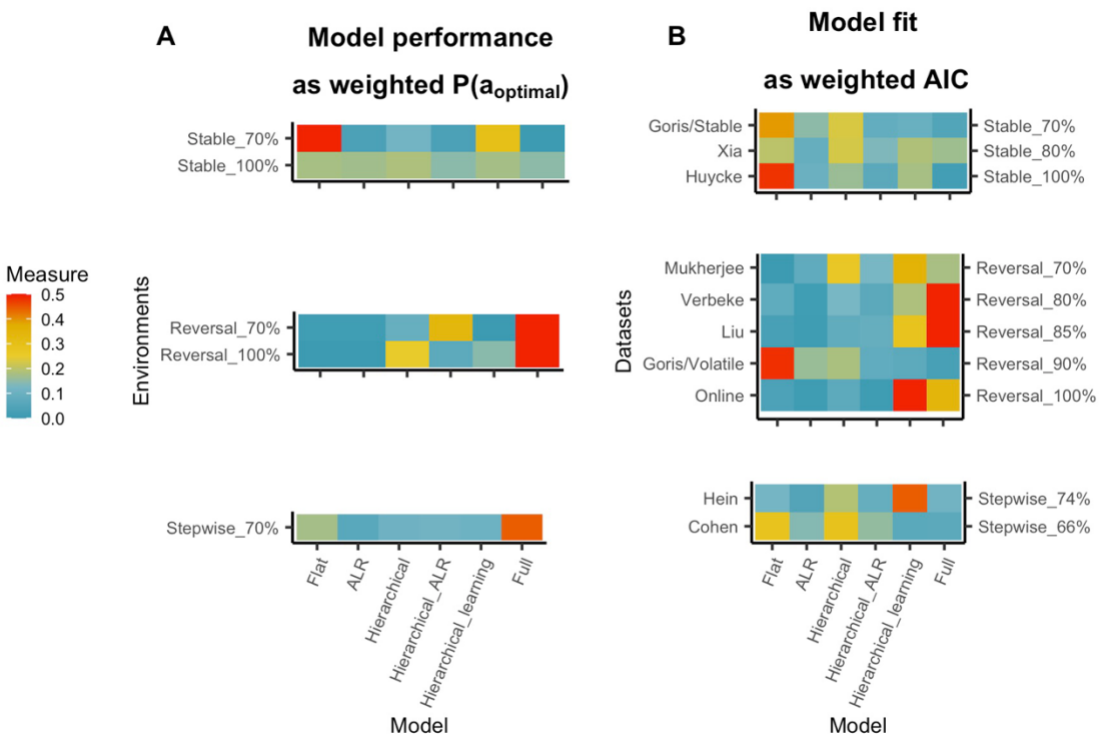

Figure 3. Simulation and fitting results. A: Model performance. Weighted accuracy (defined as $P(a_{optimal})$) for model simulations in each environment. Accuracy is weighted to allow comparison across models, each row sums to 1. B: Model fit. weighted AIC values for each model on each behavioral dataset. Model evidence is weighted to allow comparison across models, each row sums to 1. Datasets are again grouped per environment and sorted for $P(Rew|a_{optimal})$.

For Stable environments, the Flat model performs best if $P(Rew|a_{optimal})$ = 70% (Fig. 3A). If $P(Rew|a_{optimal})$ = 100%, no significant difference is observed in performance across models. Hence, also considering the fewer parameters in the Flat model, simulations predict that flat learning is most optimal in Stable environments. Consistent with this prediction from optimality, over all empirical datasets in the Stable environment there is clearly most evidence for the Flat model (Fig. 3B).

For the Reversal environment, a different picture emerges. Here, model performance (Fig. 3A) suggests that the Full model is optimal, irrespective of the level of $P(Rew|a_{optimal})$. Critically, this Full model implements all three hierarchical extensions: an adaptive learning rate, hierarchical architecture, and hierarchical learning. Consistent with this prediction from optimality, there is indeed most empirical evidence for hierarchical models (either Full or the closely related Hierarchical_learning models), with one exception. The Goris/Volatile dataset strongly prefers the Flat model. Importantly, however, for this dataset accuracy was also very low (see Fig. 2B). Hence, performance was not optimal in this dataset, potentially explaining the good fit of the less optimal flat model.

For the Stepwise environment, model simulations (Fig. 3A) predict that also here the Full model is optimal. Nevertheless, considering that the Flat model has fewest parameters, and the weighted accuracy does not consider the number of free parameters (see Methods) it is noteworthy that the Flat model is the second-best model in the simulations. This mixed evidence also appears in model fit (Fig. 3B). For the Cohen dataset, we observe mixed evidence between the Flat and Hierarchical model.

For the Hein datasets, there is most evidence for the Hierarchical_learning model, although there is evidence for the Hierarchical model as well.

In sum, results suggest that humans adaptively select the optimal model. Which model is optimal differs across environments: In the Stable environment, the Flat model is most used. When the environment gets more complex, there is also more evidence for more complex models.

Humans adaptively combine hierarchical extensions

Crucially, the main interest of the current study was to investigate the added value of each individual hierarchical extension in each of the three environments. To shed further light on this, the left column of Fig. 4 (panels A, C and E) presents the weighted accuracy of model simulations averaged for each combination of environment and hierarchical extension. As could be expected from the results in Fig. 3A, simulated model performance predicts no advantage for hierarchical extensions in the Stable environment (Fig. 4A). In the Reversal environment (Fig. 4C), an advantage is expected for each hierarchical extension. In Stepwise environments (Fig. 4E), there is a strong advantage for hierarchical learning. Although there is an advantage for hierarchical architecture and adaptive learning rates as well, this is considerably less strong compared to the hierarchical learning advantage.

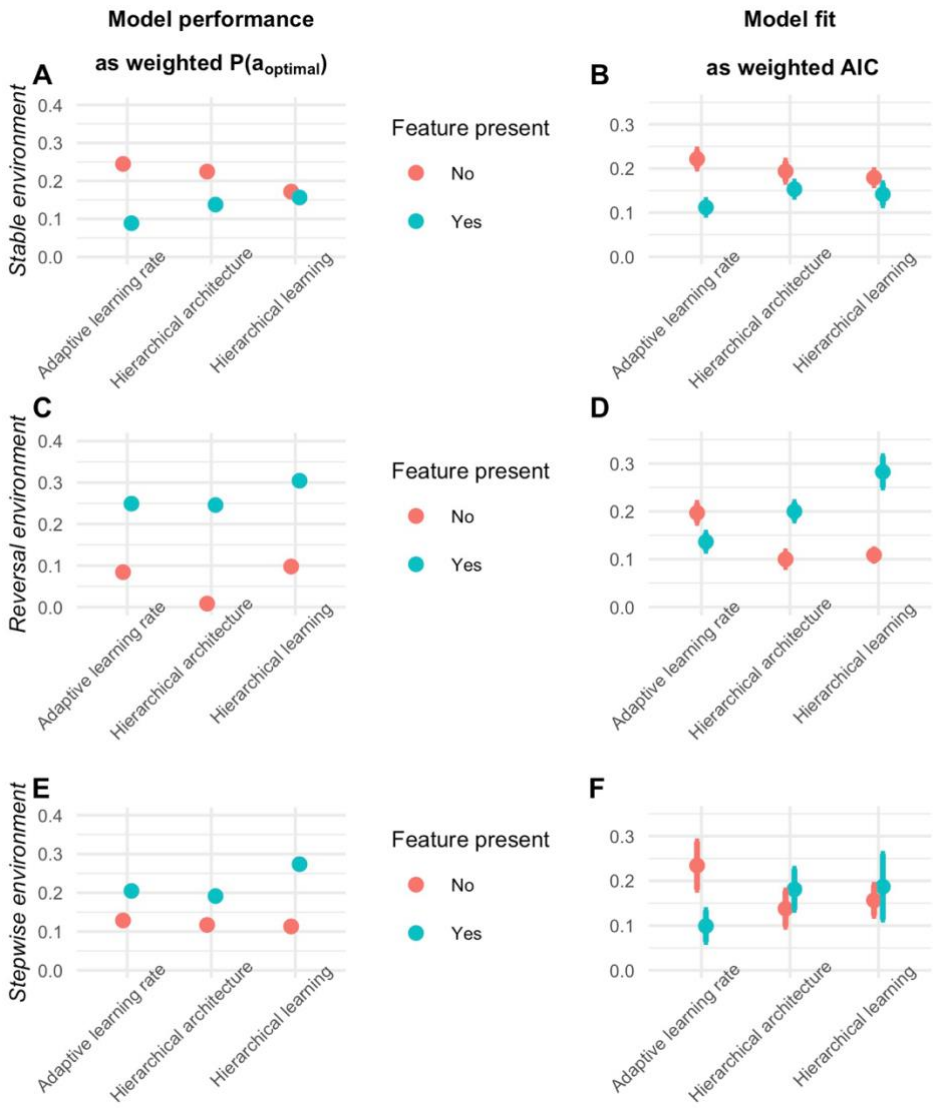

Figure 4. Hierarchical model extensions in each environment. *Each datapoint reflects the mean (weighted accuracy or model evidence) over all models with one particular feature (e.g., the ALR, Hierarchical_ALR and Full model for the adaptive learning feature) in a particular environment.* Column 1(A, C, E) illustrates the influence of each extension on model performance (as weighted $P(a_{optimal})$ on y-axis) across environments. Column 2(B, D, F) illustrates influence of each extension on model fit (as weighted AIC on y-axis) for the behavioral datasets in each environment.

To investigate the added value of each individual hierarchical extension in terms of model fit, a linear mixed-effects model was fitted with the weighted AIC for each subject in each dataset as dependent variable and environment (Stable, Reversal or Stepwise), adaptive learning rate (yes or no), hierarchical architecture (yes or no) and hierarchical learning (yes or no) as fixed independent variables. Dataset was included in the analyses as a random factor (intercept). The right column of Fig. 4 (panels B, D and F) show the interactions between environment and hierarchical extensions in terms of wAIC. Two of these interactions reached statistical significance. Although the models without adaptive learning rate (mean *wAIC* = .21) fitted better ($\chi^2(1, N = 410) = 86.376$, $p < .0001$) on average than the models with adaptive learning rate (mean *wAIC* = .123) across all three environments, the interaction between adaptive learning rate and environment did reach significance ($\chi^2(2, N = 410) = 8.791$, $p = .012$). Also for hierarchical learning, the interaction with environment reached significance

($\chi^2(2, N = 410) = 64.145$, $p < .0001$). For hierarchical architecture this interaction effect with environment did not reach significance ($\chi^2(2, N = 410) = 4.973$, $p = .094$). When investigating the interactions in a bit more detail, a similar pattern emerges as for the model performance. In Stable environments (Fig. 4B), there is no advantage for hierarchical extensions in terms of model fit. Hence, evidence is much stronger for models that do not use an adaptive learning rate ($t(936) = -7.836$, $p < .001$), or hierarchical architecture ($t(678) = -2.699$, $p = .007$) or hierarchical learning ($t(639) = -2.442$, $p = .015$). For the Reversal environment (Fig. 4D), there is a clear advantage for using a hierarchical architecture ($t(1209) = 7.752$, $p < .001$) and hierarchical learning ($t(529) = 9.813$, $p < .001$). Although model performance predicted an advantage for adaptive learning rates as well (Fig. 4C), there was significant evidence against using adaptive learning rates for model fitting on the empirical datasets ($t(1241) = -4.204$, $p < .001$). In the Stepwise environment (Fig. 4F), there was again more evidence for models that do not use an adaptive learning rate ($t(181) = -4.139$, $p < .001$). For the other two features, results were inconclusive. There was no evidence for or against the use of a hierarchical architecture ($t(199) = 1.398$, $p = .164$). The same applies for the hierarchical learning ($t(101) = .742$, $p = .46$) extension.

In sum, humans adaptively select hierarchical extensions to the flat learning approach depending on the environment. Interestingly, they usually select the extensions that optimize performance in the current environment. Specifically, there was no evidence for hierarchical extensions of the Flat model in Stable datasets. In the Reversal group, however, we did find evidence for hierarchical architecture and hierarchical learning.

Discussion

Although flat learning via the RW rule has proven very valuable to describe human (and animal) learning (1, 31, 32), a wide range of previous work has argued that flat learning is insufficient to capture human learning in complex environments (3–5, 10, 15). Therefore, several hierarchical extensions to the flat learning approach have been proposed in a wide range of different environments and datasets (3, 6, 12–14, 24, 33). Crucially, an extensive evaluation of these hierarchical extensions over multiple reinforcement learning environments was lacking. Hence, it remained unclear what hierarchical extensions are needed to model human cognition and how this relates to variations in the environment. To remedy this, the current study compared the Flat model to five more hierarchically complex models in three environments of different complexity. Model simulations were performed to illustrate which model performed optimally in each environment. Then, we compared the fit of each model on the behavioral data of ten different datasets (across the three environments). This allowed to gain insights in what hierarchical extensions are beneficial for performance but also for fitting human behavioral data in different environments.

In general, human data fitted best with the model that also performed best in that environment. Interestingly, which model was optimal, varied over environments. Specifically, we found that Flat models are optimal (in performance and fit) for Stable environments. For Reversal environments there was strong evidence for using a hierarchical architecture and hierarchical learning. In the Stepwise environment, this evidence was more mixed. Thus, current work demonstrates that humans can adaptively select the optimal model for specific reinforcement learning environments.

This suggests that the brain might have different expert systems available, from which a higher level system decides which expert receives control over behavior for the current task, as is the case in Mixture of Experts models (34–36). This higher level system has been previously linked to human prefrontal cortex (37). Here, the prefrontal cortex functions as a gating layer in a neural network, selectively guiding activity via different expert layers of the network. Interestingly, previous work has illustrated that this approach of mixing gating and expert layers can be scaled up to large deep neural networks which can be trained end-to-end to efficiently solve multiple complex (language translation) problems (38). Nevertheless, it is important to note that the experts in classic mixture of expert networks typically only contain information on how to perform a specific task. In this sense, the experts as traditionally conceived, are more closely related to the different task sets in our hierarchical models. In this sense, one could consider a three-level hierarchy, with the task sets at level 1, the task set selection model at level 2, and the model-selecting system at level 3. However, how the human brain recruits learning strategies (i.e., models) in response to varying task demands is an open question. Importantly, the six models in the current study were incremental in the sense that they all had the same base (the Flat model) but added different parameters to that base model (see Fig. 1A). Hence, different models could easily be selected by turning different parameters on (setting value larger than zero) or off (setting value to zero). Future work should further investigate how human behavior relates to optimal model selection and parameter tuning.

The current study considered three hierarchical extensions to the Flat model. A first extension is to add a higher-order process or parameter that adaptively manipulates lower-level parameters. Such meta-learning of lower-level parameter values has received considerable interest in models of human cognition (39–41) and has been linked to human medial prefrontal cortex (41, 42), and more specifically dorsal anterior cingulate cortex (12, 39, 43). One implementation of such meta-learning, which was also implemented in current work, is an adaptive learning rate. This allows the agent to be robust to noise during stable periods (by decreasing the learning rate) and to be flexible when strong changes in stimulus-action-reward contingencies occur (by increasing the learning rate). Notably, there exist several different implementations for adaptive learning rates (3, 5, 13, 39, 41). Current work adopted the approach of (3) which provides simplicity in the sense that only one parameter is added to the classic RW learning rule and it can be easily applied to multiple environments. Nevertheless, future work should further investigate how different implementations of adaptive learning rate allow to better fit human data.

Although current work only found limited support for adaptive learning rates within a dataset, we do show in the Supplementary Materials that different learning rates are used for different datasets and environments. Additionally, we discussed before that different models are selected by adjusting the hierarchical parameters. Since weighted model fit to the human data varied over environments, there is evidence that the hierarchically higher parameters are adaptive as well. Thus, current study does not argue that humans do not adapt learning rates or other parameter values. However, rather than an online adaptation to random reversals in contingencies as we investigated in the current task, a clearer distinction (via e.g., contextual cues) might be needed between periods that require different learning rates (20, 44). Future work might further explore when and how humans adapt lower-level parameters.

A second extension that we consider is to implement a hierarchical architecture. Here, the agent learns different rule sets on the lower level and on a hierarchically higher level learns/decides when to switch between rule sets. Previous work has linked this process of switching between rule sets to midfrontal theta (4-8Hz) power (6). In general, hierarchical architectures have received increasing interest in recent years, both to model human cognition (10, 16, 17, 43, 45) and to advance artificial intelligence (18, 46, 47). Indeed, a hierarchical architecture has several computational advantages. As demonstrated in previous work (10), storing multiple rule sets allows to avoid catastrophic forgetting because a change in stimulus-action-reward contingencies does not require to overwrite old information but to just switch to another rule set. Furthermore, learning is much faster if it is possible to select from previously learned rule sets (16, 18). Hence, agents who store multiple rule sets are also better at generalizing old knowledge to novel tasks or contexts (17).

A third extension considered learning at the hierarchically higher level. Here, learning at the hierarchical higher level allowed to adaptively change the weight of negative feedback in the switch accumulation process. Specifically, without hierarchical learning, the model would always switch after a specific number of trials with negative feedback (e.g., after 2 consecutive trials with negative feedback). This is appropriate if variability in the feedback is consistent over time and known to the agent. However, in the current datasets, subjects were not instructed about the exact feedback reliability and hence needed to learn about the variability of the feedback. Furthermore, in the Stepwise environments, $P(Rew|a_{optimal})$ would also change during the time course of the task.

Interestingly, learning at the hierarchically higher level can also have multiple manifestations. For instance, instead of learning when to switch rule sets, one can also learn which rule set to select given a current task context (16, 45). This allows to select the appropriate rule set in each context. One popular framework (17, 48, 49) extends this approach and suggests that on top of grouping stimulus-action associations in rule sets, contexts can also be grouped in latent states. As a result, lower-level rule sets are mapped to latent states which represent a group of contexts. This allows for fast generalization since lower-level knowledge (stored in rule sets) can immediately be transferred to all contexts that belong to the same latent state. Hence, previous work has proposed three levels of abstractions from rule sets at the lowest level to contexts and to latent states at the highest level. It has been suggested that these levels of abstraction are organized along the rostral-caudal axis of human prefrontal cortex where latent states are represented at the most rostral parts of prefrontal cortex (50–54).

Strikingly, both the adaptive learning rate and hierarchical learning extension are especially relevant for noisy environments with varying levels of uncertainty. On the lower level of the rule sets, adaptive learning rates are employed to balance behavioral adaptation in response to two types of uncertainty; uncertainty about feedback validity when $P(Rew|a_{optimal}) < 100\%$ but also uncertainty about the identity of $a_{optimal}$ in environments where this can change (Reversal and Stepwise). Previous work has linked these two types of uncertainty to the neuromodulators norepinephrine and acetylcholine (21, 55). On the higher level, hierarchical learning is employed to adapt rule switching to the uncertainty about feedback validity. When this validity is low, the agent should be more conservative to switch between rule sets than when this validity is high.

Different combinations of the three hierarchical extensions resulted in six models (see Fig. 2A). All six models were evaluated on three environments. In Stable environments, the stimulus-action-reward contingencies did not change. Hence, the agent must learn one single rule set. For this Stable environment, the Flat model was optimal. It performed optimally in the simulations and showed the best fit to human behavioral data. Interestingly, $P(Rew|a_{optimal})$ varied from 70-100% across datasets. As described in the Introduction, high learning rates would be suboptimal in noisy environments since it would lead the agent to "chase the noise". In the Supplementary Materials, we show optimal parameter values and (scaled) estimated parameters for each dataset. Here, it is observed that indeed for more noisy datasets (lower $P(Rew|a_{optimal})$), lower learning rates are more optimal and better fit behavioral data. This suggests that Flat learning models experience significant problems if they are faced with noisy environments that require strong flexibility.

To test this, a second environment considered reversal learning tasks. Results demonstrated that the Flat model is not suited for Reversal environments. Here, the Full model performed optimally in simulations and fitted best with human behavioral data. Interestingly, this Full model contained all three hierarchical extensions. However, the regression results only indicated empirical evidence for a hierarchical architecture and hierarchical learning. Consistently, evidence for the other models containing an adaptive learning rate (ALR and Hierarchical_ALR model) was low across datasets (see Fig. 3B). This canceled out the evidence for an adaptive learning rate in the Full model. Of course, this raises the question why an adaptive learning rate is only beneficial if it is combined with hierarchical learning. What is crucially different between the simple ALR model and a model with both hierarchical learning and an adaptive learning rate is that the ALR model needs to adapt the learning rate in two directions. More specifically, it needs to decrease the learning rate in stable periods and increase the learning rate when reversals occur. For models with a hierarchical architecture, this increase of the learning rate on reversals is not necessary since it can just switch between rule sets on reversals. Hence, for the Full model, the learning rate should be high only when the model learns a certain rule set for the first time. Once it has learned the rule, it needs stability (robustness against noise) and hence the learning rate can decrease. If later a reversal occurs, the Full model does not have to increase the learning rate again but can just switch to another rule set.

Interestingly, in the Reversal environment there was also more evidence for models with both a hierarchical architecture and hierarchical learning (Hierarchical_learning and Full model) than for models with only a hierarchical architecture (Hierarchical and Hierarchical_ALR model). This suggests that switching between rule sets was more optimal when there was learning at the hierarchically higher level. As described before, learning on the hierarchically higher level is beneficial when reliability of feedback changes over time. At first sight, this seems particularly relevant for the Stepwise environment. However, feedback is also more variable when the appropriate lower-level rule set is not learned yet than when learning on the lower level has converged. Hence, while receiving negative feedback during the initial learning of a rule set probably just signals that the agent should learn this rule set better, negative feedback after learning of the rule set has a higher probability of signaling an actual reversal of stimulus-action-reward contingencies. Therefore, it is optimal to have learning on the hierarchically higher level in the Reversal environment as well.

Interestingly, the Full model did not fit best for all Reversal datasets (see Fig. 3B). In the Goris/Volatile dataset, there was clear evidence in favor of the Flat model. However, Fig. 2B shows that accuracy was also significantly lower for this dataset. Interestingly, for the Goris/Volatile dataset, reversals occurred on average after 18 trials with fastest reversals occurring after only 15 trials in this dataset. In other datasets there was considerably more time between two reversals (average reversals after at least 30 trials for tasks with $P(Rew \mid a_{optimal}) < 100\%$). Hence, the fast succession of reversals in the Goris/Volatile dataset may have caused significant problems for subjects (as evidenced by a low accuracy) and caused a return to basic learning strategies as described by the Flat model.

While in the Reversal environment the identity of the optimal action could regularly change, $P(Rew|a_{optimal})$ was stable across the entire task. Hence, we considered a third environment in which $P(Rew|a_{optimal})$ could change as well. We referred to this environment as the Stepwise environment. Model simulations indicated that also here the Full model would be optimal. Although evidence in the empirical datasets was more mixed, model fit (Fig. 3B) and feature regression (Fig. 4D) shows a small advantage for models with a hierarchical architecture relative to those without a hierarchical architecture. Nevertheless, our regression results illustrated that this was not significant.

In sum, while Stable learning environments are best approached with a Flat learning model, current work found support for hierarchical extensions in the Reversal environment. While Flat models need to constantly overwrite rule sets, hierarchically organized models can store multiple (stable) rule sets and on a hierarchical higher level flexibly decide when to switch between rule sets. This provides an important computational advantage in Reversal environments; an advantage that humans also seem to exploit. Interestingly, we found an additional improvement in fit and performance when the model was further extended with learning on the hierarchically higher level and an adaptive learning rate on lower level. The added value of these features can be explained by the fact that rule sets should still be learned at initial stages of the task. This means that the model should learn faster and be more conservative for rule switching during initial rule learning than once rule sets are learned. While model simulations predicted that a hierarchical architecture would also be beneficial in Stepwise environments, there was no significant evidence for this extension in the empirical datasets. Interestingly, for both datasets in the Stepwise environment, there were periods in which $P(Rew|a_{optimal}) = 50\%$. This is highly ambiguous since there is in fact no optimal action and it is unclear for the model which rule set should receive control over behavior. One option for modeling would be to include a third rule set; but this would add another layer of complication, as it would mean that the model should either randomly switch between three rule sets, or instead implement a decision mechanism allowing to infer to which rule set it should switch. Alternatively, future work might consider evaluating model evidence in Stepwise environments that do not have this $P(Rew|a_{optimal}) = 50\%$ condition.

Meta-analysis has proven to be a valuable tool to address the replication crisis (56–59). As we also observed across datasets in the Reversal and Stepwise environments, multiple single studies can often have varying, or even contradicting, results. Nevertheless, over many datasets, a clear pattern can arise which allows to make more reliable conclusions. Importantly, as in all meta-analyses, one should consider the publication bias (60). Because of this publication bias, possibly relevant datasets were overlooked if they were not published. However, a strength of the current study is that the datasets

that were used here were not previously used to test the same models and hypotheses as in the current analyses. Hence, the publication bias that might be present in the selection of datasets is not specific to our hypotheses.

As is demonstrated in current work, and previously also in the context of working memory (61), meta-analyses can be useful for model selection. A good model of human cognition must fit human data better than an alternative model. However, most work typically considers such a model comparison for one dataset and one cognitive problem (task). This is suboptimal because, as we have demonstrated, model evidence can differ across learning problems and even across multiple datasets investigating the same learning problem. Therefore, we carried out model comparison over multiple datasets and learning environments. This allows to gain a deeper insight in which computational features are relevant for fitting human behavior and how these features relate to different environments. Specifically, we demonstrated that in Stable learning environments, Flat learning via the RW rule is sufficient to fit human behavior. However, for more complex environments in which stimulus-action-reward contingencies can reverse or change in a stepwise manner, we find evidence for more complex models that implement hierarchical extensions of the Flat learning approach. Particularly a hierarchical model architecture which stores multiple rule sets at the lower level and decides at a higher level which rule set should receive control over behavior, is beneficial in these environments. We believe there is significant potential in extending our meta-analytic approach with different models and different environments.

Materials and methods

The models

As illustrated in Fig. 1A, all models had the Flat model as base. Here, stimulus-action-reward contingencies are learned via the RW learning rule. Thus, the expected reward of one stimulus-action pair ($Q(s,a)$), is updated on each trial (t) by

$$Q(s,a)_{t+1} = Q(s,a)_t + \alpha\,(Rew_t - Q(s,a)_t) \qquad (1)$$

in which α is the learning rate and Rew_t represents the reward received on trial t.

For action selection, a SoftMax decision rule is used,

$$P(a) = \frac{\exp\left(\frac{Q(s,a)}{\tau}\right)}{\sum_{a=1}^{2}\exp\left(\frac{Q(s,a)}{\tau}\right)} \qquad (2)$$

Here, τ is a temperature parameter which controls the degree of exploration. With higher values of τ, there is a higher probability of selecting an action that does not have the highest expected reward.

One possible hierarchical extension of the flat learning approach is to implement an adaptive lower-level parameter. Here, we consider the learning rate. Importantly, to deal with periods of stability and plasticity in the environment, the learning rate should be able to both decrease and increase. For simplicity, current work also aimed to limit the number of additional parameters and have an approach that was easily applicable on multiple datasets with or without sudden changes in volatility and that could easily be integrated with the other hierarchical extensions. Therefore, we adopt the approach of (3). Here,

$$\alpha_{t+1} = (1-\sigma)\,\alpha_t + \sigma|(Rew_t - Q(s,a)_t)| \qquad (3)$$

in which we call σ the hybrid parameter. This hybrid parameter allows to continually update the learning rate parameter in both directions. Specifically, Equation (3) implies that the learning rate increases on trials with strong prediction errors ($Rew_t - Q(s, a)_t$) and decreases with a factor σ otherwise.

The second extension to flat learning considers a hierarchical architecture which stores multiple rule sets (*r*). Therefore, the expected reward of stimulus-action pairs are learned for each rule set separately (*Q*(*s, a, r*) instead of *Q*(*s,a*)). Which rule set initially has control of behavior is determined at random on the first trial. In line with previous work (6, 10), an accumulation-to-bound process determines when the model switches to the other rule set. Here, Switch evidence S_t accumulates via

$$S_{t+1} = (1-\gamma) * S_t - \gamma * min(0, Rew_t - V(r)) \qquad (4)$$

in which *V*(*r*) is the expected reward associated with rule set *r*. Here, we use two-arm bandit tasks with two rule sets. Hence, we initialized the expected reward at .5. On every trial when the reward is smaller than *V*(*r*) (that is, on unrewarded trials), the accumulated switch evidence in *S* increases and decreases otherwise. Here, the reward was either 1 (reward) or 0 (no reward) and the γ parameter had bounds [0, 1]. If there is no hierarchical learning, *V*(*r*) does not change and hence is fixed at .5. Thus, in absence of hierarchical learning, the asymptote of Equation (4) = .5. Therefore, we implemented that the agent switches to another rule set if the switch evidence S_t reaches a threshold value of .49.

Interestingly, the model can adaptively switch with a fixed *V(r)*. Nevertheless, in some situations, adding a learning rule for *V*(*r*) might be beneficial. For instance, when uncertainty about the rule set decreases (as sometimes happens in Stepwise environments), it might be adaptive to increase the weight of negative feedback (in Equation (4)) since it is more likely that this negative feedback was the consequence of a change in the contingencies than of noise in the environment. For this purpose, we can add RW learning at the hierarchically higher level as well via

$$V(m)_{t+1} = V(m)_t + \alpha_{high}\,(Rew_t - V(m)_t) \qquad (5)$$

where α_{high} is the hierarchically higher learning rate (see also (6, 10).

As illustrated in Fig. 1A, six possible models can be constructed based on the three extensions. Notice that hierarchical learning builds on the hierarchical architecture extension. Hence, the two combinations with hierarchical learning but no hierarchical architecture are impossible, leaving just $2^3 - 2 = 6$ possible models. In the Supplementary Materials, we demonstrate a good parameter and model recovery for all models.

The environments

As described in the Introduction, the current work considers three reinforcement learning environments. All environments can be described as two-armed bandit tasks but differed in the nature of the reinforcement dynamics (see Fig. 1B). We collected existing datasets for each environment via osf.io, or by contacting the corresponding authors. Note that the probability to obtain reward given an optimal action (P(*Rew* I $a_{optimal}$)) differed across datasets and simulations. In the datasets we consider, it is also the case that P(*Rew* I $a_{suboptimal}$) = 1 - P(*Rew* I $a_{optimal}$). An overview of the datasets is

given in Fig. 1B and a detailed description of each dataset and environment is provided in the Supplementary materials. Additionally, we performed simulations for each model in each environment (also under different P(*Rew* I $a_{optimal}$)) to evaluate the computational benefit of each extension.

Analyses

Model simulations and fitting

Parameter recovery and model recovery simulations were initially carried out for each model. Results for these analyses are described in the Supplementary Materials. We next aimed to investigate which model (see Fig. 1A) was able (under optimized parameter settings) to perform with the highest accuracy in each environment. For this purpose, we constructed 300 trials of each environment. For the Stable and Reversal environments, we did this twice, once with P(*Rew* I $a_{optimal}$) = 70% and once with P(*Rew* I $a_{optimal}$) = 100%. In the Stepwise environment, we used three levels of P(*Rew* I $a_{optimal}$) = [90, 70, 50] which alternated randomly during the 300 trials. Reversals or stepwise changes always occurred after 30 trials. We used these designs to estimate the parameters that optimized reward accumulation in each environment for each model. For optimization, the differential evolution method was used as implemented by the SciPy (version 1.4.1) package in Python (version 3.7.6). The optimal parameters for each environment and each model are summarized in the Supplementary Materials. The optimal parameters were used to perform 50 simulations of each model in each environment. Here, accuracy was computed as the proportion with which the optimal action was selected (P($a_{optimal}$)). In analogy to our approach for model fitting (described below), the accuracy was averaged over simulations and transformed to a weighted measure of accuracy. Specifically,

$$weighted\ \mathrm{P}\left(a_{optimal}\right) = \frac{\exp\left(-\frac{1}{2}\times\Delta P\left(a_{optimal}\right)_M\right)}{\sum_{M=1}^{M=6}\exp\left(-\frac{1}{2}\times\Delta P\left(a_{optimal}\right)_M\right)} \qquad (6)$$

in which,

$$\Delta P(a_{optimal})_M = \max\left(P\left(a_{optimal}\right)\right) - P\left(a_{optimal}\right)_M \qquad (7)$$

Hence, the weighted measure represents how strongly the accuracy of model *M* deviated from the most optimal (max) model. Here, more optimal models have a higher weighted accuracy and the weighted accuracy across all models always sums to 1.

After obtaining the optimal model, we next aimed to investigate which model fitted best with human behavioral data. For this purpose, parameter estimation was performed for each model of Fig. 1A on each dataset in Fig. 2A. Parameter estimation was implemented by minimizing the negative loglikelihood (LL) via the differential evolution method using the SciPy (version 1.4.1) package in Python (version 3.7.6).

Again, we used weighted measures for easier comparison across models. Here,

$$wAIC = \frac{\exp\left(-\frac{1}{2}\times\Delta AIC_M\right)}{\sum_{M=1}^{M=6}\exp\left(-\frac{1}{2}\times\Delta AIC_M\right)} \qquad (8)$$

for which,

$$\Delta AIC_M = AIC_M - \min(AIC) \quad (9)$$

In which AIC is the Akaike information criterion (AIC). This AIC measure implements a small penalty for the number of parameters to the negative loglikelihood. Again, the weighted measure represents how strongly the fit of model *M* deviated from the most optimal (min) model. Here, models that fit better have a higher wAIC and the wAIC across all models always sums to 1.

All six models were simulated 50 times on each environment. Based on the simulations, we extracted the weighted average accuracy for each model in each environment. Additionally, each model was fitted for each individual subject in each dataset. Based on results of the fitting procedure, weighted AIC values were computed for each subject in each dataset.

Linear mixed models

To further evaluate the added value of each separate hierarchical extension, we performed a linear mixed model regression with weighted AIC as dependent variable and reward schedule (stable, reversal or stepwise), adaptive learning rate (yes or no), hierarchical architecture (yes or no) and Hierarchical learning (yes or no) as fixed independent variables. Dataset was included in the analyses as a random independent factor.

References

1. R. A. Rescorla, A. R. Wagner, A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Class. Cond. II Curr. Res. Theory* **21**, 64–99 (1972).
2. R. Sutton, A. G. Barto, *Reinforcement learning: an introduction* (MIT Press, 1998).
3. Y. Bai, K. Katahira, H. Ohira, Dual learning processes underlying human decision-making in reversal learning tasks: Functional significance and evidence from the model fit to human behavior. *Front. Psychol.* **5**, 1–8 (2014).
4. M. Liu, W. Dong, S. Qin, T. Verguts, Q. Chen, Electrophysiological Signatures of Hierarchical Learning. *Cereb. Cortex N. Y. N 1991* **32**, 626–639 (2022).
5. J. T. McGuire, M. R. Nassar, J. I. Gold, J. W. Kable, Functionally Dissociable Influences on Learning Rate in a Dynamic Environment. *Neuron* **84**, 870–881 (2014).
6. P. Verbeke, *et al.*, Learning to synchronize: Midfrontal theta dynamics during rule switching. *J. Neurosci.* **41**, 1–13 (2021).
7. R. C. Wilson, M. R. Nassar, J. I. Gold, A Mixture of Delta-Rules Approximation to Bayesian Inference in Change-Point Problems. *PLoS Comput. Biol.* **9** (2013).
8. R. Cools, L. Clark, A. M. Owen, T. W. Robbins, Defining the neural mechanisms of probabilistic reversal learning using event-related functional magnetic resonance imaging. *J. Neurosci. Off. J. Soc. Neurosci.* **22**, 4563–4567 (2002).
9. A. Izquierdo, J. L. Brigman, A. K. Radke, P. H. Rudebeck, A. Holmes, The neural basis of reversal learning: An updated perspective. *Neuroscience* **345**, 12–26 (2017).
10. P. Verbeke, T. Verguts, Learning to synchronize: How biological agents can couple neural task modules for dealing with the stability-plasticity dilemma. *PLoS Comput. Biol.* **15** (2019).
11. M. R. Nassar, V. Troiani, The stability flexibility tradeoff and the dark side of detail. *Cogn. Affect. Behav. Neurosci.* **21**, 607–623 (2021).
12. T. E. J. Behrens, M. W. Woolrich, M. E. Walton, M. F. S. Rushworth, Learning the value of information in an uncertain world. *Nat. Neurosci.* **10**, 1214–1221 (2007).
13. C. Foucault, F. Meyniel, Gated recurrence enables simple and accurate sequence prediction in stochastic, changing, and structured environments. *eLife* **10**, e71801 (2021).
14. J. K. Kruschke, Bayesian approaches to associative learning: From passive to active learning. *Learn. Behav.* **36**, 210–226 (2008).
15. F. Bouchacourt, S. Tafazoli, M. Mattar, T. J. Buschman, N. D. Daw, Fast rule switching and slow rule updating in a perceptual categorization task. *eLife* **11**, e82531 (2022).

16. M. M. Botvinick, Y. Niv, A. C. Barto, Hierarchically organized behavior and its neural foundations : A reinforcement learning perspective. *Cognition* **113**, 262–280 (2009).
17. A. G. E. Collins, M. J. Frank, Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychol. Rev.* **120**, 190–229 (2013).
18. T. G. Dietterich, Hierarchical reinforcement learning with the MAXQ value function decomposition. *J. Artif. Intell. Res.* **13**, 227–303 (2000).
19. M. J. Frank, D. Badre, Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: Computational analysis. *Cereb. Cortex* **22**, 509–526 (2012).
20. L. Q. Yu, R. C. Wilson, M. R. Nassar, Adaptive learning is structure learning in time. *Neurosci. Biobehav. Rev.* **128**, 270–281 (2021).
21. A. J. Yu, P. Dayan, Uncertainty, neuromodulation, and attention. *Neuron* **46**, 681–92 (2005).
22. R. Moran, P. Dayan, R. J. Dolan, Human subjects exploit a cognitive map for credit assignment. *Proc. Natl. Acad. Sci. U. S. A.* **118**, 1–12 (2021).
23. M. P. Noonan, *et al.*, Separate value comparison and learning mechanisms in macaque medial and lateral orbitofrontal cortex. *Proc. Natl. Acad. Sci.* **107**, 20547–20552 (2010).
24. M. Silvetti, R. Seurinck, T. Verguts, Value and Prediction Error in Medial Frontal Cortex: Integrating the Single-Unit and Systems Levels of Analysis. *Front. Hum. Neurosci.* **5**, 1–15 (2011).
25. J. Goris, *et al.*, Autistic traits are related to worse performance in a volatile reward learning task despite adaptive learning rates. *Autism* **25**, 440–451 (2021).
26. P. Huycke, P. Verbeke, C. N. Boehler, T. Verguts, Theta and alpha power across fast and slow timescales in cognitive control. *Eur. J. Neurosci.* **54**, 4581–4594 (2021).
27. L. Xia, *et al.*, Modeling changes in probabilistic reinforcement learning during adolescence. *PLoS Comput. Biol.* **17**, 1–22 (2021).
28. D. Mukherjee, A. L. S. Filipowicz, K. Vo, T. D. Satterthwaite, J. W. Kable, Reward and Punishment Reversal-Learning in Major Depressive Disorder. *J. Abnorm. Psychol.* **129**, 810–823 (2020).
29. M. X. Cohen, C. E. Elger, C. Ranganath, Reward expectation modulates feedback-related negativity and EEG spectra. *NeuroImage* **35**, 968–978 (2007).
30. T. P. Hein, J. de Fockert, M. H. Ruiz, State anxiety biases estimates of uncertainty and impairs reward learning in volatile environments. *NeuroImage* **224**, 117424 (2021).
31. P. W. Glimcher, Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proc. Natl. Acad. Sci.* **108**, 15647–15654 (2011).
32. E. E. Steinberg, *et al.*, A causal link between prediction errors, dopamine neurons and learning. *Nat. Neurosci.* **16**, 966–973 (2013).
33. C. Mathys, J. Daunizeau, K. Friston, K. Stephan, A Bayesian Foundation for Individual Learning Under Uncertainty. *Front. Hum. Neurosci.* **5** (2011).
34. M. I. Jordan, R. A. Jacobs, Hierarchical Mixtures of Experts and the EM Algorithm. *Neural Comput.* **6**, 181–214 (1994).
35. J. P. O'Doherty, *et al.*, Why and how the brain weights contributions from a mixture of experts. *Neurosci. Biobehav. Rev.* **123**, 14–23 (2021).
36. Z. Chen, Y. Deng, Y. Wu, Q. Gu, Y. Li, Towards Understanding Mixture of Experts in Deep Learning (2022) (January 27, 2023).
37. B. Tsuda, K. M. Tye, H. T. Siegelmann, T. J. Sejnowski, A modeling framework for adaptive lifelong learning with transfer and savings through gating in the prefrontal cortex. *Proc. Natl. Acad. Sci.* **117**, 29872–29882 (2020).
38. N. Shazeer, *et al.*, Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer (2017) https:/doi.org/10.48550/arXiv.1701.06538 (January 27, 2023).
39. M. Silvetti, E. Vassena, E. Abrahamse, T. Verguts, Dorsal anterior cingulate-midbrain ensemble as a reinforcement meta-learner. *PLOS Comput. Biol.*, 1–32 (2017).
40. J. X. Wang, Meta-learning in natural and artificial intelligence. *Curr. Opin. Behav. Sci.* **38**, 90–95 (2021).
41. J. X. Wang, *et al.*, Prefrontal cortex as a meta-reinforcement learning system. *Nat. Neurosci.* **21**, 860–868 (2018).

42. M. Khamassi, P. Enel, P. F. Dominey, E. Procyk, “Chapter 22 - Medial prefrontal cortex and the adaptive regulation of reinforcement learning parameters” in *Progress in Brain Research*, Decision Making., V. S. C. Pammi, N. Srinivasan, Eds. (Elsevier, 2013), pp. 441–464.
43. C. B. Holroyd, T. Verguts, The best laid plans: Computational principles of ACC. *Trends Cogn. Sci.* **25**, 316–329 (2021).
44. J. Simoens, T. Verguts, S. Braem, Meta-learning environment-specific learning rates (2023) https:/doi.org/10.31234/osf.io/ypd45 (January 25, 2023).
45. C. B. Holroyd, S. M. McClure, Hierarchical control over effortful behavior by rodent medial frontal cortex: A computational model. *Psychol. Rev.* **122**, 54–83 (2015).
46. O. Çatal, T. Verbelen, T. Van de Maele, B. Dhoedt, A. Safron, Robot navigation as hierarchical active inference. *Neural Netw.* **142**, 192–204 (2021).
47. J. Chung, S. Ahn, Y. Bengio, Hierarchical Multiscale Recurrent Neural Networks (2017) https:/doi.org/10.48550/arXiv.1609.01704 (January 9, 2023).
48. S. J. Gershman, Y. Niv, Exploring a latent cause theory of classical conditioning. *Learn. Behav.* **40**, 255–268 (2012).
49. Y. Niv, Learning task-state representations. *Nat. Neurosci.* **22**, 1544–1553 (2019).
50. D. Badre, A. S. Kayser, M. D’Esposito, Frontal Cortex and the Discovery of Abstract Action Rules. *Neuron* **66**, 315–326 (2010).
51. D. Badre, D. E. Nee, Frontal Cortex and the Hierarchical Control of Behavior. *Trends Cogn. Sci.* **22**, 170–188 (2018).
52. E. D. Boorman, T. E. J. Behrens, M. W. Woolrich, M. F. S. Rushworth, 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).
53. E. Koechlin, C. Ody, F. Kouneiher, The architecture of cognitive control in the human prefrontal cortex. *Science* **302**, 1181–1185 (2003).
54. K. Nakahara, T. Hayashi, S. Konishi, Y. Miyashita, Functional MRI of Macaque Monkeys Performing a Cognitive Set-Shifting Task. *Science* **295**, 1532–1536 (2002).
55. G. Aston-Jones, J. D. Cohen, An Integrative Theory of Locus Coeruleus-Norepinephrine function: Adaptive Gain and Optimal Performance. *Annu. Rev. Neurosci.* **28**, 403–450 (2005).
56. E. C. Carter, F. D. Schönbrodt, W. M. Gervais, J. Hilgard, Correcting for Bias in Psychology: A Comparison of Meta-Analytic Methods. *Adv. Methods Pract. Psychol. Sci.* **2**, 115–144 (2019).
57. S. E. Maxwell, M. Y. Lau, G. S. Howard, Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? *Am. Psychol.* **70**, 487–498 (2015).
58. Open Science Collaboration, An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science. *Perspect. Psychol. Sci.* **7**, 657–660 (2012).
59. P. E. Shrout, J. L. Rodgers, Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis. *Annu. Rev. Psychol.* **69**, 487–510 (2018).
60. A. Thornton, P. Lee, Publication bias in meta-analysis: its causes and consequences. *J. Clin. Epidemiol.* **53**, 207–216 (2000).
61. R. van den Berg, E. Awh, W. J. Ma, Factorial Comparison of Working Memory Models. *Psychol. Rev.* **121**, 124–149 (2012).

Supplementary Materials

Model evaluation

We used simulations and parameter estimation via the differential evolution method implemented in the SciPy package in Python to evaluate the models. Specifically, we performed parameter and model recovery.

Simulations were informed by our previous work (1). Here, a subset (3/6) of models (see Table 1 in main text) were fitted on a reversal learning task (480 trials with 15 reversals). Based on this previous study (1), we determined plausible value distributions for each parameter in each model. From these distributions, 540 values for each parameter were drawn and used to simulate 540 datasets on the design of the reversal learning task. This resulted in 3240 (540 x 6 models) simulated datasets on the reversal learning task.

Parameter recovery

Traditionally, the appropriateness of a model to describe empirical data is evaluated by performing parameter recovery. In Fig. S1, we show the true parameter distributions that were used for simulations and the parameter distributions that were estimated based on this simulated data. As can be observed, these distributions are largely similar for most of the parameter values across all models. Notably, there is an important discrepancy between the true and estimated distributions for the learning rate in models with an adaptive learning rate (ALR, Hierarchical_ALR and Full model). Indeed, this value is much harder to estimate in these models than in the other models since the learning rate here only represents the starting value of the learning rate on the first trial. After this first trial, the learning rate is changing with a rate determined by the Hybrid parameter. The rate at which the learning rate is changing (i.e., the Hybrid parameter) is stable over trials and is thus easier to estimate. Importantly, this also means that the learning rate in these models is of lower importance to model behavior than in other models.

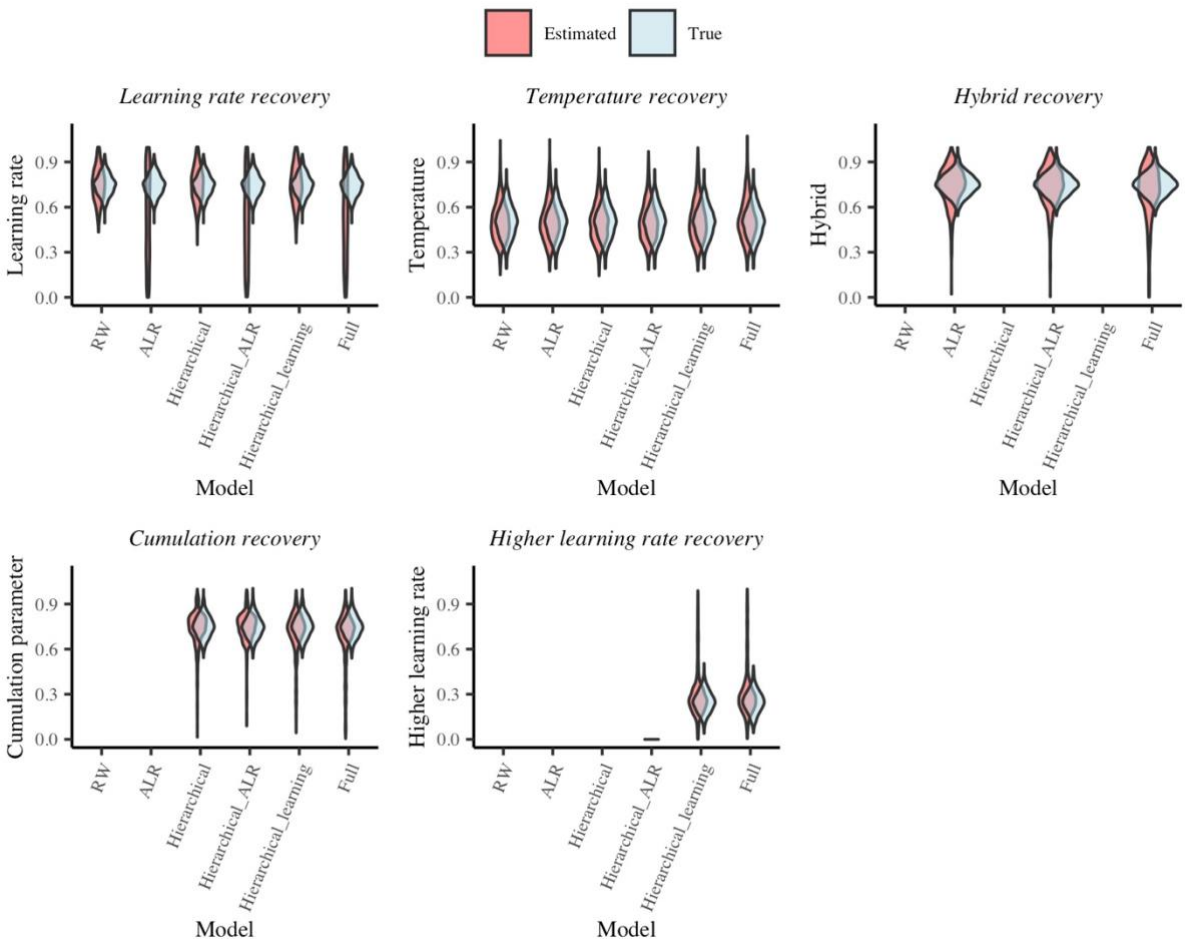

Figure S1.**True and estimated distributions for parameter values in parameter recovery analyses**. If there is no distribution in the plot it is because this parameter is absent in that particular model.

The Pearson correlation coefficient between the true and estimated parameters is illustrated in Fig. S2A. Here, white cells represent parameters that are not present in that model. As expected from Fig. S1, it can be observed that the correlation for the learning rate is indeed very low for models with an adaptive learning rate. For other parameters, the correlations are reasonably high. Apart from the learning rate in models with an adaptive learning rate, the lowest observed correlation value is .293 which is still highly significant ($p < .0001$).

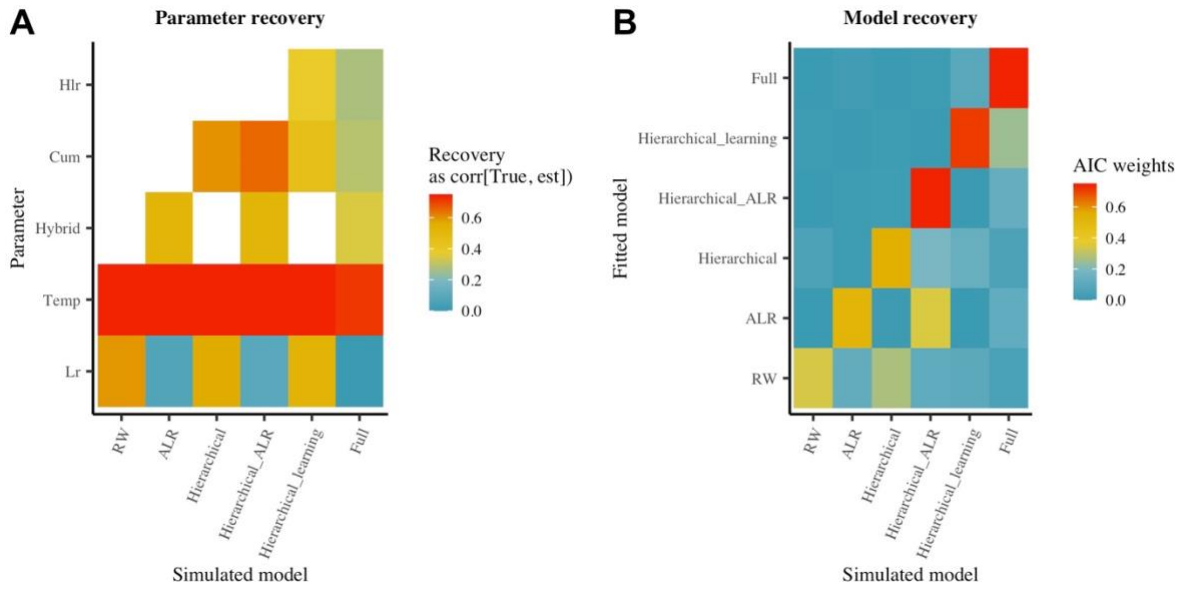

Figure S2. Recovery results. A shows parameter recovery as the Pearson correlation between the true and estimated parameters. Here, White cells indicate parameters that are absent for that particular model. B shows model recovery results.

Model recovery

Importantly, the main goal of the current study is to identify what computational strategy is most likely to be underlying a specific behavioral dataset. To validate this, we performed model recovery analyses. As shown in Fig. S2B, we can indeed clearly distinguish which model was responsible for a particular dataset. The more parameters in the model, the more distinct behavioral dataset it produces and hence the stronger the evidence for a model is for data that is simulated with that model.

Parameter results

We present parameter values for all models in all environments and/or datasets. Fig. S3 provides parameter values as optimized for performance (in terms of accumulated reward) in each environment. Fig. S4 provides (scaled) parameter values as optimized in terms of model fit to empirical data from the datasets. Values are scaled to allow for easier comparison.

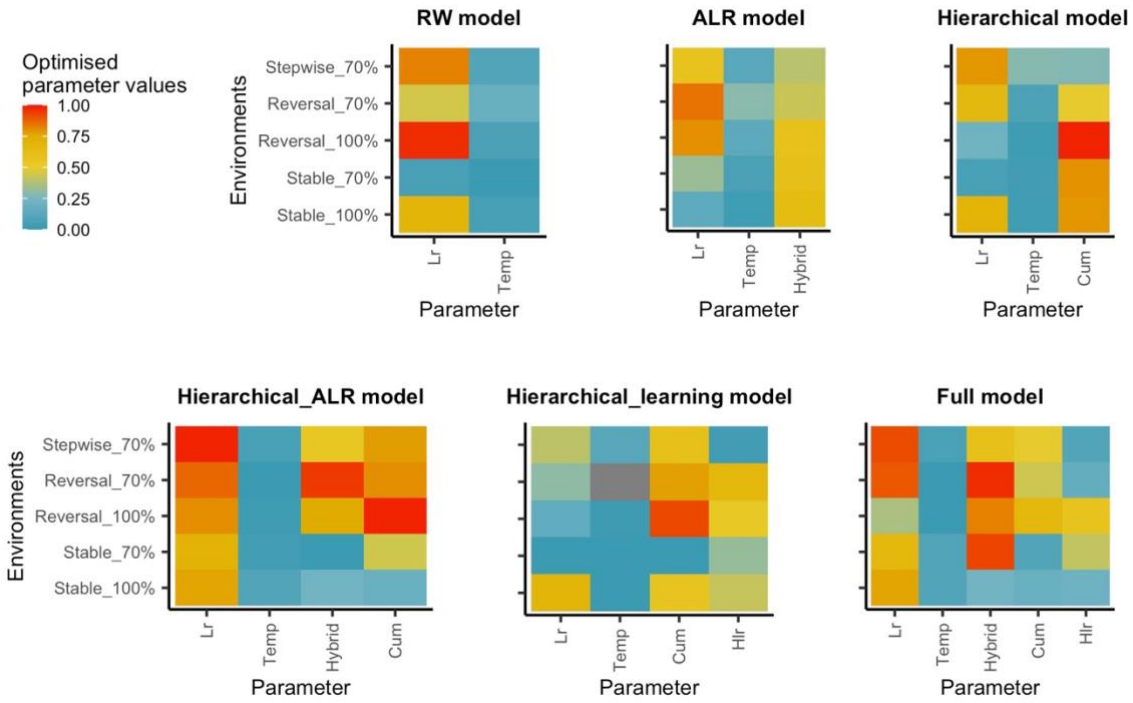

Figure S3. Optimized parameter values for each model in each environment. Parameters were optimized based on accumulated reward across 300 trials of each environment.

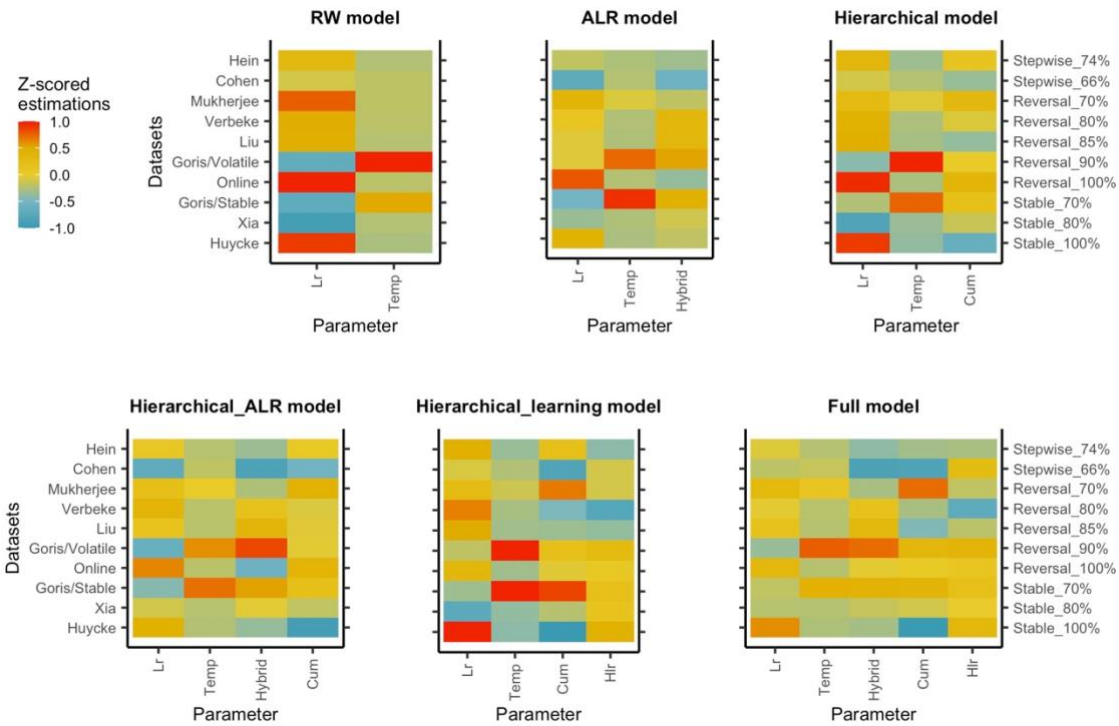

Figure S4. Scaled estimated parameter values for each model in each dataset. Parameters were estimated by minimizing the negative log likelihood of response via differential evolution. Parameter values were Z-scored for easier comparison across datasets.

Detailed description of the environments and datasets

The environments

In the Stable environment, stimulus-action-reward contingencies remain stable over time. Hence, the identity of $a_{optimal}$ as well as P(*Rew* | $a_{optimal}$) remains stable over time (over trials). Notably, P(*Rew* I $a_{optimal}$) could vary (ranging from 70 - 100%) across the three datasets within the Stable environment (see also Fig. 2A). Simulations were performed with the most extreme P(*Rew* I $a_{optimal}$). Hence, we performed simulations on the Stable environment with P(*Rew* I $a_{optimal}$) = 70% and with P(*Rew* I $a_{optimal}$) = 100%.

In the Reversal environment, stable periods alternate with sudden reversals of the stimulus-action-reward contingencies. Hence, at several time points during the task, the identity of the optimal action could change. Nevertheless, also here, P(*Rew* I $a_{optimal}$) remains stable over time within one dataset. Across the five datasets (see Fig. 2A), P(*Rew* I $a_{optimal}$) could vary (ranging from 70 – 100%). Notice that the deterministic version (P(*Rew* I $a_{optimal}$) = 100%) of the reversal learning task is a special case since there is no noise in the feedback. As a result, flat learning will experience less problems since a constant high learning rate cannot lead to chasing the noise when noise is absent. There is no stability problem. To control for this special case, we considered it necessary to include such a dataset. Nevertheless, we did not find a dataset that was available and suited for comparison (in terms of design, subjects, …) with the other datasets. For this purpose, an additional (not previously published) dataset was collected in the Reversal environment with P(*Rew* I $a_{optimal}$) = 100%. Details are described below. As in the Stable environment, also simulations for each model were performed in the Reversal environment with P(*Rew* I $a_{optimal}$) = 70% and P(*Rew* I $a_{optimal}$) = 100%.

In the Stepwise environment, the identity of the optimal action also changed over time. However, in these datasets, also P(*Rew* I $a_{optimal}$) changes over trials. An example of the reinforcement dynamics in each environment is shown in Fig. 1B. In the two Stepwise datasets (see Fig. 2A), P(*Rew* I $a_{optimal}$) varied from 50 to 90%. The same range was used for Simulations.

In total, ten different datasets were analyzed. Nine of these datasets are already described in previous publications (2–9). We collected one additional dataset via prolific (www.prolific.co) for reasons that we described above. For the specifics of the datasets, we direct readers to the original manuscript. Furthermore, a summary of all datasets is presented in Fig. 2A, and a short description of each study is provided below.

The datasets

Huycke et al., 2021 (Stable 100% reward).

In this dataset (5), subjects had to learn mappings between abstract stimuli and left- or right responses. These mappings remained stable across the entire task. Once the subject gave a response, feedback was presented (correct / incorrect). Here, p(*Rew* I $a_{optimal}$) = 100%. In total, there were 36 unique stimuli which were presented in groups of 4 stimuli per block. While 32 stimuli were only repeated 8 times (in one block), the other 4 stimuli were repeated 64 times (8 times across 8 blocks). Subjects did not receive monetary reward based on performance. All six models were fitted for all subjects (*N* = 24) across all trials (*Trials* = 512).

Xia et al., 2021 (Stable 80% reward).

In this dataset (9), subjects had to learn what butterfly preferred a particular flower. These preferences remained stable across the entire task. There were two flowers which were repeated every trial. The locations of the flowers (left or right) were randomized over trials. On each trial, one of four possible butterflies was presented, and the subject had to direct the butterfly to the left- or right flower with a video game controller. Each butterfly was repeated 30 times across 4 blocks. After a response, feedback was presented (win / lose). Here, p(*Rew* I $a_{optimal}$) = 80%. Subjects did not receive monetary reward based on performance. The original study was interested in age-related differences in reinforcement learning. Therefore, there was a strong variability in age. For comparison with other datasets, we fitted the models on the group of subjects that reached adulthood (age >= 18 years). Additionally, we excluded 3 subjects who scored below chance level. As a result, all six models were fitted on *N* = 104 subjects across all trials (*Trials* = 120).

Goris et al., 2021 (Stable 70% reward).

In this dataset (3), subjects were presented on each trial with a set of 2 pictures and had to use a left or right button press to select respectively the left or right picture. Two sets of pictures were randomly presented across trials. After response, feedback was presented in the form of a picture of money that was crossed when feedback was negative. Subjects received small monetary reward based on performance. The full task consisted of three different conditions. Here, we selected what was called the stationary low noise condition. To avoid carry-over from previous conditions, we only analyzed data of subjects that started with our condition of interest. In this condition, there was one picture for each set that elicited the highest reward across all trials and p(*Rew* I $a_{optimal}$) = 70%. The original study was interested in the influence of autistic traits for reinforcement learning. For comparison with other datasets, we excluded 4 subjects who scored (sub-)clinical on the autism scales

used in the original study as well as 6 subjects who scored below chance level. As a result, all six models were fitted on *N* = 38 subjects for all trials (*Trials* = 90).

Online (Reversal 100% reward).

We collected an additional dataset to obtain the same range for P(*Rew* I $a_{optimal}$) across studies in the Reversal group as in the Stable group. Jspsych-6 (10) was used to program the task and prolific (www.prolific.co) was used to recruit 50 subjects. One subject was excluded due to chance level performance (*N* = 49). Since this dataset was not published before, we provide a more elaborate description.

As in previous work (9, 11), subjects had to direct a butterfly to its preferred flower. Each trial started with the presentation of a fixation cross for one second. Then, one (out of two) butterfly was presented together with two flowers. Butterflies differed in color. These stimuli were presented for 5 seconds or until response. On each trial, the same two flowers were presented. The locations of the flowers (left or right) were randomized over trials. Subjects had to select the left or right flower by respectively pressing the f- (left) or j- (right) key on their keyboard. Following response, the subjects received feedback for 1 second in the form of +10 or +0 points.

Since this dataset belongs to the Reversal environment, the preference of the butterflies would change during the experiment. Specifically, over the time course of 220 trials, there were 10 reversals. Here, rule reversals were drawn from the uniform distribution with limits 15 and 25. Notice that here P(*Rew* I $a_{optimal}$) = 100%. Thus, subjects always received reliable feedback and could hence very quickly detect reversals.

Subjects were informed that each butterfly preferred one flower but that this preference could change during the experiment. After a practice phase of 40 trials with a red and green butterfly and one rule reversal, subjects were presented a two-question survey to check whether they performed and understood the task correctly. Specifically, subjects were asked (1) which butterfly preferred the white flower at the end (green or red) and they were asked (2) how many times the preference of the green butterfly changed (0, 1 or 5 times). If both questions were answered correctly, subjects could continue to the experimental phase of 220 trials with 10 rule reversals. Here, a yellow and blue butterfly (but the same flowers) were used to eliminate transfer of learning from the practice phase. The mean completion time of the study was 17 minutes for which subjects received £4 and an additional reward based on performance ranging from £1.5 to £2.5.

Goris et al., 2021 (Reversal 90% reward).

As mentioned before, the Goris et al., (2021) dataset contained three different conditions. One of these conditions was the volatile condition. Here, the same two sets of pictures were used but the optimal picture reversed every 18 trials. Additionally, p(*Rew* I $a_{optimal}$) was increased to 90%. Subjects received small monetary reward based on performance. Again, we only analyzed data of subjects that started with our condition of interest. Additionally, we excluded 5 subjects that scored (sub-)clinical on the autism scales used in the original study and 7 subjects that scored below chance level. All six models were fitted on *N* = 49 subjects for all trials (*Trials* = 90).

Liu et al., 2022 (Reversal 85% reward).

In this dataset (6), subjects were presented on each trial with a picture of either an animal or a piece of clothing. After this picture, two gratings were presented, one with a horizontal orientation

and one with vertical orientation. The locations (left or right) of the gratings were randomized across trials. Subjects had to indicate with a left or right button press, which grating matched the initially presented picture. The grating that matched each picture was reversed after either 80 or 20 trials. During the short stable periods of 20 trials, p(*Rew* I $a_{optimal}$) = 90% while during the longer stable periods of 80 trials, p(*Rew* I $a_{optimal}$) = 80%. Although p(*Rew* I $a_{optimal}$) was not stable during the whole experiment, we believed changes were too small to categorize this dataset in the Stepwise group. Subjects did not receive monetary reward based on performance. Again, all six models were fitted for all *N* = 23 subjects on all trials (*Trials* = 320).

Verbeke et al., 2021 (Reversal 80% reward).

In this dataset (8), Subjects were presented either a horizontal or vertically oriented grating. Subjects had to learn which response, right (j-key) or left (f-key) matched with each grating. There were 15 reversals and the response that matched each grating was reversed every 20-40 trials. Here, p(*Rew* I $a_{optimal}$) = 80% across the entire task. Subjects received small monetary reward based on performance. All six models were fitted for all *N* = 30 subjects on all trials (*Trials* = 480).

Mukherjee et al., 2020 (Reversal 70% reward).

In this dataset (7), subjects were presented with two fractals, one on each side of the screen. The locations of the fractals were randomized across trials. Subjects chose one of the fractals by pressing a keyboard button. After the response, feedback was presented in the form of a coin (reward) or a red dot (absence of reward). Here, p(*Rew* I $a_{optimal}$) = 70% across the entire task. Subjects received small monetary reward based on performance. In the original study, there was a reward and a punishment condition. Additionally, two groups of subjects were used. One group with a major depressive disorder and a control group. For comparison with the other datasets, we only used data from the reward condition for the control group. Additionally, 8 subjects were excluded due to below chance level performance. As a result, all six models were fitted on the data of *N* = 56 subjects for (*Trials* = 90) trials.

Cohen et al., 2007 (Stepwise).

In this dataset, the subject was presented with two grey squares, one on each side of the screen. Subjects simply had to press either a right or left button on their keyboard. After the response, feedback was presented in the form of +10 or -10 cents. Every 80-150 trials, the probability of reward given a particular button press changed. Here, p(*Rew* I $a_{optimal}$) could be either 75% or 50%. Subjects received small monetary reward based on performance. Two subjects were removed from the analyses due to chance level accuracy. All six models were fitted for the remaining *N* = 15 subjects on all trials (*Trials* = 1200).

Hein et al., 2021 (Stepwise).

In this dataset, subjects were presented with two fractals, one on each side of the screen. The locations of the fractals were randomized across trials. By making a button press on the keyboard, subjects needed to select one of the fractals. After the response, subjects received feedback (win 5 pence or lose 5 pence). Every 26-38 trials the probability of reward given a particular fractal changed. Here, p(*Rew* I $a_{optimal}$) could be 90%, 70% or 50%. Subjects received small monetary reward based on performance. In the original study, subjects were split into two groups based on their score for an anxiety trait questionnaire. For comparison with the other datasets, we only used data from the

control (not anxious) group. As a result, all six models were fitted on the data of *N* = 20 subjects for *Trials* = 400.

References

1. P. Verbeke, *et al.*, Learning to synchronize: Midfrontal theta dynamics during rule switching. *J. Neurosci.* **41**, 1–13 (2021).
2. M. X. Cohen, C. E. Elger, C. Ranganath, Reward expectation modulates feedback-related negativity and EEG spectra. *NeuroImage* **35**, 968–978 (2007).
3. J. Goris, *et al.*, Autistic traits are related to worse performance in a volatile reward learning task despite adaptive learning rates. *Autism* **25**, 440–451 (2021).
4. T. P. Hein, J. de Fockert, M. H. Ruiz, State anxiety biases estimates of uncertainty and impairs reward learning in volatile environments. *NeuroImage* **224**, 117424 (2021).
5. P. Huycke, P. Verbeke, C. N. Boehler, T. Verguts, Theta and alpha power across fast and slow timescales in cognitive control. *Eur. J. Neurosci.* **54**, 4581–4594 (2021).
6. M. Liu, W. Dong, S. Qin, T. Verguts, Q. Chen, Electrophysiological Signatures of Hierarchical Learning. *Cereb. Cortex N. Y. N 1991* **32**, 626–639 (2022).
7. D. Mukherjee, A. L. S. Filipowicz, K. Vo, T. D. Satterthwaite, J. W. Kable, Reward and Punishment Reversal-Learning in Major Depressive Disorder. *J. Abnorm. Psychol.* **129**, 810–823 (2020).
8. P. Verbeke, *et al.*, Learning to synchronize: Midfrontal theta dynamics during rule switching. *J. Neurosci.* **41**, 1–13 (2021).
9. L. Xia, *et al.*, Modeling changes in probabilistic reinforcement learning during adolescence. *PLoS Comput. Biol.* **17**, 1–22 (2021).
10. J. R. de Leeuw, jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behav. Res. Methods* **47**, 1–12 (2015).
11. J. Y. Davidow, K. Foerde, A. Galván, D. Shohamy, An Upside to Reward Sensitivity: The Hippocampus Supports Enhanced Reinforcement Learning in Adolescence. *Neuron* **92**, 93–99 (2016).