Computational evidence for hierarchically-structured reinforcement learning in humans

Maria K Eckstein
maria.eckstein@berkeley.edu
& *AnneGECollins*
annecollins@berkeley.edu
Department of Psychology
UC Berkeley
Berkeley, CA 94720

August 10, 2019

Abstract

Humans have the fascinating ability to achieve goals in a complex and constantly changing world, still surpassing modern machine learning algorithms in terms of flexibility and learning speed. It is generally accepted that a crucial factor for this ability is the use of abstract, hierarchical representations, which employ structure in the environment to guide learning and decision making. Nevertheless, it is poorly understood how we create and use these hierarchical representations. This study presents evidence that human behavior can be characterized as hierarchical reinforcement learning (RL). We designed an experiment to test specific predictions of hierarchical RL using a series of subtasks in the realm of context-based learning, and observed several behavioral markers of hierarchical RL, for instance asymmetric switch costs between changes in higher-level versus lower-level features, faster learning in higher-valued compared to lower-valued contexts, and preference for higher-valued compared to lower-valued contexts, which replicated across three samples. We simulated a classic flat and a hierarchical RL model and compared the behaviors of both to humans. As expected, only hierarchical RL reproduced all human behaviors, whereas flat RL was qualitatively unable to show some, and provided worse model fits for the rest. Importantly, individual hierarchical RL simulations showed human-like behaviors in all subtasks simultaneously, providing a unifying account for a diverse range of behaviors. This work shows that hierarchical RL, a biologically-inspired and computationally cheap algorithm, can reproduce human behavior in complex, hierarchical environments, and opens the avenue for future research in this field.

Introduction

Research in the cognitive sciences has long highlighted the importance of hierarchical structure for intelligent behavior, in domains including perception (38), learning and decision making (11; 10), planning and problem solving (12), cognitive control (42), and creativity (17), among many others (56; 30). Specific hierarchical frameworks differ, but they all share the insight that hierarchical representations—i.e., the simultaneous representation of information at different levels of abstraction—allow humans to behave adaptively and flexibly in complex, high-dimensional, and ever-changing environments.

To illustrate, consider the following situation. Mornings in your office, your colleagues are working silently or quietly discussing work-related topics. After work, they are laughing and chatting loudly at their favorite bar. In this example, a context change induced a drastic change in behavior, despite the same interaction partners (i.e., "stimuli"). Hierarchical theories of cognition capture this behavior by positing strategies over strategies, i.e., that distinct behavioral strategies (or "task-sets") are activated in the appropriate context.

Although hierarchical representations can incur additional cognitive cost (14), they have several advantages compared to exhaustive "flat" representations: Once a task-set has been selected (e.g., office), attention can be focused on a subset of environmental features (e.g., just the interaction partner) (28; 45; 39; 62). When new contexts are encountered (e.g., new workplace, new bar), entire task-sets can be reused, allowing for generalization (17; 25; 54). Old skills are not catastrophically forgotten, but learning is continuous (27). Lastly, hierarchical representations deal elegantly with incomplete information, for example supporting action selection when context information is unavailable (17; 16).

Although we know that hierarchical representations are essential for flexible behavior, how we create these representations and how we learn to use them is still poorly understood. Here, we hypothesize that hierarchical reinforcement learning (RL), in which simple RL computations are combined to simultaneously operate at different levels of abstraction, can explain both aspects.

RL theory (52) formalized how to adjust behavior based on feedback in order to maximize rewards. Standard RL algorithms estimate how much reward to expect when selecting actions in response to stimuli, and use these estimates (called "action-values") to select actions. Old action-values are updated in proportion to the "reward prediction error", the discrepancy between action-values and received reward, to produce increasingly accurate estimates (methods). Such RL algorithms (which we call "flat RL" in contrast to hierarchical RL) converge to optimal behavior, are computationally cheap, and have lead to several recent break-throughs in the field of artificial intelligence (AI) (52). Nevertheless, there are also important shortcomings to flat RL. These include the curse of dimensionality, i.e., the exponential drop in learning speed with increasing numbers of states and/or actions; the lack of flexible behavioral changes; and oftentimes difficulties with generalizing, i.e., transferring old knowledge to new situations. Hierarchical RL (35) attempts to resolve these shortcomings by nesting RL processes at different levels of temporal (9; 43; 47) or state abstraction (39; 26).

There is broad evidence suggesting that the brain implements computations similar to RL algorithms, whereby the ventral tegmental area generates reward prediction errors (49; 8), and a wide-spread network of frontal cortical regions (37) and basal ganglia (1; 55) represents action values. Specific brain circuits thereby form "RL loops" (2; 16), in which learning is implemented through the continuous updating of action values (48; 45). Interestingly, this neural circuit is multiplexed, with distinct RL loops operating at different levels of abstraction along the rostro-caudal axis (28; 6; 2; 3; 31; 34; 4; 5; 7). This neural architecture is thus compatible with the implementation of hierarchical RL.

Recent research has started to support this possibility. In a temporal abstraction paradigm, reward prediction errors at two levels of abstraction were observed in human electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) (47; 46), as well as primate anterior cingulate firing patterns (13). Similarly, a state abstraction paradigm revealed neural value signals at two levels of abstraction in human BOLD (24). The goal of the current study is to explain—using a hierarchical RL model—what role these signals play in the learning process, to relate the isolated components of prior studies to each other, and expand the model to additional tasks including behavioral flexibility, generalization, and inference.

We first introduce our hierarchical RL model, and then assess how well it accounts for human behavior in a complex learning task by comparing humans to flat and hierarchical model predictions. Our results—replicated across three independent samples—support the hierarchical RL model qualitatively and quantitatively.

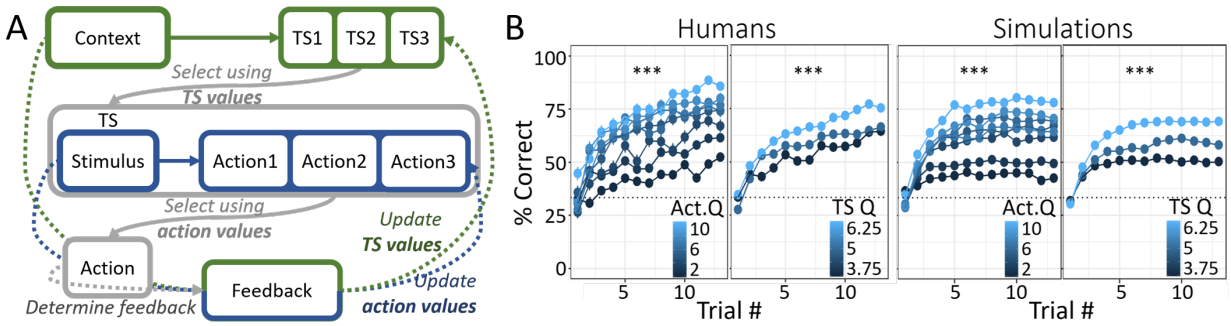

Figure 1: A) Schematic of the hierarchical RL model. A high-level RL loop (green) selects a task-set TS in response to the observed context, using learned TS values. A low-level RL loop (blue) selects an action in response to the observed stimulus, using the learned action-values of the chosen task-set. Task-set and action-values are updated based on feedback. B) Left: Human learning curves in the initial learning phase, averaged over blocks. Colors denote action-values (left) and task-set values (right). Stars denote effects of values on performance (main text). Right: Hierarchical RL simulation.

Results

Computational model

Our model is composed of two hierarchically-structured RL processes. The high-level process manages behavior at an abstract level by acquiring a policy over policies that learns which task-set to choose in each context. Within the selected task-set, the low-level process learns which action to choose in response to each stimulus (Fig. 1A). Task-sets initially select all actions with equal probability (random strategy), but learn to represent context-specific strategies through feedback-based updating of stimulus-action values (henceforth "action values"; lower-level loop). Similarly, task-sets themselves are initially selected at random, but get associated with specific contexts through feedback-based updating of context-task-set values (henceforth "task-set values"; high-level loop). In summary, the hierarchical RL model is based on two nested RL processes, which create an interplay between learning stimulus-action associations and context-task-set associations (suppl. Fig. S3).

Formally, to select an action a in response to stimulus s in context c, hierarchical RL goes through a two-step process: It first assesses the context and selects an appropriate task-set TS based on task-set values $Q(TS|c)$, using $p(TS|c) = \frac{exp(Q(TS|c))}{\sum_{TS_i} exp(\beta_{TS}\ Q(TS_i|c))}$, where the inverse temperature β_{TS} captures task-set choice stochasticity (Fig. 1A; for example behavior, see suppl. Fig. S3A). The chosen task-set TS provides a set of action-values $Q(a|s,TS)$ which are used to select an action a, according to $p(a|s,TS) = \frac{exp(Q(a|s,TS))}{\sum_{a_i} exp(\beta_a\ Q(a_i|s,TS))}$, where β_a captures action choice stochasticity (Fig. 1A; suppl. Fig. S3B). After executing action a on trial t, feedback r_t reflects the continuous amount of reward received, and is used to update the values of the selected task-set and action:

$$Q_{t+1}(TS|c) = Q_t(TS|c) + \alpha_{TS}\ (r_t - Q_t(TS|c))$$
$$Q_{t+1}(a|s,TS) = Q_t(a|s,TS) + \alpha_a\ (r_t - Q_t(a|s,TS))$$

where α_{TS} and α_a are the learning rates at the task-set and action level (Fig. 1A; suppl. Fig. S3C).

Task design

We designed a task in which participants learned to select the right action for each context and stimulus (Fig. 2A). We introduced differences in rewards between stimuli and differences in average rewards between contexts (Fig. 2B). Participants underwent an initial learning phase to acquire task-sets, and then completed several testing phases to test predictions of hierarchical RL (Fig. 2C and methods).

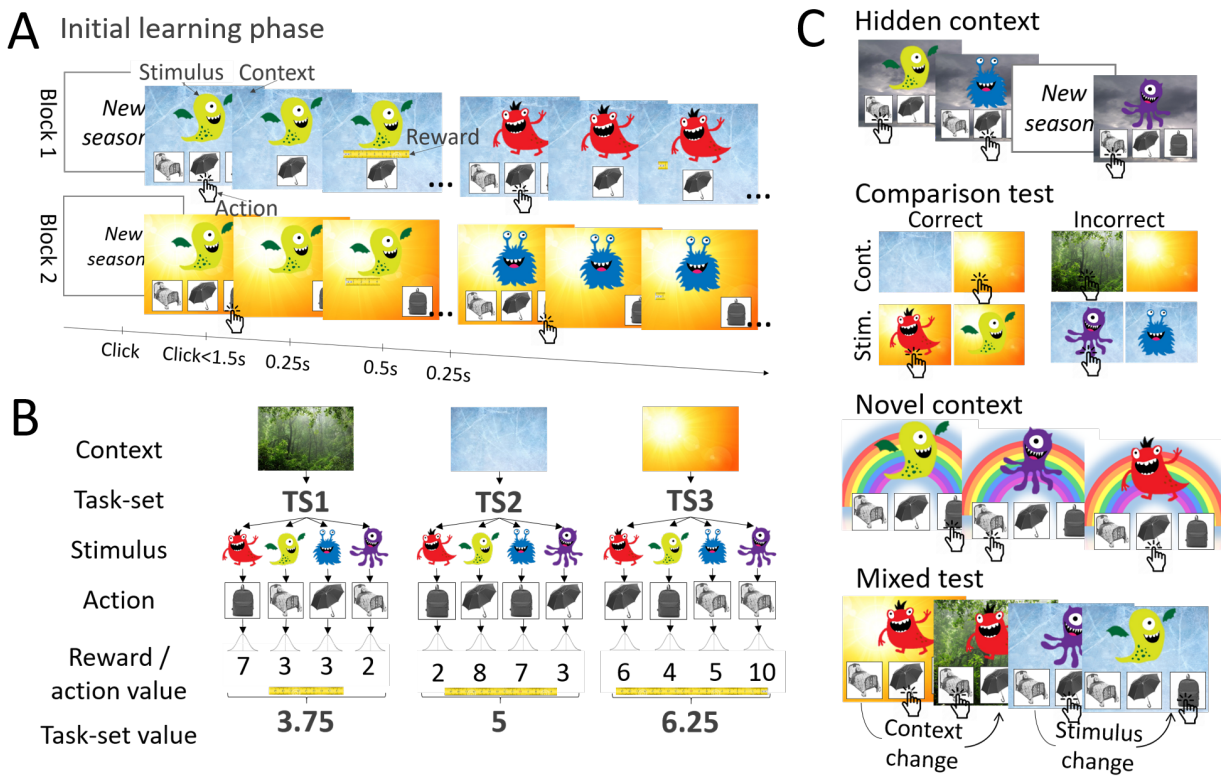

Figure 2: Task design. A) In the initial learning phase, participants saw one of three contexts (season) with one of four stimuli (alien) and had to find the correct action (item) through trial and error. Trials on the left are correct (large rewards), trials on the right are incorrect (small rewards). Contexts were presented blockwise. Feedback was deterministic, i.e., context-stimulus-action combinations always produced the same (mean$\pm$std) reward. B) Mappings between contexts, stimuli, and correct actions for each task-set TS. In each task-set, each stimulus was associated with one correct action, but the same action could be correct for multiple stimuli. Expected rewards differed between contexts, but were identical for all actions and stimuli. C) Test phases assessed behavioral markers of hierarchical RL. The hidden context phase was similar to the initial learning phase, except that contexts were unobservable (season hidden by clouds); participants knew that only previously-encountered contexts were possible. In the comparison test, at each trial, participants saw either two contexts (top row), or two stimuli in the same context (bottom row), and selected their preferred one. The novel context phase was similar to the initial learning phase, but introduced a novel context (rainbow) in extinction, i.e., without feedback. The mixed test was similar to the initial learning phase, but both stimuli and contexts were allowed change on each trial.

Learning curves and effects of reward

We first verified that participants were sensitive to the differences in rewards. Simple flat RL predicts better performance for larger rewards due to larger reward-prediction errors and faster learning. For the same reason, hierarchical RL additionally predicts better performance for higher-valued contexts due to larger reward-prediction errors and faster learning at the task-set level.

Using mean rewards for stimuli as a proxy for action-values, and mean rewards over contexts as a proxy for task-set values (Fig. 2B), we found evidence for both effects: participants performed better for stimuli with larger action values and in contexts with larger task-set values (Fig. 1B). To test for separable effects of both values, we ran a mixed-effects logistic regression model predicting trialwise accuracy from action-values, task-set values, and their interaction (fixed effects), specifying participants, trial, and block as random effects. Confirming our predictions, the model revealed unique effects of action-values, $\beta = 0.36(se = 0.05), z = 7.63, p < 0.001$, and task-set values, $\beta = 0.21(se = 0.05), z = 4.44, p < 0.001$. There was also a negative interaction, $\beta = -0.033(se = 0.01), z = -3.64, p < 0.001$, revealing smaller effects of action-values in higher-valued task-sets. This shows that participants were sensitive to differences in rewards, and provides initial evidence for hierarchical learning.

Acquisition of task-sets

We next tested whether participants created hierarchical representations, as predicted by the model. We conducted three analyses to highlight different aspects of hierarchy, assessing the asymmetry of switch costs, reactivation of task-sets, and task-set-based intrusion errors.

Asymmetric switch costs

Changes in task contingency at higher levels of abstraction (i.e., switching task-sets) are more costly than changes at lower levels (i.e., switching actions within task-sets) (44). Asymmetric switch costs can thus provide evidence for hierarchical representations (15). We assessed switch costs in the mixed phase of our task in which both contexts and stimuli were allowed to change on every trial (Fig. 2C). We compared participants' response times (RTs) between trials on which contexts switched (but same stimulus), compared to when stimuli changed (but same context). As expected, RTs were significantly slower for context switches than stimulus switches, $t(28) = 4.14$, $p < 0.001$. These asymmetric switch costs suggest that participants represented the task hierarchically, with contexts at the abstract level.

Reactivating task-sets

We next asked whether participants employed this hierarchical structure adaptively. Did they reactivate already-learned task-sets when appropriate (17; 25; 34)? We tested this in the hidden context phase in which the same three contexts were reiterated as before, but no context cues were provided, such that participants needed to infer the appropriate task-set (Fig. 2C).

Hierarchical RL makes a specific prediction about the first four trials after a (signaled) context switch in this test phase: despite the fact that each stimulus is only presented once, performance should increase because each feedback provides additional information about the correct task-set. In other words, hierarchical RL uses trial-by-trial feedback to infer the correct task-set. By contrast, flat RL treats each context-stimulus compound independently, such that feedback about one stimulus does not inform action selection for another, and performance should stay at the same level. Behavioral results supported hierarchical RL. Necessarily, participants' accuracy was at chance for the first stimulus after a context switch, but it improved steadily in the four subsequent trials, even though different stimuli were presented each time (Fig. 3A), evident in the correlation between accuracy and trial number (1-4), $r = 0.19, p = 0.038$. This shows that participants took advantage of the task structure and inferred missing information about the hidden context. This is evidence that participants formed and inferred task-sets.

We next tested whether our computational models could explain this behavior, using the following simulation approach (methods). We sampled the full parameter space of the flat and hierarchical RL models with 60,000 sets of random parameters. For each set of parameters, we simulated a population of $n = 310$ artificial agents (10 per participant), and conducted the same statistical analyses as on the human sample. We then compared the behavior of both models to humans and estimated model fits.

To investigate task-set reactivation, we used the slope in performance across the first four trials of a block as a summary statistic to compare simulations and humans (Fig. 3B). In hierarchical simulations, slopes were centered around

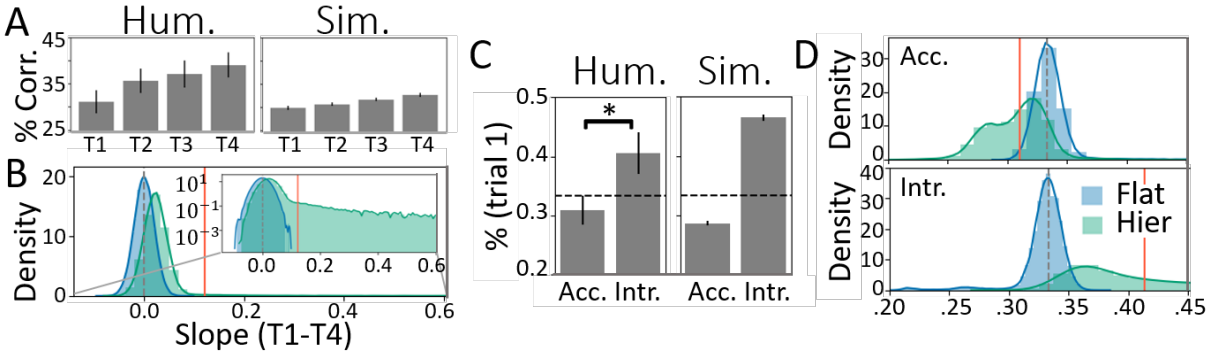

Figure 3: A)-B) Task-set reactivation in the hidden context phase. A) Left: Human performance increased across the first four trials after a context switch even though different stimuli were presented on each trial. This shows that feedback on one stimulus informed action selection for others. Right: Hierarchical RL simulation. B) Histogram of the performance slopes of 120,000 simulated populations based on hierarchical (green) and flat (blue) RL. Human slope from part A in red. Inset shows the histogram on a log scale to highlight long tails. C)-D) Intrusion errors in the initial learning phase. C) Left: Participants' accuracy in the current context (Acc.) and percent intrusion errors from the previous context (Intr.), based on first trials after a context switch. Participants more often erroneously employed the previous task-set than the current one. Star denotes significance in a repeated-measures t-test. Right: Hierarchical RL simulation. D) Accuracy and intrusion errors for hierarchical (Hier; green) and flat simulations (Flat; blue), with human behavior in red. Dotted lines indicates chance.

0.03 ($sd = 0.05$) and significantly positive, $t(59,999) = 136, p < 0.001$, suggesting significant task-set reactivation. By contrast, slopes in the flat model were centered around 0.00 ($sd = 0.02$), $t(59,999) = 0.38, p = 0.70$, reflecting the predicted lack of reactivation, with a significant difference between both, $t(59,999) = 140, p < 0.001$.

We next quantified the fit of flat and hierarchical RL to human behavior, using simulation-based Bayesian model comparison (methods). We first estimated the marginal likelihood of each model M, which corresponds to the probability $p(data|M)$ of human behavior under each model, from model simulations. Based on 60,000 simulations, the estimated likelihood of the flat model to produce human performance slopes was 0.00, whereas the likelihood of the hierarchical model was 0.024. The Bayes factor, i.e., the ratio between both, was infinite, showing that the hierarchical model accounted far better for human behavior than the flat. This supports our hypothesis that hierarchical RL can account for human task-set formation and reactivation.

Task-set based intrusion errors

We showed that participants used hierarchy adaptively to reactivate task-sets (positive transfer). However, hierarchy can also lead to negative transfer, e.g., intrusion errors, which we define as actions that are incorrect in the current context, but would have been correct in the previous context. Intrusion errors reflect incorrect task-set selection, but correct action selection within that task-set.

Intrusion errors should be most common in the initial learning phase when task-set values are least certain, i.e., immediately after context switches. Indeed, we found that participants' performance was at chance in the first trial after a context switch, and that intrusion errors were more common than correct actions, $t(28) = 2.22$, $p = 0.035$ (Fig. 3C). After the first trial, intrusion errors decreased consistently, like expected from a slowly-integrating RL process, evident in a negative effect of trial number on intrusion errors in a logistic regression predicting intrusion errors from trial index, task-set values, and action-values, controlling for block, and specifying random effects of participants, $\beta = -6.83\%$, $z = -9.31$, $p < 0.001$. In accordance with hierarchical RL, action-values, $\beta = -14.03\%$, $z = -8.45$, $p < 0.001$, and task-set values also had significant effects on intrusion errors, $\beta = -2.43\%$, $z = -1.00$, $p < 0.001$, showing that intrusion errors were more likely for stimuli and task-sets with larger values, in accordance with hierarchical RL.

We next assessed negative transfer in model simulations, focusing on accuracy and intrusion errors on the first trial

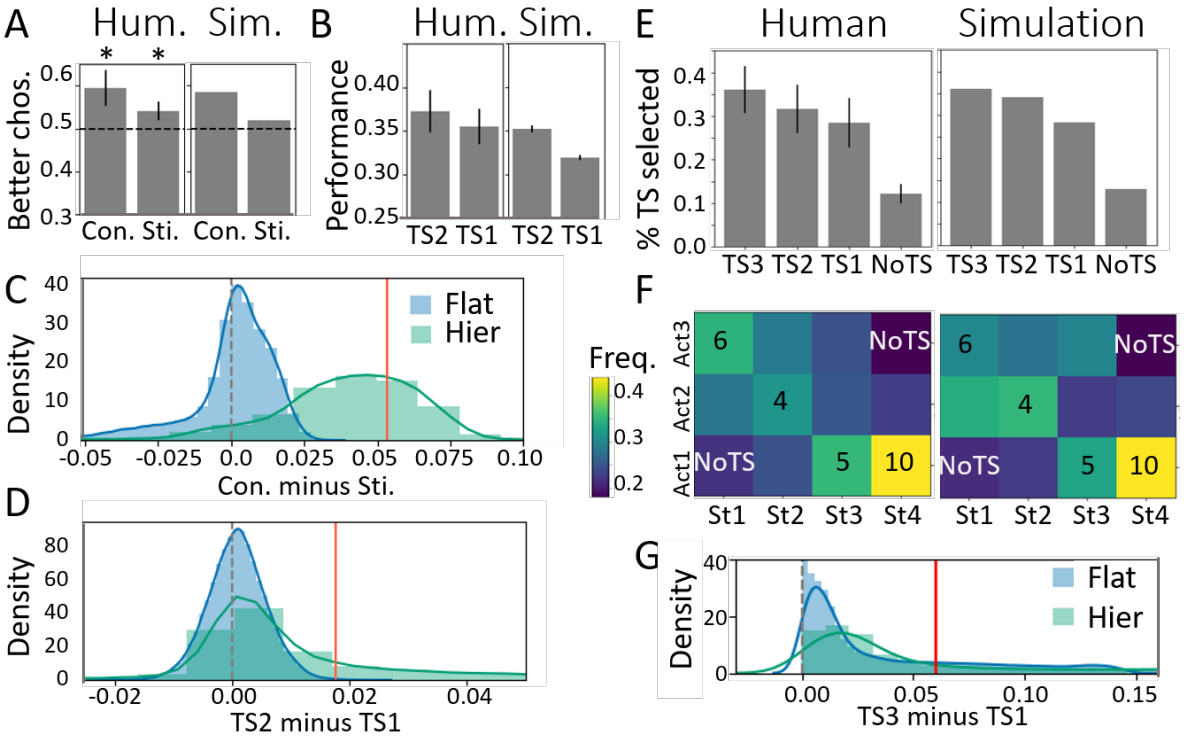

Figure 4: Effects of task-set values. A) Left: Human performance in the comparison phase for contexts and stimuli. Star indicates significant difference from chance (dotted line). Right: Hierarchical simulation. B) Left: Human performance for stimuli with equal rewards, presented in the higher-valued context TS2 or the lower-valued TS1. Performance was better in TS2, despite identical rewards. Right: Hierarchical simulation. C) Model simulations for part A), showing performance differences between contexts (Con.) and stimuli (Sti.). Human behavior in red. D) Model simulations for part B), showing performance differences between task-sets TS1 and TS2. E)-G) Novel context phase. E) Left: Human proportions of actions from each task-set. Right: Hierarchical simulations. F) Left: Raw human action frequencies; Act1, Act2, etc., indicate actions, St1, St2, etc. stimuli. Black numbers indicate action-values in the highest-valued TS3, selected frequently. "NoTS" indicates actions that were incorrect in all task-sets, selected rarely. Right: Hierarchical simulation. G) Difference between choosing TS3 and TS1 for hierarchical and flat simulations, human behavior in red, dotted line for chance.

after a context switch (Fig. 3D). Average accuracy over all flat simulations was above chance, $mean = 0.36$ ($sd = 0.089$), and intrusion errors occurred right at chance, $mean = 0.33$ ($sd = 0.020$), as expected; the marginal likelihoods for the flat model were 0.011 for accuracy and 0.00 for intrusion errors. Average accuracy over all hierarchical simulations, on the other hand, was below chance, $mean = 0.31$ ($sd = 0.028$), while intrusion errors occurred above chance, $mean = 0.44$ ($sd = 0.090$), like in humans. Model likelihoods were 0.49 and 0.50, respectively, resulting in Bayes factors of 44.54 for accuracy and infinity for intrusion errors, in favor of hierarchical RL. An adapted version of flat RL that learned action values for all stimuli but ignored contexts showed intrusion errors, but failed to show other markers of hierarchy, as expected. Thus, the hierarchical model provided a better account of human intrusions errors than the flat model, and confirms our prediction that a hierarchical RL process can underlie transfer effects.

Effects of task-set values

So far, we have shown evidence that participants used, reactivated, and transferred task-sets. Nevertheless, various structure learning models make this prediction (e.g., Bayesian structure learning models) (17; 34). In the following, we therefore test predictions that are unique to hierarchical RL. Specifically, we will assess the role of task-set val-

ues, a construct unique to hierarchical RL, and show that participants acquired these values and that they affected performance.

Effect of task-set values on subjective preference

We first assessed subjective preferences as a proxy for task-set values (32), employing the comparison phase. Here, participants saw either two contexts (context condition) or two stimuli in a context (stimulus condition), and selected their preferred one. Given our task design, participants had experience assessing stimuli in contexts (stimulus condition), but not contexts alone (context condition). We therefore reasoned that value-based context preferences would provide evidence for the acquisition of task-set values.

In the stimulus condition, participants preferred stimuli with larger action-values to those with smaller action-values, confirming the acquisition of action-values (repeated-measures t-test, $t(28) = 2.23, p = 0.034$, Fig. 4A). Similarly in the context condition, participants preferred contexts with larger task-set values, $t(28) = 2.30, p = 0.029$. This is evidence for the acquisition of task-set values, as predicted by hierarchical RL, because it shows that participants acquired preferences at a high level of abstraction, which was not strictly necessary.

The comparison phase also provided another test for hierarchical RL, because it predicts that retrieving action-values requires two steps, whereas retrieving task-set values requires just one. This implies, counter-intuitively, better performance for contexts than stimuli. Human behavior indeed showed this pattern, with faster RTs for contexts than stimuli, $t(28) = 2.10, p = 0.044$, consistent with the idea that participants needed two steps to retrieve stimulus preferences, but had direct access to context preferences. Note that a process model of flat RL would predict the opposite pattern because here, context-stimulus values are directly available, whereas context values need to be inferred.

A similar pattern also emerged for accuracy, such that participants performed better on contexts than stimuli (Fig. 3), although the difference was not significant in the current sample, $t(28) = 1.13, p = 0.27$ (significant in the replication, see suppl. Table S1). Because our models do not simulate RTs, we focused on accuracy differences for model comparison. Average differences were small in flat RL, $mean = -0.38\%$ ($sd = 2.59\%$), as expected, but large in hierarchical RL, $mean = 4.16\%$ ($sd = 3.06\%$), with a significant difference between both, $t(59,999) = 300, p < 0.01$. Hierarchical simulations were centered around human behavior, whereas no flat population showed differences of the size observed in humans (Fig. 4C). Estimated marginal likelihoods were 0.32 for hierarchical RL and 0.00 for flat, resulting in an infinite Bayes factor. This is strong evidence that hierarchical RL captured human subjective preferences better, and supports our hypothesis that hierarchical RL underlies structured reasoning.

Effect of task-set values on performance

We showed in the beginning that both action-values and task-set values were associated with better performance, such that participants learned faster after receiving larger rewards, but also during higher-valued contexts in general. To obtain a summary measure of this effect, we restricted our analysis to stimuli with equal rewards in different contexts, identifying two pairs that occurred in TS1 and TS2 (methods). Despite identical rewards, participants performed 1.77% better in TS2 (Fig. 4B), although the difference was not significant in the current sample, $t(28) = 0.54, p = 0.59$ (significant in the replication, see suppl. Table S1).

In simulations, hierarchical RL performed on average 0.88% better in TS2 than TS1 on the selected stimuli, whereas differences were centered around 0.09% for flat RL (Fig. 4D), with a significant difference between both, $t(59,999) = 104, p < 0.001$. The hierarchical model yielded a marginal likelihood of 0.22, the flat model of 0.00, with a Bayes factor of 228.86 in favor of the hierarchical model. This is compelling evidence that hierarchical RL accounted better for human performance differences based on task-set values, and shows that hierarchical RL might underlie interesting idiosyncrasies in human learning.

Effect of task-set values on generalization

We next assessed whether participants used task-set values when generalizing previous knowledge to new contexts. Specifically, were higher-valued task-sets preferred to lower-valued ones?

In the novel context phase, participants responded to known stimuli in a new context. Despite the lack of feedback, and contrary to self-reports, participants showed consistent preferences for certain stimulus-action combinations over others (Fig. 4F). Stimulus-action combinations of previous contexts were more frequent than new combinations, $t(28) = 2.74$, $p = 0.011$, and within these, combination were more frequent that applied in two contexts than in one,

$t(28) = 2.70$, $p = 0.012$. This confirms that participants reactivated old task-sets in new contexts, in accordance with prior findings (16).

We next assessed the role of task-set values, categorizing actions as correct in task-sets TS3, TS2, TS1, task-sets 3 and 1 (TS3and1), task-sets 2 and 1 (TS2and1), or not correct in any task-set (NoTS). As expected, participants selected TS3 more than TS1, $t(28) = 2.94, p = 0.007$, TS2 more than TS1 $t(28) = 2.52, p = 0.018$, and TS3 more than TS2, $t(28) = 0.65, p = 0.52$, although the last difference was not significant (Fig. 4E). Raw action probabilities were strongly correlated between the hierarchical model and human participants, $r = 0.91, p < 0.001$ (Fig. 4F), confirming that human generalization was captured closely by the model.

In the simulations, hierarchical RL selected actions in two steps like before, whereas flat RL had to create a new set of values (methods). We chose the difference in selecting TS3 versus TS1 as a summary statistic for model comparison (Fig. 4G). Flat RL selected TS3 more often than TS1, $mean = 3.61\%$ ($sd = 4.00\%$), but this preference was stronger in hierarchical RL, $mean = 8.09\%$ ($sd = 8.95\%$), difference, $t(59,999) = 124, p < 0.001$. Likelihoods were 0.19 for flat and 0.34 for hierarchical RL, with a Bayes factor of 1.82. This shows that hierarchical structure provided an additional nudge in explaining human behavior in the novel context, and supports the hypothesis that hierarchical value learning supported generalization in humans.

Relating different markers of hierarchical RL

We showed that for all behavioral markers of hierarchical RL in humans, Bayes factors favored the hierarchical model over the flat. Nevertheless, these analyses could be misleading because different parameters might be necessary to produce each result. As a final step, we therefore assessed all markers simultaneously, and confirmed that the same set of parameters produced human-like behavior in all measures. The right panels of Figs. 3A, 3C, 4A, 4B, 4E, and 4F show the behavior of one such parameter set. Model parameters of all identified sets are shown in suppl. Table S2.

Discussion

The goal of the current study was to assess whether human flexible behavior could be explained as a process involving hierarchical reinforcement learning (RL), i.e., the concurrent use of RL at different levels of abstraction (10; 23). This has been predicted by previous studies (47; 46; 13; 24) and models (16). We proposed a hierarchical RL model in which RL underlies the acquisition of low-level strategies or "task-sets" as well as the high-level selection of these task-sets in appropriate contexts. Employing only simple RL computations, the model showed flexible behavioral switches, continual learning, generalization, and inference based on incomplete information, hallmarks of human intelligent behavior. To assess whether humans employ hierarchical RL, we designed a context-based learning task in which different subtasks targeted at specific aspects of hierarchical RL. Participants' behavior in all subtasks was in accordance with predictions of hierarchical RL.

One prediction was that participants would create a hierarchical representation of the task using context-specific task-sets. Several behavioral markers supported this claim, including asymmetric switch costs, intrusion errors, and a marker of task-set reactivation. This supports the hypothesis that participants formed distinct (switch costs) and coherent task-sets (task-set reactivation), and that the association between contexts and task-sets was itself learned over time (intrusion errors).

Unique predictions of hierarchical RL, which are not shared with flat RL or hierarchical non-RL models, concerned the existence of task-set values and their influence on 1) performance, 2) subjective preference, and 3) generalization. Human behavior showed all three effects: 1) When participants earned identical rewards for stimuli in different contexts, they performed better on the stimulus in the higher-valued context. This behavior is a direct consequence of hierarchical RL because higher-valued task-sets contain more reliable action-values, and therefore allow for more accurate action selection. 2) When asked to pick their preferred contexts, participants selected those more often that were associated with higher-valued task-sets. Because participants had no prior practice comparing contexts, this suggests that they had implicitly learned task-set values, as proposed by hierarchical RL. Participants thereby performed better when selecting contexts than stimulus-context compounds, in accordance with the "blessing of abstraction" of hierarchical representations (29; 33). Hierarchical RL provides an elegant explanation for the latter result because task-set values are retrieved in one step, whereas action-values require two. 3) When faced with a new context, participants more often selected actions from higher-valued than lower-valued task-sets, suggesting that task-set values

influenced generalization. The reuse of entire strategies is an important advantage of hierarchical models, and these results suggest the additional advantage of indicating which strategies should be reused for hierarchical RL.

In sum, human behavior showed several patterns that were uniquely predicted by hierarchical RL. Nevertheless, even though hierarchical RL predicted the observed patterns, it does not necessarily provide the best account for human behavior. We therefore complemented the behavioral analyses with modeling, and tested if our hierarchical RL model fit human behavior better than a standard flat one. Classical model comparison approaches (20; 51; 61) were unavailable in our case because model likelihoods were computationally intractable due to the latent task-set structure. We therefore compared models using Bayes factors, estimating model likelihoods from exhaustive model simulations. This method instantiated an implicit Occam's razor that punished the increased complexity of the hierarchical model. It also allowed us to test whether each model was qualitatively capable of showing human behavior, without restriction to a single (best-fit) set of parameters as in classical model comparison, in addition to providing quantitative model fits. Furthermore, the method allowed us to assess different subtasks and measures independently, revealing interesting differences between them. Finally, it offered more detailed insights into the abilities of each model than parameter fitting alone could have provided.

To exemplify this, remember that hierarchical RL predicts stronger subjective preferences for contexts than for context-stimulus combinations. Note that such a difference need not arise from a hierarchical RL process, but could be caused by other factors as well: Concretely, the task had three contexts, but twelve different context-stimulus combinations, such that selecting contexts might just be easier than selecting stimuli, without reference to hierarchical RL. Our simulation approach resolves this concern because flat and hierarchical simulations were faced with the exact same task, such that behavioral differences must arise from differences in cognitive architecture. In other words, we could assess the ability of each model to explain human behavior while controlling for potential confounds in task design.

Bayes factors confirmed that hierarchical RL accounted better for human behavior than flat RL. For two behavioral markers—the relationships between task-set values and performance and task-set values and generalization—the difference was quantitative, i.e., human behavior was more likely under the hierarchical model. For the remaining three markers, differences were qualitative: Only the hierarchical model was able to show task-set reactivation (hidden context phase), intrusion errors (initial learning phase), and context preferences (comparison phase) of the same strength as humans, but not a single flat simulation showed either of these patterns. This suggests that these three markers provide particularly strong evidence for hierarchical RL—a result that could not have been obtained through pure behavioral analyses or classic model fitting.

The results so far showed that individually, each human behavior was more likely under the hierarchical model than under the flat, but the goal should be a model that explains all behaviors using the same underlying process. To assess whether our hierarchical RL model achieved this, we tested whether individual sets of parameters reproduced all human behaviors simultaneously. Several sets of parameters achieved this, showing that seemingly different behaviors, such as trial-and-error learning (initial learning phase), inference of missing information (hidden context phase), subjective preferences (comparison phase), and generalization (novel-context phase), can be explained by the same underlying learning and decision making process.

Computational models differ in the level of analysis (41) at which they describe cognitive function. Our hierarchical RL model is situated at the algorithmic level, aiming to describe mechanistically the cognitive steps that humans employ when making decisions in complex environments. The model is also inspired by the neural organization of human learning circuits (2; 3; 4), thereby reaching out to the implementational level. Many previous models of hierarchical cognition, on the other hand, were situated at the computational level of analysis, assessing behavior in terms of the organism's fundamental (evolutionary) goals. Computational-level models often employ Bayesian inference, highlighting the rationality and statistical optimality of human behavior (56; 57; 50). Bayesian-RL hybrids have combined Bayesian inference at the abstract level with RL at the lower level (17; 28). Lastly, resource-rationality is an approach that combines principles of rationality and cognitive constrains (40), and would explain our results through the lens of effort allocation.

Computational models at different levels of analysis are not mutually exclusive. Bayesian inference offers a perspective based on optimality, but performing Bayesian inference is often extremely expensive or even computationally intractable, making it unlikely that human cognition uses this algorithm. Hierarchical RL, on the other hand, is computationally cheap, making it a promising candidate for the algorithm of human cognition, and is inspired by neural structure, highlighting its potential as a model of implementation. Recent research showed that a neural network that implemented hierarchical RL approximated Bayesian inference and lead to optimal behaviors using simpler computations (16). Hierarchical RL might therefore be an algorithmic model that does not give up the optimality of Bayesian

inference.

Other related models include models that aim to integrate model-based ("goal-directed") and model-free ("habitual") RL. One model proposed a model-free controller at the abstract level and a model-based one at the lower level (19), another suggested the inverse arrangement (22). Both share with our model the intuition that hierarchically-structured RL can explain complex behavior that goes beyond simple stimulus-response learning. Our model extends these ideas from the domain of temporal abstraction to the domain of state abstraction, and also shows how these ideas play out when hierarchical RL is also used for inference, generalization, etc.

Although goals differ between AI and computational cognitive modeling, ideas from both fields have fundamentally shaped each other (52; 36; 18). Hierarchical RL was originally proposed in AI (58; 35), but several frameworks have since been used as models of human cognition, including the options framework (47; 46; 53), successor representation (43; 21), feature-based RL (26), and meta-learning (59; 60). Our model is a very general implementation of hierarchical RL, and not tied to one of these frameworks. The success of all these models in AI and the cognitive sciences alike shows the strength of this approach and its likely adaptation in human cognition.

Classic "flat" RL has been a powerful model for simple decision making in animals and humans, but it cannot explain hallmarks of intelligence like flexible behavioral change, continual learning without catastrophic forgetting, generalization, and inference in the face of missing information. Recent advances in AI have proposed hierarchical RL as a solution to these and other shortcomings, and we found that human behavior showed many signs of hierarchical RL, which were captured better by our hierarchical RL model than a competing flat one.

There is no debate that achieving goals and receiving punishment are some of the most fundamental motivators that shape our learning and decision making. Nevertheless, almost all decisions we face pose more complex problems than what can be achieved by flat RL. Structured hierarchical representations have long been proposed as a solution to this problem, and our hierarchical RL model uses only simple RL computations, known to be implemented in our brains, to solve complex problems that have traditionally been tackled with intractable Bayesian inference. This research aims to model complex behaviors using neurally plausible algorithms, and provides a step toward modeling human-level, everyday-life intelligence.

Methods

Participants

We tested three groups of participants, with approval from UC Berkeley's institutional review board. All were university students, gave written informed consent, and received course credit for participation. The first sample had 51 participants (26 women; mean age$\pm$sd: 22.1 ± 1.5). Due to a technical error, data were not recorded in the comparison phase. The second sample had 31 participants (22 women; mean age$\pm$sd: 20.9 ± 2.1). We added the mixed testing phase for this sample. Two participants were excluded because average performance in the initial learning phase was below 35% (chance is 33%). The third sample had 32 participants (15 women; mean age$\pm$sd $= 20.8 \pm 5.0$) in an EEG-adapted version of the task. Three participants did not reach the final section of the experiment and were excluded from the corresponding analyses. This paper presents the second sample because it is the first sample that includes all phases. Results replicated across the three samples (Suppl. Fig. S1 and Suppl. Table S1).

Task design

The task consisted of several subtasks (Fig. 2). Participants were first trained on context-stimulus-action-associations in an initial learning phase, then underwent several testing phases which assessed specific aspects of hierarchical RL. To alleviate carry-over effects and forgetting between phases, these were interleaved by refresher blocks, shorter versions of the initial learning phase (120 trials).

Participants were told that their goal was to "feed aliens to help them grow as much as possible". A brief tutorial with instructed trials followed, before participants practiced a **simplified task** without (season) contexts. On each trial, participants saw one of four (alien) stimuli and selected one of three (item) actions by pressing J, K, or L on the keyboard (Fig. 2A). Feedback was given in form of a measuring tape whose length indicated the amount of reward. Correct actions were rewarded with long (mean=5.0) and incorrect actions with short tapes (mean=1.0). Feedback was deterministic, i.e., the same action for the same alien always led to the same mean length, but reward amounts were noisy, i.e., lengths were drawn from a Gaussian distribution with sd=0.5 (Fig. 2B). Participants received 10 training

trials per stimulus (40 total). Order was pseudo-randomized such that each stimulus appeared once in four trials, and the same stimulus never appeared twice in a row.

The **initial learning phase** had the same structure as training, but. three contexts were introduced, each with a unique mapping between stimuli and actions (Fig. 2B). Contexts were presented in blocks. In addition, reward amounts varied between 2-10 for correct actions (2-10, Fig. 2B); rewards for incorrect actions remained 1. Feedback was deterministic but noisy, to ensure participants paid attention to the reward amounts rather than treating feedback as binary correct / incorrect. Reward values (Fig. 2B) were chosen to maximize differences between contexts, while controlling for differences between stimuli and actions. Maximum response time was reduced to 1.5 seconds to limit deliberation. The initial learning phase had 468 trials, distributed across 9 blocks separated by self-paced breaks (3 blocks per context; 52 trials per block). Context changes were signaled.

The **hidden context phase** was identical to the initial learning phase and participants knew they would encounter the same stimuli and contexts as before, but this time, contexts were "hidden" (Fig. 2C). There were 9 blocks with 10 trials per stimulus (360 total). Context switches were signaled. In the **comparison phase**, participants were shown two contexts (context condition), or two stimuli in a context (stimulus condition), and selected their preferred one (Fig. 2C). Participants saw each of three pairs of contexts 5 times, and each of 66 pairs of stimuli 3 times, for a total of $15 + 198 = 213$ trials. Participants had 3 seconds to respond. The **novel context phase** was identical to the initial learning phase, except that it introduced a new context in extinction, i.e., without feedback (Fig. 2C). Participants received 3 trials per stimulus (12 total). The **mixed phase** was identical to the initial learning phase, except that contexts as well as stimuli could change on every trial. Participants received 3 blocks of 84 trials (252 total), each with 7 repetitions per stimulus-context combination.

Behavioral analyses

We used repeated-measures t-tests and mixed-effects regression models where applicable. For section , we identified the following pairs of stimuli: Reward 3: stimuli 2 and 3 in TS1 versus stimulus 4 in TS2. Reward 7: stimulus 1 in TS1 versus stimulus 3 in TS2

RL Models

In the **initial learning phase**, flat RL learns one action-value for each context-stimulus combination ($3 \times 3 \times 4 = 36$ values total; Suppl. Fig. S2A), but uses the same value updating and action selection as the hierarchical model. Action selection: $p(a|s,c) = \frac{exp(Q(a|s,c))}{\sum_{a_i} exp(\beta\ Q(a_i|s,c))}$. Value updating: $Q_{t+1}(a|s,c) = Q_t(a|s,c) + \alpha\ (r - Q_t(a|s,c))$. The model had three free parameters α, β, and f (see below).

The hierarchical model is described in the main text. It has 9 task-set and 36 action-values (45 total), and six free parameters α_a, α_{TS}, β_a, β_{TS}, f_a, and f_{TS}. Both models include forgetting parameters f (or f_a and f_{TS}) to accommodate for value decay between blocks: $Q_{t+1} = (1-f)\ Q_t + f\ Q_{init}$. Q-values were initialized at the expected reward for chance performance, $Q_{init} = 1.67$. Test phases utilized the values that had been acquired at the end of the initial learning phase.

Because context information was unavailable in the **hidden context phase**, models could not use context-based values $Q(a|c,s)$ (flat) and $Q(TS|c)$ (hierarchical). Models therefore used new values $Q(a|s,c_{hidden})$ (flat) and $Q(TS|c_{hidden})$ (hierarchical), which were initialized at Q_{init} and updated as usual. For the flat model, this resulted in learning a totally new set of values; for the hierarchical model, it resulted in uncertainty about which task-set was in place, while retaining old action-values. Context-based values were re-initialized for both models after each context switch.

In the **comparison phase**, models selected options by deriving selection probabilities $p(c,s)$ (stimulus condition) and $p(c)$ (context condition) from RL values $V(c,s)$ and $V(c)$. For the flat model, stimulus values $V(c,s)$ were the values of the best action within a context-stimulus pair (c,s), similar to "state values" (52): $V(c,s) = max_a\ Q(a|c,s)$. Context values $V(c)$ were calculated as averages: $V(c) = mean_s\ V(c,s)$. In the hierarchical model, both stimulus values $V(c,s)$ and context values $V(c)$ were "state values": $V(c) = max_{TS}\ Q(TS|c)$ and: $V(c,s) = \max_a\ Q(a|s,TS)\ p(TS|c)$.

Faced with the unknown context in the **novel context phase**, the flat model created new action-values by averaging old action-values across contexts, $Q(a|c_{new},s) = mean_a\ Q(a|c,s)$. This allowed the model to use past experience rather than learning from scratch. The hierarchical model applied the previously highest-valued task-set to the new context: $Q(TS|c_{new}) = max_c\ Q(TS|c)$. Because each simulated population consisted of many individual agents, this leads to similar results as probabilistic selection of task-sets according to task-set values.

Simulations and model comparison

Simulation parameters were chosen by sampling each parameter independently and uniformly from its allowed range: $0 < \alpha_a, \alpha_{TS}, f_a, f_{TS} < 1$; $1 < \beta_a, \beta_{TS} < 20$. We simulated 10 agents per participant to obtain consistent results across simulation runs.

Bayesian model comparison is based on Bayes factors, which quantify the support for one model M_1 compared to another model M_2 by assessing the ratio between their marginal likelihoods, $F = \frac{p(data|M_1)}{p(data|M_2)}$. Likelihoods are marginal when model parameters θ are marginalized out: $p(data|M) = \int p(data|M,\theta)\ p(\theta)\ d\theta$. Due to the uniform sampling of parameters, $p(\theta)$ was uniform in our case, rendering marginalization unnecessary. In other words, likelihoods $p(data|M)$ could be directly estimated from empirical data distributions: For each behavioral marker v, $p(v|M) = \frac{n_{m \leq h}}{60{,}000}$, where $n_{m \leq h}$ is the number of simulated populations with more extreme values than the human sample.

To identify simulations with human-like behavior in all markers, we selected those in which all markers were within 75% to 150% of human behavior. Suppl. Table S2 shows the parameters of nine identified simulations. The first row refers to the simulation presented in figures Fig. 1, Fig. 3, and Fig. 4.

Acknowledgements

We thank Lucy Whitmore and Sarah Master for their enormous contributions to this project.

References

[1] Abler, B., Walter, H., Erk, S., Kammerer, H., and Spitzer, M. (2006). Prediction error as a linear function of reward probability is coded in human nucleus accumbens. *NeuroImage*, 31(2):790–795.

[2] Alexander, G., DeLong, M., and Strick, P. (1986). Parallel Organization of Functionally Segregated Circuits Linking Basal Ganglia and Cortex. *Annual Review of Neuroscience*, 9(1):357–381.

[3] Alexander, W. H. and Brown, J. W. (2015). Hierarchical Error Representation: A Computational Model of Anterior Cingulate and Dorsolateral Prefrontal Cortex. *Neural Computation*, 27(11):2354–2410.

[4] Badre, D. (2008). Cognitive control, hierarchy, and the rostrocaudal organization of the frontal lobes. *Trends in Cognitive Sciences*, 12(5):193–200.

[5] Badre, D. and D'Esposito, M. (2009). Is the rostro-caudal axis of the frontal lobe hierarchical? *Nature Reviews Neuroscience*, 10(9):659–669.

[6] Badre, D. and Frank, M. J. (2012). Mechanisms of Hierarchical Reinforcement Learning in CorticoStriatal Circuits 2: Evidence from fMRI. *Cerebral Cortex*, 22(3):527–536.

[7] Balleine, B. W., Dezfouli, A., Ito, M., and Doya, K. (2015). Hierarchical control of goal-directed action in the corticalbasal ganglia network. *Current Opinion in Behavioral Sciences*, 5:1–7.

[8] Bayer, H. M. and Glimcher, P. W. (2005). Midbrain Dopamine Neurons Encode a Quantitative Reward Prediction Error Signal. *Neuron*, 47(1):129–141.

[9] Botvinick, M. (2012). Hierarchical reinforcement learning and decision making. *Current Opinion in Neurobiology*, 22(6):956–962.

[10] Botvinick, M., Niv, Y., and Barto, A. C. (2009). Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition*, 113(3):262–280.

[11] Botvinick, M., Weinstein, A., Solway, A., and Barto, A. (2015). Reinforcement learning, efficient coding, and the statistics of natural tasks. *Current Opinion in Behavioral Sciences*, 5:71–77.

[12] Chase, W. G. and Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, 4(1):55–81.

[13] Chiang, F.-K. and Wallis, J. D. (2018). Neuronal Encoding in Prefrontal Cortex during Hierarchical Reinforcement Learning. *Journal of Cognitive Neuroscience*, 30(8):1197–1208.

[14] Collins, A. G. (2017). The cost of structure learning. *Journal of cognitive neuroscience*, 29(10):1646–1655.

[15] Collins, A. G., Cavanagh, J. F., and Frank, M. J. (2014). Human EEG Uncovers Latent Generalizable Rule Structure during Learning. *The Journal of Neuroscience*, 34(13):4677–4685.

[16] Collins, A. G. and Frank, M. J. (2013). Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychological Review*, 120(1):190–229.

[17] Collins, A. G. and Koechlin, E. (2012). Reasoning, Learning, and Creativity: Frontal Lobe Function and Human Decision-Making. *PLOS Biology*, 10(3):e1001293.

[18] Collins, A. G. E. (2019). Reinforcement learning: bringing together computation and cognition. *Current Opinion in Behavioral Sciences*, 29:63–68.

[19] Cushman, F. and Morris, A. (2015). Habitual control of goal selection in humans. *Proceedings of the National Academy of Sciences*, page 201506367.

[20] Daw, N. D. (2011). Trial-by-trial data analysis using computational models. *Decision Making, Affect, and Learning: Attention and Performance XXIII*.

[21] Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624.

[22] Dezfouli, A., Lingawi, N. W., and Balleine, B. W. (2014). Habits as action sequences: hierarchical action control and changes in outcome value. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1655):20130482–20130482.

[23] Diuk, C., Schapiro, A., Crdova, N., Ribas-Fernandes, J., Niv, Y., and Botvinick, M. (2013a). Divide and Conquer: Hierarchical Reinforcement Learning and Task Decomposition in Humans. In *Computational and Robotic Models of the Hierarchical Organization of Behavior*, pages 271–291. Springer, Berlin, Heidelberg.

[24] Diuk, C., Tsai, K., Wallis, J., Botvinick, M., and Niv, Y. (2013b). Hierarchical Learning Induces Two Simultaneous, But Separable, Prediction Errors in Human Basal Ganglia. *Journal of Neuroscience*, 33(13):5797–5805.

[25] Donoso, M., Collins, A. G., and Koechlin, E. (2014). Foundations of human reasoning in the prefrontal cortex. *Science*, 344(6191):1481–1486.

[26] Farashahi, S., Rowe, K., Aslami, Z., Lee, D., and Soltani, A. (2017). Feature-based learning improves adaptability without compromising precision. *Nature Communications*, 8(1):1768.

[27] Flesch, T., Balaguer, J., Dekker, R., Nili, H., and Summerfield, C. (2018). Comparing continual task learning in minds and machines. *Proceedings of the National Academy of Sciences*, 115(44):E10313–E10322.

[28] Frank, M. J. and Badre, D. (2012). Mechanisms of Hierarchical Reinforcement Learning in Cortico-Striatal Circuits 1: Computational Analysis. *Cerebral Cortex*, 22(3):509–526.

[29] Gershman, S. J. (2017). On the Blessing of Abstraction. *Quarterly Journal of Experimental Psychology*, 70(3):361–365.

[30] Griffiths, T. L., Callaway, F., Chang, M. B., Grant, E., Krueger, P. M., and Lieder, F. (2019). Doing more with less: meta-reasoning and meta-learning in humans and machines. *Current Opinion in Behavioral Sciences*, 29:24–30.

[31] Haruno, M. and Kawato, M. (2006). Heterarchical reinforcement-learning model for integration of multiple cortico-striatal loops: fMRI examination in stimulus-action-reward association learning. *Neural Networks*, 19(8):1242–1254.

[32] Jocham, G., Klein, T. A., and Ullsperger, M. (2011). Dopamine-Mediated Reinforcement Learning Signals in the Striatum and Ventromedial Prefrontal Cortex Underlie Value-Based Choices. *Journal of Neuroscience*, 31(5):1606–1613.

[33] Kemp, C., Perfors, A., and Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, 10(3):307–321.

[34] Koechlin, E. (2016). Prefrontal executive function and adaptive behavior in complex environments. *Current Opinion in Neurobiology*, 37:1–6.

[35] Konidaris, G. (2019). On the necessity of abstraction. *Current Opinion in Behavioral Sciences*, 29:1–7.

[36] Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40.

[37] Lee, D., Seo, H., and Jung, M. W. (2012). Neural Basis of Reinforcement Learning and Decision Making. *Annual review of neuroscience*, 35:287–308.

[38] Lee, T. S. and Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *JOSA A*, 20(7):1434–1448.

[39] Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., and Niv, Y. (2017). Dynamic Interaction between Reinforcement Learning and Attention in Multidimensional Environments. *Neuron*, 93(2):451–463.

[40] Lieder, F. and Griffiths, T. L. (2019). Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, pages 1–85.

[41] Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc., New York, NY, USA.

[42] Miller, E. K. and Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24:167–202.

[43] Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M., Daw, N. D., and Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behaviour*, 1(9):680–692.

[44] Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, 7(3):134–140.

[45] Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., and Wilson, R. C. (2015). Reinforcement Learning in Multidimensional Environments Relies on Attention Mechanisms. *Journal of Neuroscience*, 35(21):8145–8157.

[46] Ribas Fernandes, J., Shahnazian, D., Holroyd, C. B., and Botvinick, M. (2018). Subgoal-and Goal-Related Prediction Errors in Medial Prefrontal Cortex. *bioRxiv*, page 245829.

[47] Ribas Fernandes, J., Solway, A., Diuk, C., McGuire, J. T., Barto, A. G., Niv, Y., and Botvinick, M. (2011). A Neural Signature of Hierarchical Reinforcement Learning. *Neuron*, 71(2):370–379.

[48] Schultz, W. (2013). Updating dopamine reward signals. *Current Opinion in Neurobiology*, 23(2):229–238.

[49] Schultz, W., Dayan, P., and Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, 275(5306):1593–1599.

[50] Solway, A., Diuk, C., Crdova, N., Yee, D., Barto, A. G., Niv, Y., and Botvinick, M. (2014). Optimal behavioral hierarchy. *PLoS computational biology*, 10(8):e1003779.

[51] Sunnaker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., and Dessimoz, C. (2013). Approximate Bayesian Computation. *PLOS Computational Biology*, 9(1):e1002803.

[52] Sutton, R. S. and Barto, A. G. (2017). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA; London, England, 2 edition.

[53] Sutton, R. S., Precup, D., and Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1):181–211.

[54] Taatgen, N. A. (2013). The nature and transfer of cognitive skills. *Psychological Review*, 120(3):439–471.

[55] Tai, L.-H., Lee, A. M., Benavidez, N., Bonci, A., and Wilbrecht, L. (2012). Transient stimulation of distinct subpopulations of striatal neurons mimics changes in action value. *Nature Neuroscience*, 15(9):1281–1289.

[56] Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to Grow a Mind: Statistics, Structure, and Abstraction. *Science*, 331(6022):1279–1285.

[57] Tomov, M. S., Yagati, S., Kumar, A., Yang, W., and Gershman, S. J. (2019). Discovery of Hierarchical Representations for Efficient Planning. *bioRxiv*, page 499418.

[58] Vezhnevets, A. S., Osindero, S., Schaul, T., Heess, N., Jaderberg, M., Silver, D., and Kavukcuoglu, K. (2017). FeUdal Networks for Hierarchical Reinforcement Learning. *arXiv:1703.01161*.

[59] Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., Hassabis, D., and Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, 21(6):860–868.

[60] Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., Blundell, C., Kumaran, D., and Botvinick, M. (2016). Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*.

[61] Wilson, R. C. and Collins, A. (2019). Ten simple rules for the computational modeling of behavioral data. *arxiv*.

[62] Wilson, R. C. and Niv, Y. (2012). Inferring Relevance in a Changing World. *Frontiers in Human Neuroscience*, 5.