PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Deterministic response strategies in a trial-and-error learning task

Holger Mohr*, Katharina Zwosta, Dimitrije Markovic, Sebastian Bitzer, Uta Wolfensteller, Hannes Ruge

Department of Psychology, Technische Universität Dresden, Dresden, Germany

* holger.mohr@tu-dresden.de

OPEN ACCESS

Citation: Mohr H, Zwosta K, Markovic D, Bitzer S, Wolfensteller U, Ruge H (2018) Deterministic response strategies in a trial-and-error learning task. PLoS Comput Biol 14(11): e1006621. https://doi.org/10.1371/journal.pcbi.1006621

Editor: Cory Inman, Emory University, UNITED STATES

Received: February 13, 2018

Accepted: November 2, 2018

Published: November 29, 2018

Data Availability Statement: Matlab analysis scripts and human data are publicly available on GitHub, URL: https://github.com/holger-m/Trial-and-error-learning-modeling-scripts.

Funding: This work was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG), SFB 940, subprojects Z2, A2, A9. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Trial-and-error learning is a universal strategy for establishing which actions are beneficial or harmful in new environments. However, learning stimulus-response associations solely via trial-and-error is often suboptimal, as in many settings dependencies among stimuli and responses can be exploited to increase learning efficiency. Previous studies have shown that in settings featuring such dependencies, humans typically engage high-level cognitive processes and employ advanced learning strategies to improve their learning efficiency. Here we analyze in detail the initial learning phase of a sample of human subjects (N = 85) performing a trial-and-error learning task with deterministic feedback and hidden stimulus-response dependencies. Using computational modeling, we find that the standard Q-learning model cannot sufficiently explain human learning strategies in this setting. Instead, newly introduced deterministic response models, which are theoretically optimal and transform stimulus sequences unambiguously into response sequences, provide the best explanation for 50.6% of the subjects. Most of the remaining subjects either show a tendency towards generic optimal learning (21.2%) or at least partially exploit stimulus-response dependencies (22.3%), while a few subjects (5.9%) show no clear preference for any of the employed models. After the initial learning phase, asymptotic learning performance during the subsequent practice phase is best explained by the standard Q-learning model. Our results show that human learning strategies in the presented trial-and-error learning task go beyond merely associating stimuli and responses via incremental reinforcement. Specifically during initial learning, high-level cognitive processes support sophisticated learning strategies that increase learning efficiency while keeping memory demands and computational efforts bounded. The good asymptotic fit of the Q-learning model indicates that these cognitive processes are successively replaced by the formation of stimulus-response associations over the course of learning.

Author summary

Humans and other animals can learn how to respond to novel stimuli by incrementally strengthening or weakening associations between stimuli and responses based on feedback. Q-learning, which is based on a delta learning rule, has been established as the

standard computational model for associative learning. By comparing the Q-learning model with alternative computational models, we investigate human learning strategies in a simple trial-and-error learning task, where stimuli mapped onto responses one-to-one and correct responses were invariably rewarded. We find that humans can learn more efficiently than predicted by the Q-learning model in this setting. Specifically, we show that some subjects systematically went through the response options and made inferences across stimuli to improve their learning speed and avoid unnecessary errors during the initial learning phase. However, after the initial learning phase, the Q-learning model provided a better prediction than the competing models. We conclude that human learning behavior in our experimental task can be best explained as a mixture of sophisticated learning strategies involving high-level cognitive processes at the beginning of learning, and associative learning facilitating further performance improvements at later learning stages.

Introduction

Learning rewarded stimulus-response associations via trial-and-error can be a powerful strategy, which has been employed successfully in complex learning tasks [1]. However, human learning strategies in trial-and-error learning tasks typically go beyond merely associating stimuli and responses via reinforcement. Instead, it has been shown that humans employ high-level cognitive capabilities like working memory and attention to make learning more efficient by exploiting hidden or overt structure in the environment [2–6]. For example, it was shown that subjects can quickly reactivate previously learned response strategies [7] and incorporate information on unselected response options to improve learning efficiency [8–10]. Building on a long history of research on associative learning [11, 12], recent studies increasingly employed advanced modeling approaches like reinforcement learning or Bayesian and Hidden Markov models to explain human learning strategies in various learning tasks [13–15]. Specifically, Q-learning models have been adapted or extended to account for high-level cognitive processes engaged during learning. For instance, Collins et al. have shown in a series of studies that by adding a working memory module to the standard Q-learning model, human learning can be better explained than by pure associative learning [2, 16, 17], see also [18]. Selective attention also plays an important role in human learning, as demonstrated in studies employing extended reinforcement learning models to capture attention-related processes in multidimensional environments [19, 20]. For example, Leong et al. showed that an extended reinforcement learning model with separate weights for different stimulus dimensions can capture attention-related processes in a trial-and-error learning task [21]. Moreover, several studies have shown that humans incorporate implicit relations and hidden task structure into their learning strategy to make learning more efficient [4, 22–24]. Specifically in probabilistic settings, it was shown that when updating internal beliefs about reward probabilities, humans integrate information about unchosen stimuli-response pairs into the updating process both in tasks overtly presenting the outcome of the unchosen options and in tasks with implicit outcome contingencies [8–10, 14, 15, 25–27]. Using probabilistic reward schemes, including fluctuating reward probabilities or dependencies, these studies showed that modified Q-learning, Bayesian or Hidden Markov models, approximating optimal performance in the respective learning tasks, outperformed the standard Q-learning model serving as a baseline for comparison with the more sophisticated models.

Here we show that even in a simple learning task with deterministic feedback, human learning strategies can be surprisingly complex. Specifically, we introduce novel deterministic

response pattern models to test whether subjects explore response options in a fixed order during the initial learning phase. These deterministic models are compared to three alternative models, which are the standard Q-learning model reflecting pure associative learning, a generic optimal model that fully exploits hidden stimulus-response dependencies, and an intermediate model that exploits dependencies less efficiently than the optimal learning model but more efficiently than the Q-learning model.

Methods

Ethics statement

Our sample consisted of healthy human subjects performing a behavioral task in front of a computer (that is, no fMRI, no TMS or so involved). Since participation in this task was not associated with any physical/emotional risk or discomfort, according to our funding agency (German Research Association, DFG) and German law we did not require an approval by our local review board. All participants were informed about the purpose and the procedure of the study and gave written informed consent prior to the experiment.

Experimental task

Subjects performed a simple stimulus-response learning task with deterministic feedback (N = 85), see also [28]. All subjects (31.8% male, mean age 24.3 years, with a range from 18 to 36 years) were informed about the purpose and procedure of the experiment and gave written informed consent prior to taking part in the experiment, in accordance with the Declaration of Helsinki. Subjects were mainly recruited from a pool of students from the Technische Universität Dresden and were paid a fixed amount of 5€ or received credit points for their participation. In each learning block, a novel set of four stimuli was introduced and subjects had to learn the correct responses to the four stimuli (see Fig 1). The set of responses remained constant across blocks and consisted of the four keys *d, f, k, l* on a computer keyboard, corresponding to the left middle, left index, right index and right middle finger. Each stimulus was associated with a unique correct response, i.e. stimuli mapped onto responses one-to-one. Before performing the task, subjects were instructed that each learning block comprises four different symbols and that responses can be given with the four fingers, but subjects were not informed about the one-to-one property of the stimulus-response mappings. See S1 Text for detailed information on the task instructions. Feedback was given deterministically, i.e. correct/incorrect responses were invariably indicated by positive/negative feedback.

Q-learning

The standard Q-learning model served as a baseline for comparison with more sophisticated models [29]. In Q-learning, associations between stimuli and responses are expressed as Q-values (action values or associative weights), which were set to zero initially and were updated after each trial with learning rate $\alpha \in (0, 1]$ based on the following learning rule:

After positive feedback for stimulus-response pair S_i, R_j:

$$q_{ij}^{t+1} = (1 - \alpha)q_{ij}^{t} + \alpha$$

After negative feedback for stimulus-response pair S_i, R_j:

$$q_{ij}^{t+1} = (1 - \alpha)q_{ij}^{t} - \alpha$$

Response probabilities were determined via the softmax response selection rule with noise parameter $\tau \geq 0$:

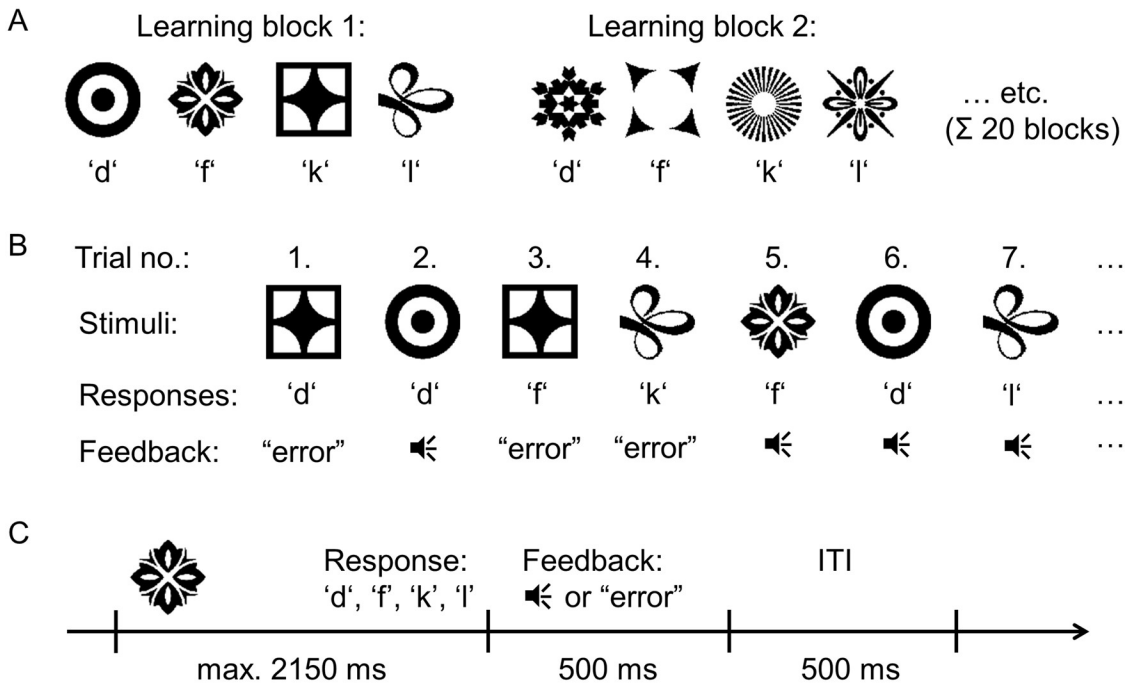

Fig 1. The trial-and-error learning task. A: In each learning block, subjects had to learn the correct responses to four novel stimuli (N = 85). Stimuli mapped onto responses one-to-one, i.e. each stimulus was associated with a unique correct response. Each subject performed 20 learning blocks. B: Stimuli were presented in randomized order, and subjects responded with one of the four keys *d, f, k, l* on a computer keyboard. After response selection, subjects were provided with feedback indicating a correct response via auditory feedback or an incorrect response via the word 'error' written on the screen. Blocks ended when each stimulus had been performed correctly eight times or maximally after 70 trials. C: Response times were limited to 2150 ms, feedback was presented for 500 ms, followed by an inter-trial interval of 500 ms.

https://doi.org/10.1371/journal.pcbi.1006621.g001

Given S_i, the probability for selecting response R_j was:

$$p_{ij} = \exp\left(\frac{q_{ij}}{\tau}\right) \Big/ \sum_k \exp\left(\frac{q_{ik}}{\tau}\right)$$

For the special case $\tau = 0$ (noise-free response selection), responses were selected uniformly among the responses with maximal Q-values.

Note that the Q-learning model updates its associative weights for each stimulus-response (S-R) pair separately, i.e. independently of the other stimulus-response pairs. Hence, this model cannot directly capture dependencies among different stimulus-response-outcome (S-R-O) combinations. Specifically, Q-learning cannot exploit the one-to-one property of the S-R mappings, i.e. the fact that once a response has been associated with a stimulus, this response can be excluded for the other three stimuli.

Free optimal play (FOP)

Based on the literature discussed in the introduction, we hypothesized that subjects may show a tendency towards optimal behavior, i.e. exploit the dependencies among S-R pairs, rather than learning S-R associations independently via reinforcement. In order to maximize expected reward while concurrently minimizing expected uncertainty, the following optimal learning strategy can be employed: Given the 4 stimuli and 4 responses, there are 4! = 24 possible S-R mappings. At the beginning of a learning block, there is no evidence against any of

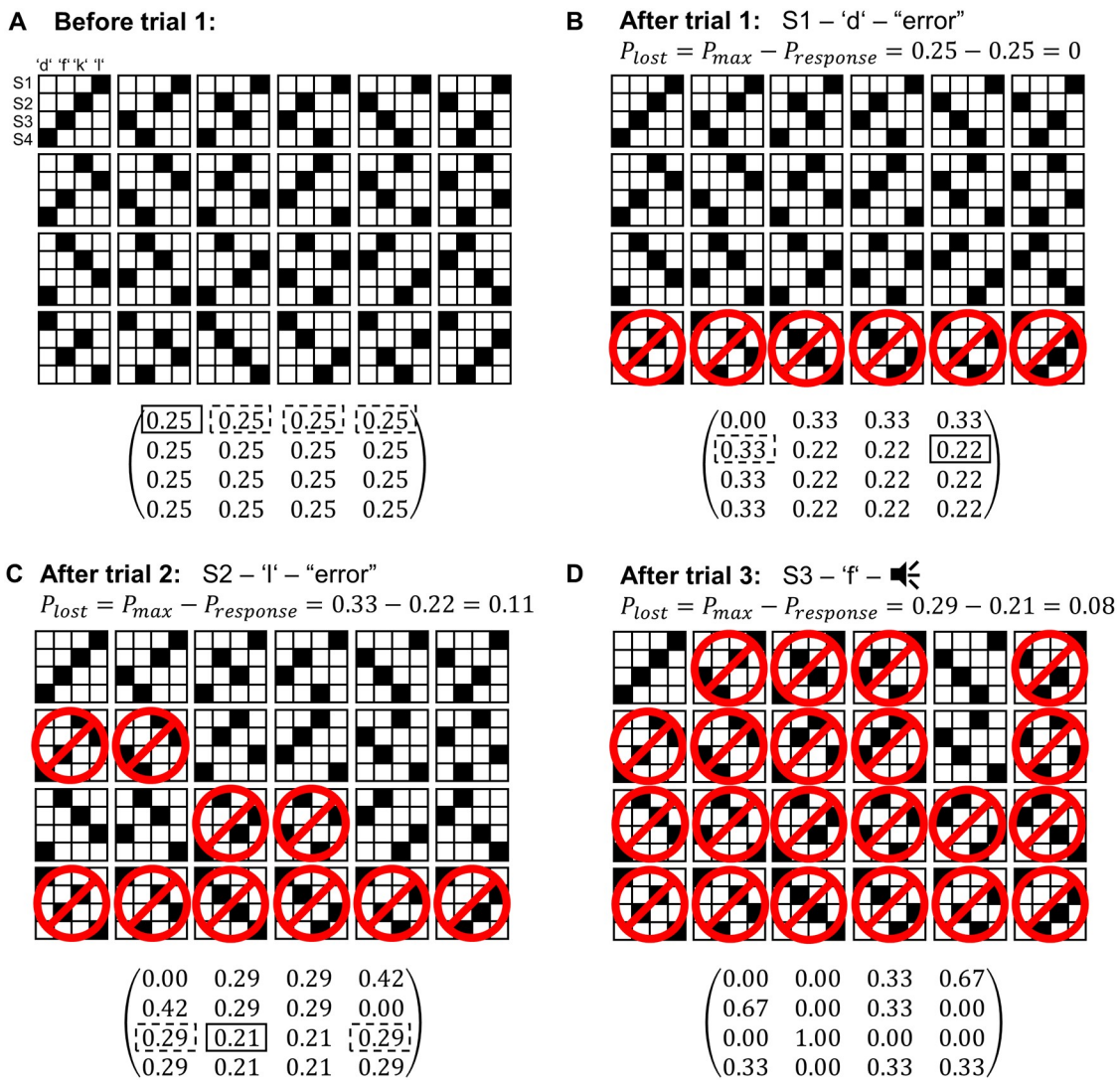

Fig 2. Computation of response probabilities. The four stimuli can map one-to-one onto the four responses in 24 different ways, depicted by the 24 matrices, with rows corresponding to stimuli and columns to responses. As feedback was deterministic, reward was delivered either with probability zero or one, indicated by the white and black squares, respectively. Overall probabilities (shown below the binary matrices) could be computed by averaging across the mappings that were consistent with the S-R-O history. A: At the beginning of a learning block, all 24 mappings were included in the set of consistent mappings. In the presented example, the subject chose response *d* in the first trial (solid box), which is optimal (i.e. provides the maximal likelihood of being rewarded), as were the other three response options (dashed boxes). B: The resulting negative feedback led to the exclusion of all S-R mappings that mapped stimulus S1 onto response *d* (indicated by the red no sign). In the next trial, the subject responded to stimulus S2 with *l*, resulting in negative feedback again. This response was not optimal, as response *d* was more likely than response *l*. C: Based on the feedback information, four additional mappings could be excluded. In the third trial, the subject responded with *f* to stimulus S3, which was correct (but not optimal a-priori). D: Only three S-R mappings are consistent with the S-R-O history at this point. Eventually, only the correct S-R mapping will remain. See S1 Appendix for a technical discussion of the procedure.

https://doi.org/10.1371/journal.pcbi.1006621.g002

these 24 mappings, thus the probability for each mapping is assumed to be 1/24 (see Fig 2). After each trial, the set of S-R mappings that are consistent with the observed S-R-O history is updated. For each S-R pair, the probability of being correct can be computed by averaging across the set of consistent S-R mappings. Selecting the most likely responses according to this procedure maximizes expected reward and minimizes expected uncertainty, hence this strategy is optimal for the presented task, see S1 Appendix for a technical discussion. Moreover,

FOP is the most liberal optimal strategy in the sense that any response sequence generated by an optimal strategy can also be generated by FOP.

The strategy of selecting a response that is maximally likely to be correct is termed free optimal play (FOP) in the following. Note that several responses can be maximally likely, i.e. this learning strategy does not necessarily determine a unique response. As this procedure required tracking the consistency of all 24 S-R mappings and computing averages across subsets of S-R mappings, it seemed unlikely that the subjects implemented this strategy. Yet, we hypothesized that there might be a trend towards this optimal strategy. Indeed, if subjects occasionally exploited the one-to-one property of the S-R mappings, free optimal play might provide a better fit to the data than Q-learning.

As in Q-learning, response selection probabilities in the FOP model were determined by a softmax rule:

Given S_i, the probability for selecting response R_j was:

$$p_{ij} = \exp\left(\frac{\hat{p}_{ij}}{\tau}\right) \Big/ \sum_k \exp\left(\frac{\hat{p}_{ik}}{\tau}\right)$$

with $\hat{p}_{ij}$ denoting the probabilities as computed by the FOP scheme (Fig 2).

Binarized play (BP)

To test whether the subjects tracked the fine-grained differences between response probabilities as provided by FOP, or alternatively, only excluded responses that had already been assigned to a different stimulus, we implemented a simpler version of free optimal play, termed binarized play (BP), that was no longer optimal. The probabilities $\hat{p}_{ij}$ as computed by the FOP model were transformed into a simplified distribution by making all nonzero probabilities uniform, i.e. for a given stimulus S_i, the BP probabilities $\bar{p}_{ij}$ were defined as

$$\bar{p}_{ij} = \begin{cases} 1/\sum_k 1_{\hat{p}_{ik}>0} & \text{if } \hat{p}_{ij} > 0 \\ 0 & \text{if } \hat{p}_{ij} = 0 \end{cases}$$

For example, for a given stimulus S_i, a vector of response probabilities $\hat{p}_i = (0.6, 0, 0.3, 0.1)$, computed according to FOP, was transformed into $\bar{p}_i = (0.33, 0, 0.33, 0.33)$.

Response selection probabilities were again computed via the softmax rule:

Given S_i, the probability for selecting response R_j was:

$$p_{ij} = \exp\left(\frac{\bar{p}_{ij}}{\tau}\right) \Big/ \sum_k \exp\left(\frac{\bar{p}_{ik}}{\tau}\right)$$

Deterministic response patterns (DRPs)

Instead of tracking all 24 S-R mappings as required by FOP, the task could also be optimally performed with reduced memory and computational demands by means of deterministic response strategies. In contrast to FOP, responses are tested in a fixed order for all stimuli, for instance by going from left to right on the keyboard (*dfkl*). In case of negative feedback, the next response according to the response order is tested at the subsequent presentation of the stimulus. Alternatively, if the response is correct, it is logged in for the respective stimulus, and the response is excluded for the remaining stimuli, i.e. the next response to test in the fixed order is the next response that has not yet been assigned to any stimulus (see Fig 3).

Trial no.	Stimulus	Actual response	Feedback	Correct responses	Designated responses
				[-, -, -, -]	[‘d‘, ‘d‘, ‘d‘, ‘d‘]
1.		‘d‘	“error”	[-, -, -, -]	[‘f‘, ‘d‘, ‘d‘, ‘d‘]
2.		‘d‘		[-, ‘d‘, -, -]	[‘f‘, -, ‘f‘, ‘f‘]
3.		‘f‘		[‘f‘, ‘d‘, -, -]	[-, -, ‘k‘, ‘k‘]
4.		‘k‘	“error”	[‘f‘, ‘d‘, ‘l‘, ‘k‘]	[-, -, -, -]
5.		‘k‘		[‘f‘, ‘d‘, ‘l‘, ‘k‘]	[-, -, -, -]

Fig 3. Example for a deterministic response pattern. Using the response order *dfkl*, the deterministic response pattern unfolds in the following way: Before trial 1, all four stimuli are to be responded by the first response of the response order, which is response *d* in this example. Due to the negative feedback in trial 1, the designated response for stimulus S1 is set to the next response according to the response order, which is *f* in this example. In the second trial, responding with *d* to stimulus S2 results in positive feedback, thus response *d* is logged in as the correct response for this stimulus. Importantly, due to the one-to-one property of the S-R mappings, the response *d* is blocked for the other three stimuli, thus the designated responses for stimuli S3 and S4 are set to the next unoccupied response, which is *f*. In trial 3, response *f* is logged in for stimulus S1, and the designated responses for stimuli S3, S4 are set to the next response according to the response order, which is response *k* in this example. In trial 4, responding with *k* to stimulus S3 results in negative feedback. At this point, again due to the one-to-one property of the S-R mapping, one can conclude that the correct response to stimulus S3 must be *l*. Moreover, although stimulus S4 has not yet been presented at this point, its correct response *k* can already be inferred.

https://doi.org/10.1371/journal.pcbi.1006621.g003

From a theoretical point of view, the order by which the responses are tested is arbitrary, i.e. any of the 24 possible response orders could be used to perform the task. However, we hypothesized that from a human perspective, certain response orders, like *dfkl* (going from left to right) or *lkfd* (going from right to left), might be easier to implement than others.

The deterministic response pattern (DRP) models were implemented as follows: For a given stimulus S_i, the response R_j determined by the respective response order (either the designated or correct response) was set to probability one (i.e. $\tilde{p}_{ij} = 1$) and the other three responses were set to probability zero (i.e. $\tilde{p}_{ik} = 0$ for $k \neq j$). This degenerate distribution was transformed into a response selection probability distribution via the softmax rule:

Given S_i, the probability for selecting response R_j was:

$$p_{ij} = \exp\left(\frac{\tilde{p}_{ij}}{\tau}\right) \Big/ \sum_k \exp\left(\frac{\tilde{p}_{ik}}{\tau}\right)$$

Under the presence of response selection noise ($\tau > 0$), the updating procedure was defined in the following way: If the selected response deviated from the designated response due to response selection noise, only positive feedback led to an update, whereas negative feedback left the internal state of the model unchanged. Although this implementation does probably not fully capture human behavior, it was selected to keep the updating procedure of the DRP models as simple as possible. Moreover, this procedure was motivated by the fact that a deviation from the DRP could either occur as a backward deviation with respect to the response

order, in which case the deviating response had been falsified before and updating was not necessary, or as a forward deviation, and in this case an update based on negative feedback would involve an invalid jump over possibly correct responses, thereby potentially corrupting the DRP procedure.

Remarks on the computational models

The Q-learning, BP, FOP and DRP models cover different types of learning ranging from low-level associative learning to more sophisticated learning strategies involving inferences across different S-R pairs. The Q-learning model, which represents low-level associative learning, only strengthens or weakens the association between the currently presented stimulus and selected response based on the provided feedback information without drawing inferences from this information about the remaining S-R pairs. Associative learning is suboptimal on the presented learning task, as the one-to-one property of the S-R mappings is not utilized by this learning approach. In contrast, the FOP model optimally exploits feedback information by excluding all S-R mappings that are incompatible with the information. While solving the learning task optimally, the FOP strategy requires storing which of the 24 S-R mappings are consistent with the observed history of S-R-O combinations as well as computing response probabilities by averaging across consistent S-R mappings. It seems unlikely that the subjects could accurately implement this strategy during the relatively fast-paced experiment. However, the FOP model might capture an unspecific tendency towards optimality, that is, this model might provide a better explanation than the Q-learning model for response data of subjects that at least occasionally drew inferences based on the one-to-one property of the S-R mappings. Similarly, the BP model, which represents a more coarse-grained version of the FOP model, might provide the best explanation for response data of subjects which only partially exploited the one-to-one property of the S-R mappings without taking subtle differences between response probabilities into account. Both the FOP and the BP models are rather employed with the intention to capture unspecific tendencies towards optimality in the human response data than to demonstrate that the subjects accurately implemented the respective strategies. Instead of implementing the cognitively demanding FOP strategy, we hypothesized that some subjects might have implemented DRP strategies, which retain optimality while requiring less cognitive resources than FOP. Specifically, instead of storing consistency/inconsistency for 24 S-R mappings, DRP strategies only require storing the currently designated test responses and correct responses for four stimuli in working memory. Moreover, instead of computing response probabilities by averaging across different S-R mappings, the DRP procedures only require step-by-step testing of the responses using a fixed response order. Thus, we hypothesized that some subjects might have implemented a DRP strategy in order to minimize the number of errors while keeping working memory and computational efforts bounded.

Maximum likelihood estimates

All models were fitted to the data by maximizing the log-likelihood of the data given the models, i.e. parameters were selected such that the actual responses were maximally likely given the models. The models were fitted on data of the initial learning phase of the learning blocks 6 to 20 (i.e. excluding blocks 1 to 5) to ensure that learning strategies had stabilized, since subjects had not been instructed on the one-to-one property of the S-R mappings before performing the task and thus had to adapt their learning strategies within the first few blocks. Models were fitted on data of the initial learning phase, which started at trial 1 and ended when all four stimuli were performed correctly at least once, i.e. the trial in which the fourth stimulus was

performed correctly for the first time marked the end of the initial learning phase in each learning block. This was motivated by the fact that the DRP, FOP and BP models make no specific predictions for the subsequent practice phase following initial learning, besides the general prediction that correct responses are selected, up to a certain degree of fidelity determined by the noise parameter τ. Note that in contrast to the other models, the Q-learning model does make specific predictions for the practice phase, since S-R association strengths continuously increase with every correct repetition during the practice phase. Model parameters, consisting of the response selection noise $\tau \in (0, 1/6.0, 1/5.8, 1/5.6, \ldots, 1/0.2)$ for the DRP, FOP, BP and Q-learning models, and the learning rate $\alpha \in (0.05, 0.10, 0.15, \ldots, 1.00)$ for the Q-learning model, were fitted separately for each subject on the initial learning phases of the learning blocks 6 to 20.

Model comparisons

Based on the maximum likelihood estimates, we determined for each subject which model provided the best fit to the initial learning phase, i.e. which model obtained the highest log-likelihood score (see S1 Fig, for additional group-level analyses see S2 Text). As expected, the response orders *dfkl* (going from left to right on the keyboard) and *lkfd* (from right to left) were ranked first and second among the DRP models, while the third-ranked response order was *kfdl*, which corresponds to the rather implausible sequence right index finger, left index finger, left middle finger, right middle finger, indicating a false positive hit for this response pattern. Thus, to be on the conservative side and avoid excessive statistical testing, we discarded all response orders but *dfkl* and *lkfd* and constrained our model space to the five models DRP *dfkl*, DRP *lkfd*, FOP, BP and Q-learning for subsequent analyses.

Only reporting which model scored the highest likelihood is in general not very informative, since the difference between the log-likelihood scores of the best and second best model can be arbitrarily small. Thus, we tested subject-wise whether one of the five models fitted the initial learning phase significantly better than competing models by conducting nonparametric Wilcoxon signed-rank tests across the log-likelihood values of the 15 blocks of interest. The five models were compared in an order corresponding to the quality of their predictions: As the DRP models make the most specific predictions on the learning strategy, we first tested for each subject whether either the DRP *dfkl* or DRP *lkfd* model fitted significantly better than the respective competing four models by conducting four pairwise one-sided Wilcoxon signed-rank tests across the 15 blocks of interest, using a significance threshold of $p < 0.05$. That is, if the DRP *dfkl* model fitted the initial learning phase of a given subject significantly better than the DRP *lkfd*, FOP, BP and Q-learning models (all four tests resulted in $p < 0.05$), the respective subject was assigned to the DRP subsample. Subjects were also assigned to the DRP subsample if the DRP *lkfd* model fitted significantly better than the DRP *dfkl*, FOP, BP and Q-learning models. If neither the DRP *dfkl* nor the DRP *lkfd* model outperformed the competing models, we tested whether the FOP model fitted significantly better than the BP and Q-learning models, as the FOP model makes more specific predictions than BP and Q-learning. That is, subjects were assigned to the FOP subsample if the FOP model fitted the initial learning phase significantly better than the BP and Q-learning models (both tests resulted in $p < 0.05$). For the remaining subjects, we tested whether the BP model fitted significantly better than Q-learning on the initial learning phase, and subjects were assigned to the BP subsample if this was the case. Finally, we tested whether Q-learning fitted significantly better than FOP or BP on the remaining subjects, and those subjects were assigned to the Q-learning subsample.

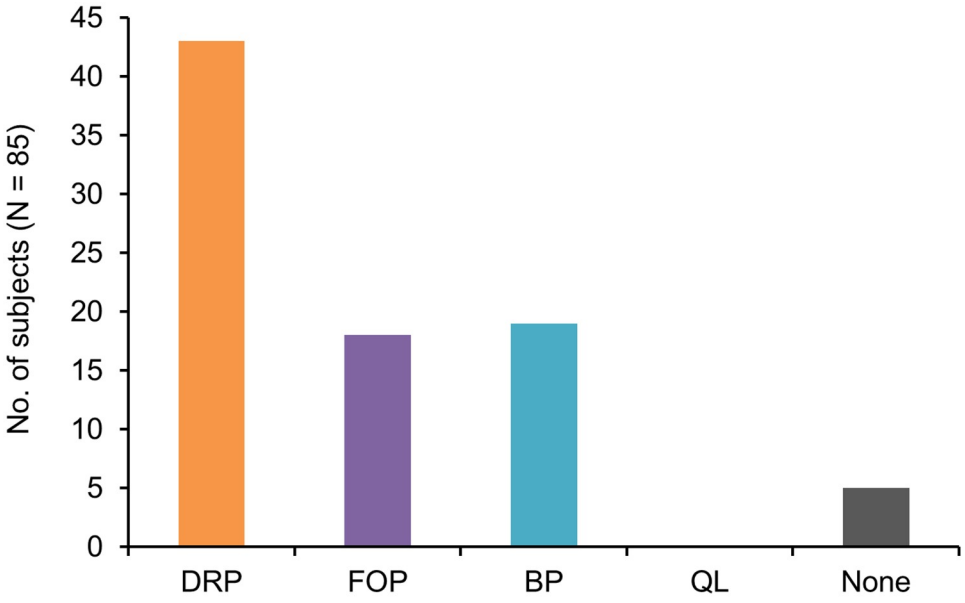

Fig 4. Result of the model comparison procedure. The trial-and-error learning task was performed by N = 85 subjects. For each subject, it was tested in descending order (see main text for details) which model provided the best fit for the initial learning phase. For 43 subjects (50.6%), the DRP models outperformed the FOP, BP and Q-learning models, with 36 subjects following the *dfkl* response pattern and 7 subjects following the *lkfd* pattern. Of the remaining subjects, 18 subjects (21.2%) showed a tendency towards generic optimal learning, while 19 subjects (22.3%) partially exploited stimulus-response dependencies. Q-learning was never significantly better than FOP or BP on the initial learning phase. Five subjects (5.9%) could not be assigned to a model-specific subsample.

https://doi.org/10.1371/journal.pcbi.1006621.g004

Results

Based on this model comparison procedure, we found that the DRP models provided the best fit for 43 of 85 subjects (50.6%), with 36 subjects following the *dfkl* pattern and 7 subjects following the *lkfd* pattern (see Fig 4). A tendency towards generic optimal learning, as expressed by a better fit of FOP than BP and Q-learning, was found for 18 subjects (21.2%), while 19 subjects (22.3%) exploited the stimulus-response dependencies at least partially, as indicated by a better fit of BP than Q-learning. The remaining 5 subjects (5.9%) were assigned to none of the model-specific subsamples. Specifically, Q-learning did not fit significantly better than FOP or BP on the initial learning phase for any subject. Model parameter estimates are shown, separately for the three subsamples, in S2 Fig.

Learning curves

Besides predictiveness, the generative performance of computational models is an important indicator for their ability to explain effects observed in the actual data [30]. To evaluate the generative performance of the five models, we generated response data with N = 1000 repetitions for each block, using the respective subject-specific maximum likelihood model parameters. To evaluate the generative performance of the models in terms of learning dynamics, we compared the learning curves generated by the models with the actual learning curves of the subjects (see Fig 5). While the FOP and BP models provided better fits than Q-learning for the initial learning phase on all three subsamples, the DRP models further improved the fit compared to FOP and BP within the first few trials on the DRP subsample. The Q-learning model provided the best asymptotic fit on all three subsamples, as the DRP, FOP and BP models

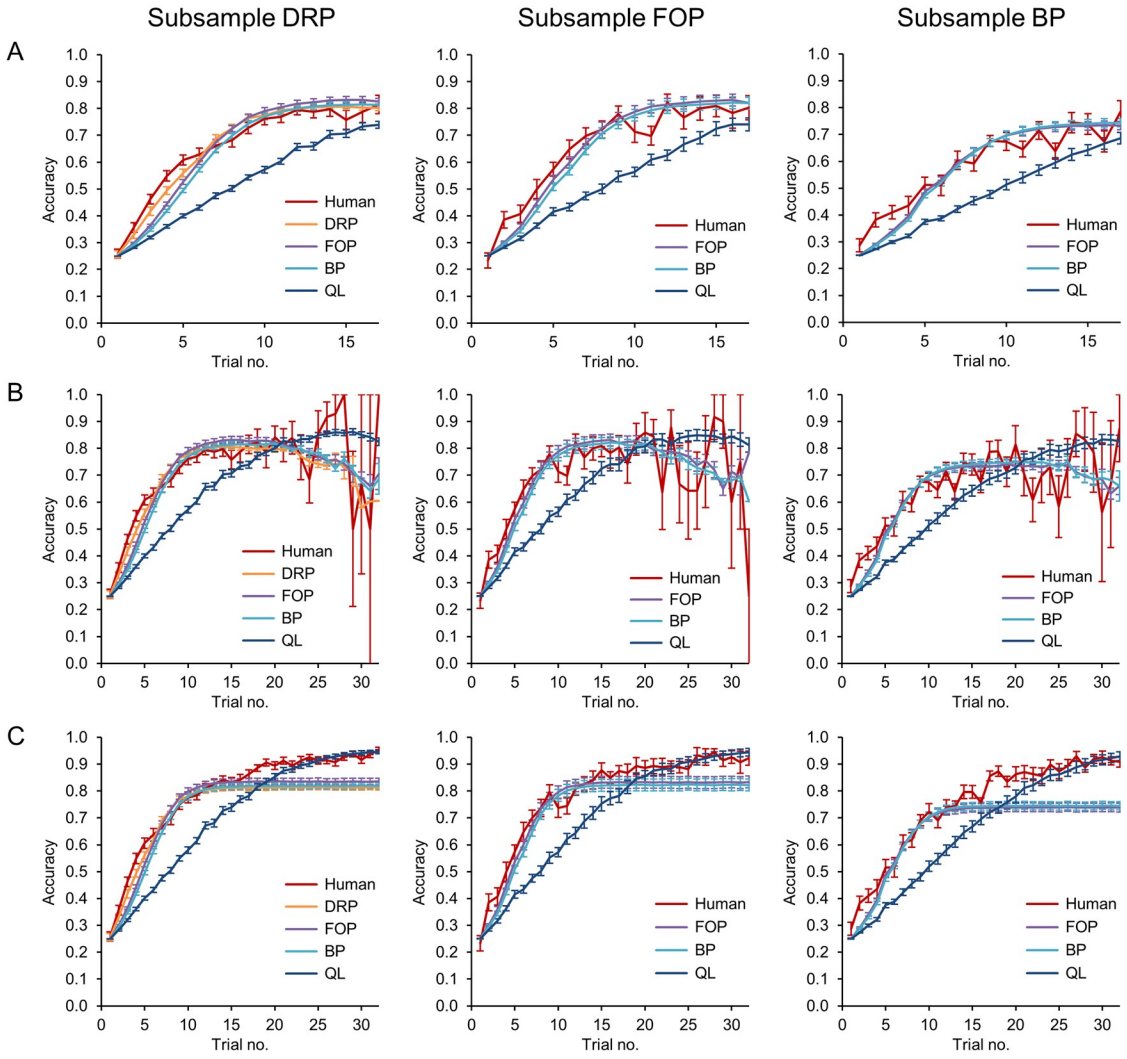

Fig 5. Learning curves for the three subsamples. A: Learning curves of the initial learning phase from trial 1 to 17. For the DRP subsample, the DRP, FOP and BP models provided a markedly better fit to the human learning curve than the Q-learning model. The DRP models improved the fit compared to the FOP and BP models for the first few trials. Within the FOP subsample, again both FOP and BP outperformed Q-learning, with the FOP model providing a marginally better fit than the BP model. For the BP subsample, the FOP and BP learning curves were indistinguishable but again fitted markedly better than Q-learning. Vertical lines indicate standard errors of the mean. B: Learning curves of the initial learning phase from trial 1 to 32. These data are shown for the sake of completeness in addition to the truncated learning curves shown in A. As the initial learning phase ended in 75% of the blocks before trial 18, estimates became increasingly unreliable after trial 17, see also S5 Fig. C: Learning curves including trials of the initial learning phase and the subsequent practice phase. While the DRP, FOP and BP models became stationary when the initial learning phase ended, the Q-learning model further strengthened its associations between stimuli and responses, resulting in the best asymptotic fit on all three subsamples. Note that maximum likelihood estimates of the response selection noise parameter τ were consistently larger than zero, thus the asymptotic performance of the DRP, FOP and BP models was below 100%.

https://doi.org/10.1371/journal.pcbi.1006621.g005

made no specific predictions for the practice phase following initial learning beyond the general prediction that correct responses are selected with a certain degree of fidelity determined by the response selection noise parameter τ.

Optimal and suboptimal errors

While learning curves are typically used to characterize the temporal dynamics of learning processes, they are rather uninformative in terms of the circumstances by which different

types of errors occurred during learning. To evaluate the generative performance of the models in terms of their ability to reproduce specific types of errors that occurred during initial learning, the errors were assigned to different categories. The first category consisted of 'optimal errors', defined as errors that occurred although the subject (or model) had chosen an optimal response, i.e. a response with maximal probability according to optimal (noise-free) FOP. The second category consisted of 'suboptimal errors', defined as errors that occurred for responses with nonzero, but not maximal probability according to noise-free FOP. Using these definitions, we found that the DRP models generated error distributions similar to those actually observed in the DRP subsample, whereas the FOP, BP and Q-learning models could not reproduce the actual distributions (see Fig 6). Specifically, the variability of the number of optimal errors generated by the FOP, BP and Q-learning models was much lower than actually observed. Moreover, these three models produced considerably more suboptimal errors than actually observed within the DRP and FOP subsamples. The results of an extended analysis of error types, including errors that could have been avoided completely with optimal play, can be found in S3 Fig.

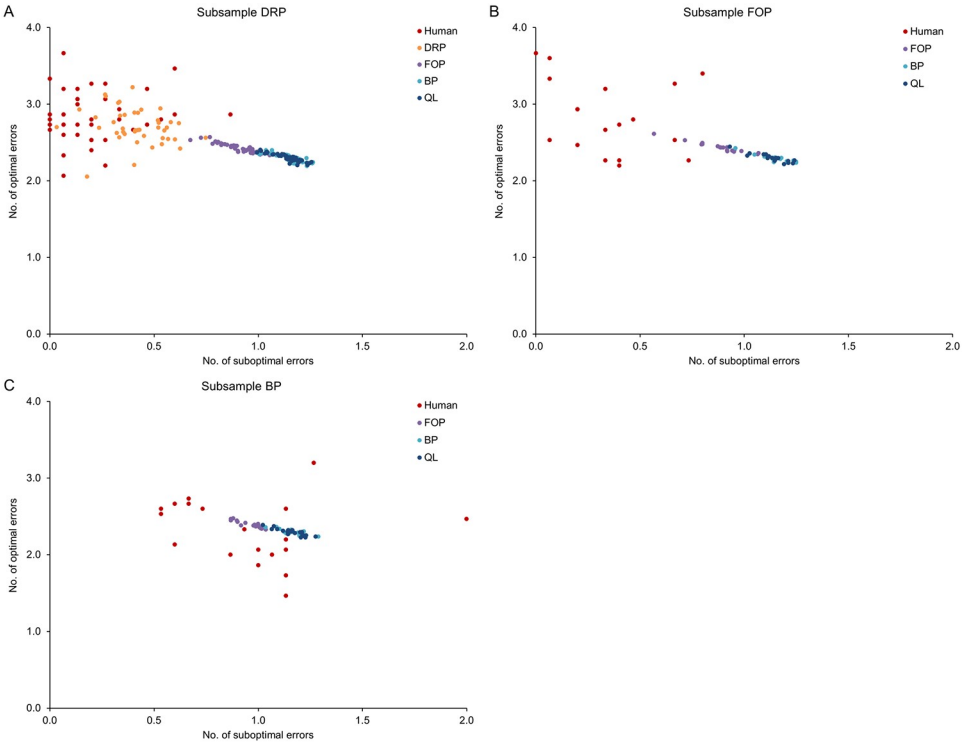

Fig 6. Joint distributions of optimal and suboptimal errors for the three subsamples. Optimal errors were defined as errors occurring when a response with maximum probability of being correct was selected and followed by negative feedback. Suboptimal errors were defined as errors occurring when a response with nonzero probability of being correct, but not maximum probability of being correct, was selected and followed by negative feedback. For each subject, the actual and modeled number of errors was averaged across blocks, i.e. each data point represents mean values of an individual subject. A: For the DRP subsample, the DRP models generated error distributions similar to those produced by the subjects, whereas the variability of optimal errors and the average number of suboptimal errors produced by the FOP, BP and Q-learning models were markedly different from the observed human data. B: Within the FOP subsample, the FOP, BP and Q-learning models again failed to reproduce the variability of optimal errors and the average number of suboptimal errors observed in the actual data. C: For the BP subsample, the three models generated approximately the same number of suboptimal errors as the subjects, but again failed to reproduce the variability of optimal errors.

https://doi.org/10.1371/journal.pcbi.1006621.g006

Why did the FOP, BP and Q-learning models only poorly fit the optimal and suboptimal errors? The Q-learning model acquired stimulus-response associations independently for each S-R pair, hence it could not distinguish between optimal and suboptimal errors, as the computation of response probabilities required inferences across S-R pairs. Moreover, the BP model could also not distinguish between optimal and suboptimal errors, as by definition differences between optimal and suboptimal response probabilities were removed before response selection. While the FOP model could distinguish between optimal and suboptimal errors, and indeed produced slightly better fits for these two error types than the BP and Q-learning models (see S3 Fig), the variability of optimal errors was still considerably reduced compared to the actual data, but also compared to the DRP models (see Fig 6 and S3 Fig). The reason for this reduced variability is that the number of optimal errors produced by FOP is independent of the stimulus sequences. More specifically, under noise-free FOP, the distribution of the number of optimal errors invariably converges towards the distribution shown in S4B Fig for any stimulus sequence. In contrast, for the DRP models, the number of optimal errors varies as a function of the stimulus sequences, as shown in S4A Fig.

Discussion

Using computational models to analyze the initial learning phase of a trial-and-error learning task with deterministic feedback and hidden stimulus-response dependencies, we found that about 50% of the subjects employed deterministic response patterns to increase learning efficiency. Most of the remaining subjects either showed a tendency towards generic optimal learning, or performed better than predicted by pure associative learning by partially exploiting stimulus-response dependencies. A detailed analysis of specific error types showed that only the DRP model could generate the variability found in the human data, whereas the other three models were unable to reproduce this variability.

We followed a modeling approach that has been employed by a variety of studies before [2, 5, 8–10, 19, 20]: The standard Q-learning model served as a baseline for comparison with more sophisticated models that either partially exploited task structure (BP) or approximated optimal performance (FOP), and found that the more sophisticated models provided a better fit to the data than the standard reinforcement learning model. This finding is in line with other studies that have compared pure associative learning with more sophisticated learning strategies in settings with probabilistic feedback [7, 14, 15, 25, 27], deterministic feedback [17, 23], or both types of feedback [31]. Specifically, we found that the BP model provided a better fit to the data than the Q-learning model for a significant fraction of the subjects, which can be unambiguously attributed to certain inferences based on the one-to-one property of the stimulus-response mappings. More specifically, the BP model differed from Q-learning with respect to errors that could have been avoided by excluding responses that had been assigned to other stimuli in previous trials, as indicated by marked differences in specific error categories between these two models across all three subsamples (that is, error categories 'correct for different stimulus', 'both correct for different stimulus and repeatedly wrong' and 'neither correct for different stimulus nor repeatedly wrong', see S3 Fig). Hence, there is good evidence that these subjects exploited the fact that once a response had been assigned to a stimulus, it could be excluded for other stimuli. Similar findings have been reported before for trial-and-error learning tasks featuring two stimuli and probabilistic feedback [8–10, 14, 15, 25, 26].

More surprisingly perhaps, the BP model was outperformed by the FOP model on another significant fraction of the sample. This can be unambiguously attributed to differences in optimal and suboptimal errors, since the two models performed similarly with respect to other error types. These differences indicate that subjects of the FOP subsample did not only exclude

responses previously assigned to other stimuli, as reflected by BP, but also exploited more subtle S-R-O dependencies corresponding to FOP; for instance the fact that when a response was rejected for some stimulus, its probability of being correct for one of the remaining open stimuli increased compared to the other available responses (cf. 'After trial 1' in Fig 2). Similar trends towards optimal task performance based on the integration of task structure into learning strategies have been reported before [3–5, 8].

The novel contribution of the results presented here is that they demonstrate that human learning strategies can be characterized beyond a general trend towards the optimal learning strategy. For 50% of the subjects, the initial learning phase was better explained by the DRP models than by FOP. Thus, these subjects did not select responses arbitrarily from the set of theoretically optimal responses, as predicted by FOP, but instead implemented a response selection procedure that determined a unique response in every trial. On the presented trial-and-error learning task with deterministic feedback, this was a highly adaptive learning strategy: Although being equivalent to FOP from a theoretical point of view, DRPs were more efficient from the human perspective as they considerably reduced working memory and computational demands. Indeed, using DRPs, only the correct or designated response for each stimulus had to be maintained in working memory, whereas FOP required tracking all 24 S-R mappings. Computational costs were also significantly reduced, as the DRPs only required counting up to the next free response in case of negative feedback or storing the correct response in case of positive feedback, whereas FOP required computing response probabilities by averaging across all S-R mappings consistent with the S-R-O history. Moreover, subjects could choose their preferred response order, which was arbitrary from a theoretical point of view, but not from the human perspective, as evidenced by the strongly non-uniform distribution across response orders (S1 Fig).

In order to successfully employ DRPs, subjects were required to reliably update their internal representations of task states on a trial-to-trial basis. Such an explicit and rapid updating of S-R-O contingencies, involving high-level cognitive processes and especially short-term maintenance of S-R-O information in working-memory, has also been reported before in studies on instruction-based and one-shot learning [5, 32–35]. In these learning paradigms, subjects were either explicitly instructed on S-R contingencies [36–39], or had to infer instantaneously the correct response [40–42] or S-O causalities [5] in a single trial. Specifically, by investigating different types of S-R-O contingencies [43, 44] and learning conditions [45, 46], several studies have shown that the explicit instruction of S-R-O contingencies facilitates an almost error-free task performance right from the start of the practice phase. In the light of these studies, the findings presented here suggest that subjects employing DRPs might have divided the trial-and-error learning task implicitly into an exploration phase where they established the correct S-R links (equivalent to the initial learning phase defined for the computational models), and a subsequent practice phase where the S-R associations were consolidated via repetition. Together with the good asymptotic fit of the Q-learning model on the practice phase, our findings suggest that the high-level cognitive system supporting stimulus-response processing during initial learning is successively replaced by an associative system performing automatized, low-level stimulus-response transformations.

Limitations and open questions

In the present study, a trial-and-error learning task with deterministic feedback was analyzed in detail using different computational models. While the employed computational models provided novel insights into human learning strategies in this specific setting, it remains an open question to which extent similar learning strategies can be detected in modified versions

of the task. For example, it remains unclear how learning strategies are impacted by changes of the number of stimuli, type of feedback, assignment of response keys and other factors. Specifically, manipulating feedback probabilities would help to assess whether the deterministic nature of the task is central for the presented findings or not. In this context, it would also be interesting to investigate how a variable bonus (as implemented in other studies on associative learning) would impact learning strategies, compared to the fixed payments used in the current study. Moreover, the computational models employed here did not provide a unified theory about human learning in the investigated setting but instead only covered separate aspects of the involved learning processes. Specifically, the DRP, FOP and BP models could explain human learning better than Q-learning during the initial learning phase, whereas asymptotic learning performance was better predicted by the Q-learning model. Thus, further progress could be made by constructing a unified model that is able to predict the entire learning curve. This might be achieved by combining the models employed here using a mixture parameter that is estimated on the data (c.f. [2, 16, 18]). Another interesting avenue for future research might be to compare the DRP, FOP and BP models not only to the most basic version of Q-learning as employed here, but to more sophisticated associative learning models that allow the integration of task structure, as proposed by Gershman [47]. Further model extensions might incorporate response time data, maybe in the form of drift-diffusion models [48, 49]. Moreover, limitations and open questions of the current study not only concern the computational models per se, but also potential connections between the computational models and more general measures of cognitive performance. Given that only some subjects implemented DRPs, it is conceivable that whether or how accurate DRPs are implemented might correlate with interindividual measures of cognitive capacities like working memory capacity or fluid intelligence across subjects. In summary, the findings presented here might be seen as a first step towards a better understanding of human learning strategies in specific deterministic settings.

Conclusion

Using a computational modeling approach, we showed that the subjects performed the presented trial-and-error learning task using highly adaptive and efficient learning strategies. While 50% of the subjects implemented deterministic response strategies in order to optimize task performance while keeping memory and computational demands bounded, most of the remaining subjects showed a general tendency to exploit hidden stimulus-response dependencies. These sophisticated learning strategies go beyond the incremental reinforcement of stimulus-response associations via feedback, and instead reflect the engagement of high-level cognitive processes during the initial learning phase.

Supporting information

S1 Fig. Preliminary model comparison including all 24 DRP response orders and the FOP, BP and Q-learning models. For each subject, it was determined which of the 27 models provided the largest log-likelihood score based on response data of the initial learning phase. Most subjects were best fitted either by the DRP *dfkl*, DRP *lkfd*, FOP or BP models. The response orders *dfkl* and *lkfd* correspond, respectively, to going from left to right and from right to left on the computer keyboard, which seem to be reasonable response strategies from a human perspective (while from a theoretical perspective, all 24 response orders are equivalent). In contrast, the third-ranked DRP response order *kfdl* corresponds to the rather implausible sequence right index finger, left index finger, left middle finger, right middle finger, and the fourth-ranked response order *fkld* corresponds also to an implausible sequence (left index finger, right index finger, right middle finger, left middle finger). Note that the preliminary

model comparison reported here is based on the best-ranked model for each subject, with the difference between the best and second best model potentially being arbitrarily small. Thus, the fact that for a few subjects some implausible response orders obtained the highest log-likelihood score seems to reflect a bias in the model comparison procedure: Simply by submitting a larger number of models from the same class to the model comparison procedure, it becomes more likely that a model of this class obtains the highest score. To remove this bias from subsequent analyses, we constrained the model space to the five models DRP *dfkl*, DRP *lkfd*, FOP, BP and Q-learning, and conducted statistical tests for model comparison, reported in the main text.
(TIF)

S2 Fig. Maximum likelihood estimates of the model parameters, shown separately for the three subsamples. Response selection noise τ was fitted for all four models DRP, FOP, BP and Q-learning, while the learning rate α was only included in the Q-learning model. Response selection noise τ was optimized along the range 0, 1/6, 1/5.8, . . ., 1/0.2 (31 values), and the learning rate α was selected from the range 0.05, 0.10, . . ., 1.0 (20 values). Parameters were fitted separately for each subject on response data of the initial learning phase.
(TIF)

S3 Fig. Extended analysis of error types of the initial learning phase. Errors were categorized into 7 different types. Optimal errors were defined as errors that occurred when a response with maximum probability of being correct was selected. Suboptimal errors were defined as errors that occurred when a response with nonzero probability, but not maximal probability, was selected. Errors were categorized as 'repeatedly wrong' if negative feedback had been received before for the respective S-R pair. Errors were categorized as 'correct for a different stimulus' if a response was selected that had been assigned to a different stimulus in earlier trials. Errors were categorized as 'both repeatedly wrong and correct for a different stimulus' if both criteria were fulfilled. Errors were categorized as 'neither repeatedly wrong nor correct for a different stimulus' if an indirect inference would have led to the correct response, as for example in step 5 of Fig 3, where the correct response for the fourth stimulus was inferred based on the one-to-one property of the S-R mappings. Finally, errors were categorized as 'after first correct' if the respective S-R pair had been performed correctly before. For each subject, the number of errors of each type was averaged across the 15 blocks of interest. The plots show median, first- and third quartile, and minimum and maximum values across the subjects of the respective sample. A: Data of the DRP subsample. As also depicted in Fig 6 of the main text, the DRP models performed considerably better than the other models in terms of optimal and suboptimal errors. The FOP model showed at least a tendency towards the actual data for these two error types, but all three models (FOP, BP and Q-learning) failed to reproduce the high variability of optimal and suboptimal errors found in the actual data. Moreover, the Q-learning model was unable to exploit the one-to-one property of the S-R mappings, as can be seen by the high rate of 'correct for a different stimulus', 'both repeatedly wrong and correct for a different stimulus' and 'neither repeatedly wrong nor correct for a different stimulus' errors. B: Data of the FOP subsample. As in A, the FOP model showed a slightly better fit in terms of optimal and suboptimal errors than the BP and Q-learning models, but none of the three models could reproduce the high variability of these error types. C: Data of the BP subsample. Again, all three models showed much lower variability in terms of optimal and suboptimal errors than observed in the actual data.
(TIF)

S4 Fig. Distributions of optimal errors under optimal (noise-free) play. A: For any stimulus sequence, the number of optimal errors produced by the 24 response orders invariably resulted

in the shown distribution. B: For a large number of repetitions, the number of errors under free optimal play converged to the same distribution as in A on any stimulus sequence.
(TIF)

S5 Fig. Histograms of block length and initial learning phase length. A: Histogram of the overall block length, including blocks 6 to 20 from all subjects (N = 85). The maximum of 70 trials was never reached after block 5. B: Histogram of the initial learning phase length for blocks 6 to 20. The third quartile (trial no. 17) was taken as cut-off in Fig 5A, indicated by the black/gray shading.
(TIF)

S1 Text. Task instructions.
(PDF)

S2 Text. Additional group-level analyses.
(PDF)

S1 Appendix. Optimality of FOP.
(PDF)

Author Contributions

Conceptualization: Holger Mohr, Uta Wolfensteller, Hannes Ruge.

Formal analysis: Holger Mohr, Dimitrije Markovic.

Methodology: Holger Mohr, Katharina Zwosta, Sebastian Bitzer, Hannes Ruge.

Software: Holger Mohr.

Supervision: Uta Wolfensteller, Hannes Ruge.

Visualization: Holger Mohr.

Writing – original draft: Holger Mohr.

Writing – review & editing: Katharina Zwosta, Dimitrije Markovic, Sebastian Bitzer, Uta Wolfensteller, Hannes Ruge.

References

1. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. Nature. 2015; 518:529–533. https://doi.org/10.1038/nature14236 PMID: 25719670
2. Collins AGE, Frank MJ. How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis: Working memory in reinforcement learning. European Journal of Neuroscience. 2012; 35(7):1024–1035. https://doi.org/10.1111/j.1460-9568.2011.07980.x PMID: 22487033
3. Collins A, Koechlin E. Reasoning, Learning, and Creativity: Frontal Lobe Function and Human Decision-Making. PLoS Biology. 2012; 10(3):e1001293. https://doi.org/10.1371/journal.pbio.1001293 PMID: 22479152
4. Collins AGE, Frank MJ. Cognitive control over learning: Creating, clustering, and generalizing task-set structure. Psychological Review. 2013; 120(1):190–229. https://doi.org/10.1037/a0030852 PMID: 23356780
5. Lee SW, O'Doherty JP, Shimojo S. Neural Computations Mediating One-Shot Learning in the Human Brain. PLOS Biology. 2015; 13(4):e1002137. https://doi.org/10.1371/journal.pbio.1002137 PMID: 25919291
6. Le Pelley ME, Mitchell CJ, Beesley T, George DN, Wills AJ. Attention and associative learning in humans: An integrative review. Psychological Bulletin. 2016; 142(10):1111–1140. https://doi.org/10.1037/bul0000064 PMID: 27504933

7. Donoso M, Collins AGE, Koechlin E. Foundations of human reasoning in the prefrontal cortex. Science. 2014; 344(6191):1481–1486. https://doi.org/10.1126/science.1252254 PMID: 24876345
8. Hampton AN, Bossaerts P, O'Doherty JP. The Role of the Ventromedial Prefrontal Cortex in Abstract State-Based Inference during Decision Making in Humans. Journal of Neuroscience. 2006; 26(32): 8360–8367. https://doi.org/10.1523/JNEUROSCI.1010-06.2006 PMID: 16899731
9. Behrens TEJ, Woolrich MW, Walton ME, Rushworth MFS. Learning the value of information in an uncertain world. Nature Neuroscience. 2007; 10(9):1214–1221. https://doi.org/10.1038/nn1954 PMID: 17676057
10. Boorman ED, Behrens TEJ, Woolrich MW, Rushworth MFS. How Green Is the Grass on the Other Side? Frontopolar Cortex and the Evidence in Favor of Alternative Courses of Action. Neuron. 2009; 62(5):733–743. https://doi.org/10.1016/j.neuron.2009.05.014 PMID: 19524531
11. Thorndike EL. Animal intelligence: An experimental study of the associative processes in animals. The Psychological Review: Monograph Supplements. 1898; 2(4):i–109.
12. Rescorla R, Wagner A. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. Classical Conditioning II: Current Research and Theory. 1972; Vol. 2.
13. Brovelli A, Laksiri N, Nazarian B, Meunier M, Boussaoud D. Understanding the Neural Computations of Arbitrary Visuomotor Learning through fMRI and Associative Learning Theory. Cerebral Cortex. 2008; 18(7):1485–1495. https://doi.org/10.1093/cercor/bhm198 PMID: 18033767
14. Hampton AN, Adolphs R, Tyszka JM, O'Doherty JP. Contributions of the Amygdala to Reward Expectancy and Choice Signals in Human Prefrontal Cortex. Neuron. 2007; 55(4):545–555. https://doi.org/10.1016/j.neuron.2007.07.022 PMID: 17698008
15. Behrens TEJ, Hunt LT, Woolrich MW, Rushworth MFS. Associative learning of social value. Nature. 2008; 456(7219):245–249. https://doi.org/10.1038/nature07538 PMID: 19005555
16. Collins AGE, Brown JK, Gold JM, Waltz JA, Frank MJ. Working Memory Contributions to Reinforcement Learning Impairments in Schizophrenia. Journal of Neuroscience. 2014; 34(41):13747–13756. https://doi.org/10.1523/JNEUROSCI.0989-14.2014 PMID: 25297101
17. Collins AGE, Ciullo B, Frank MJ, Badre D. Working Memory Load Strengthens Reward Prediction Errors. The Journal of Neuroscience. 2017; 37(16):4332–4342. https://doi.org/10.1523/JNEUROSCI.2700-16.2017 PMID: 28320846
18. Viejo G, Khamassi M, Brovelli A, Girard B. Modeling choice and reaction time during arbitrary visuomotor learning through the coordination of adaptive working memory and reinforcement learning. Frontiers in Behavioral Neuroscience. 2015; 9. https://doi.org/10.3389/fnbeh.2015.00225 PMID: 26379518
19. Wilson R, Niv Y. Inferring Relevance in a Changing World. Frontiers in Human Neuroscience. 2012; 5:189. https://doi.org/10.3389/fnhum.2011.00189 PMID: 22291631
20. Niv Y, Daniel R, Geana A, Gershman SJ, Leong YC, Radulescu A, et al. Reinforcement Learning in Multidimensional Environments Relies on Attention Mechanisms. Journal of Neuroscience. 2015; 35(21): 8145–8157. https://doi.org/10.1523/JNEUROSCI.2978-14.2015 PMID: 26019331
21. Leong YC, Radulescu A, Daniel R, DeWoskin V, Niv Y. Dynamic Interaction between Reinforcement Learning and Attention in Multidimensional Environments. Neuron. 2017; 93(2):451–463. https://doi.org/10.1016/j.neuron.2016.12.040 PMID: 28103483
22. Collins AGE, Frank MJ. Motor Demands Constrain Cognitive Rule Structures. PLOS Computational Biology. 2016; 12(3):e1004785. https://doi.org/10.1371/journal.pcbi.1004785 PMID: 26966909
23. Collins AGE. The Cost of Structure Learning. Journal of Cognitive Neuroscience. 2017;. https://doi.org/10.1162/jocn_a_01128
24. Gershman SJ, Niv Y. Learning latent structure: carving nature at its joints. Current Opinion in Neurobiology. 2010; 20(2):251–256. https://doi.org/10.1016/j.conb.2010.02.008 PMID: 20227271
25. Gläscher J, Hampton AN, O'Doherty JP. Determining a Role for Ventromedial Prefrontal Cortex in Encoding Action-Based Value Signals During Reward-Related Decision Making. Cerebral Cortex. 2009; 19(2):483–495. https://doi.org/10.1093/cercor/bhn098 PMID: 18550593
26. Li J, Daw ND. Signals in Human Striatum Are Appropriate for Policy Update Rather than Value Prediction. Journal of Neuroscience. 2011; 31(14):5504–5511. https://doi.org/10.1523/JNEUROSCI.6316-10.2011 PMID: 21471387
27. Boorman ED, Behrens TE, Rushworth MF. Counterfactual Choice and Learning in a Neural Network Centered on Human Lateral Frontopolar Cortex. PLoS Biology. 2011; 9(6):e1001093. https://doi.org/10.1371/journal.pbio.1001093 PMID: 21738446
28. Ruge H, Karcz T, Mark T, Martin V, Zwosta K, Wolfensteller U. On the efficiency of instruction-based rule encoding. Acta Psychologica. 2017;. https://doi.org/10.1016/j.actpsy.2017.04.005 PMID: 28427713
29. Sutton RS, Barto AG. Introduction to Reinforcement Learning. MIT Press; 1998.

30. Palminteri S, Wyart V, Koechlin E. The Importance of Falsification in Computational Cognitive Modeling. Trends in Cognitive Sciences. 2017; 21(6):425–433. https://doi.org/10.1016/j.tics.2017.03.011 PMID: 28476348

31. Bakic J, Jepma M, De Raedt R, Pourtois G. Effects of positive mood on probabilistic learning: Behavioral and electrophysiological correlates. Biological Psychology. 2014; 103:223–232. https://doi.org/10.1016/j.biopsycho.2014.09.012 PMID: 25265572

32. Wolfensteller U, Ruge H. Frontostriatal Mechanisms in Instruction-Based Learning as a Hallmark of Flexible Goal-Directed Behavior. Frontiers in Psychology. 2012; 3:192. https://doi.org/10.3389/fpsyg.2012.00192 PMID: 22701445

33. Ruge H, Wolfensteller U. Towards an understanding of the neural dynamics of intentional learning: Considering the timescale. NeuroImage. 2016; 142:668–673. https://doi.org/10.1016/j.neuroimage.2016.06.006 PMID: 27288320

34. Cole MW, Laurent P, Stocco A. Rapid instructed task learning: A new window into the human brain's unique capacity for flexible cognitive control. Cognitive, Affective, & Behavioral Neuroscience. 2013; 13(1):1–22. https://doi.org/10.3758/s13415-012-0125-7

35. Cole MW, Braver TS, Meiran N. The task novelty paradox: Flexible control of inflexible neural pathways during rapid instructed task learning. Neuroscience & Biobehavioral Reviews. 2017; 81:4–15. https://doi.org/10.1016/j.neubiorev.2017.02.009

36. Ruge H, Wolfensteller U. Rapid Formation of Pragmatic Rule Representations in the Human Brain during Instruction-Based Learning. Cerebral Cortex. 2010; 20(7):1656–1667. https://doi.org/10.1093/cercor/bhp228 PMID: 19889712

37. Ruge H, Wolfensteller U. Functional integration processes underlying the instruction-based learning of novel goal-directed behaviors. NeuroImage. 2013; 68:162–172. https://doi.org/10.1016/j.neuroimage.2012.12.003 PMID: 23246992

38. Meiran N, Pereg M, Kessler Y, Cole MW, Braver TS. The power of instructions: Proactive configuration of stimulus–response translation. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2015; 41(3):768–786. https://doi.org/10.1037/xlm0000063 PMID: 25329082

39. Mohr H, Wolfensteller U, Betzel RF, Misic B, Sporns O, Richiardi J, et al. Integration and segregation of large-scale brain networks during short-term task automatization. Nature Communications. 2016; 7:13217 EP. https://doi.org/10.1038/ncomms13217 PMID: 27808095

40. Cole MW, Bagic A, Kass R, Schneider W. Prefrontal Dynamics Underlying Rapid Instructed Task Learning Reverse with Practice. Journal of Neuroscience. 2010; 30(42):14245–14254. https://doi.org/10.1523/JNEUROSCI.1662-10.2010 PMID: 20962245

41. Cole MW, Reynolds JR, Power JD, Repovs G, Anticevic A, Braver TS. Multi-task connectivity reveals flexible hubs for adaptive task control. Nature Neuroscience. 2013; 16:1348–1355. https://doi.org/10.1038/nn.3470 PMID: 23892552

42. Cole MW, Ito T, Braver TS. The Behavioral Relevance of Task Information in Human Prefrontal Cortex. Cerebral Cortex. 2016; 26(6):2497–2505. https://doi.org/10.1093/cercor/bhv072 PMID: 25870233

43. Mohr H, Wolfensteller U, Frimmel S, Ruge H. Sparse regularization techniques provide novel insights into outcome integration processes. NeuroImage. 2015; 104:163–176. https://doi.org/10.1016/j.neuroimage.2014.10.025 PMID: 25467302

44. Frimmel S, Wolfensteller U, Mohr H, Ruge H. The neural basis of integrating pre- and post-response information for goal-directed actions. Neuropsychologia. 2016; 80:56–70. https://doi.org/10.1016/j.neuropsychologia.2015.10.035 PMID: 26522619

45. Ruge H, Wolfensteller U. Distinct contributions of lateral orbito-frontal cortex, striatum, and fronto-parietal network regions for rule encoding and control of memory-based implementation during instructed reversal learning. NeuroImage. 2016; 125:1–12. https://doi.org/10.1016/j.neuroimage.2015.10.005 PMID: 26471057

46. Mohr H, Wolfensteller U, Ruge H. Large-scale coupling dynamics of instructed reversal learning. NeuroImage. 2018; 167:237–246. https://doi.org/10.1016/j.neuroimage.2017.11.049 PMID: 29175610

47. Gershman SJ. A Unifying Probabilistic View of Associative Learning. PLOS Computational Biology. 2015; 11(11):1–20. https://doi.org/10.1371/journal.pcbi.1004567

48. Ratcliff R, Smith PL, Brown SD, McKoon G. Diffusion Decision Model: Current Issues and History. Trends in Cognitive Sciences. 2016; 20(4):260–281. https://doi.org/10.1016/j.tics.2016.01.007 PMID: 26952739

49. Bitzer S, Park H, Blankenburg F, Kiebel S. Perceptual decision making: drift-diffusion model is equivalent to a Bayesian model. Frontiers in Human Neuroscience. 2014; 8:102. https://doi.org/10.3389/fnhum.2014.00102 PMID: 24616689