



OPEN

Asymmetric reinforcement learning facilitates human inference of transitive relations

Simon Ciranka ^{1,2,4}, Juan Linde-Domingo ^{1,4}, Ivan Padezhki¹, Clara Wicherz¹, Charley M. Wu ^{1,3} and Bernhard Spitzer ^{1,2}

Humans and other animals are capable of inferring never-experienced relations (for example, A > C) from other relational observations (for example, A > B and B > C). The processes behind such transitive inference are subject to intense research. Here we demonstrate a new aspect of relational learning, building on previous evidence that transitive inference can be accomplished through simple reinforcement learning mechanisms. We show in simulations that inference of novel relations benefits from an asymmetric learning policy, where observers update only their belief about the winner (or loser) in a pair. Across four experiments ($n=145$), we find substantial empirical support for such asymmetries in inferential learning. The learning policy favoured by our simulations and experiments gives rise to a compression of values that is routinely observed in psychophysics and behavioural economics. In other words, a seemingly biased learning strategy that yields well-known cognitive distortions can be beneficial for transitive inferential judgements.

Humans routinely infer relational structure from local comparisons. For instance, learning that boxer Muhammad Ali defeated George Foreman can let us infer that Ali would probably win against other boxers that Foreman had defeated. More formally, generalizing from relational observations to new, unobserved relations (for example, knowing that A > B and B > C leads to A > C) is commonly referred to as transitive inference^{1–4}. Transitive inference is not a uniquely human capacity⁵ but can also be observed in non-human primates^{6–8}, rats⁹ and birds^{10–12}.

In the laboratory, transitive inference can be observed after teaching participants the relations between neighbouring elements from an ordered set of arbitrary stimuli (Fig. 1a). The neighbour relations are typically taught through pairwise choice feedback (Fig. 1b) where the relational information is deterministic (that is, if A > B, in our sporting analogy, A would never lose a match against B). Various theories have been proposed to describe how observers accomplish transitive inferences of non-neighbour relations (for example, A > D) in such settings. One class of models posits that observers learn implicit value representations for each individual element (A, B, C and so on), which then enables judgements of arbitrary pairings^{3,13,14}. Alternatively, transitive inference could be accomplished through more explicit, hippocampus-based memory processes^{15–18}, which we will return to below.

Before turning to transitive inference, we consider relational learning in a full-feedback scenario (Fig. 1b) where choice feedback is provided for every possible pairing of items, such that no transitive inference is required. We model implicit value learning in this setting through a simple reinforcement learning (RL) mechanism (Q-learning; Methods) by which relational feedback (for example, ‘correct’ when selecting A over B) may increase the perceived value (Q) of item A and decrease that of item B (Model Q1, Fig. 2a). In this simple RL model, relational feedback symmetrically updates (with opposite signs) the value estimates for both items in a pair. For instance, if Muhammad Ali beat George Foreman, it seems rational to attribute this outcome to Ali’s greater skill as much as to Foreman’s

deficit. We show in simulations that symmetric value updating is in fact optimal in the full-feedback setting. An alternative model with asymmetric learning rates ($\alpha^+ \neq \alpha^-$) applied to the winner and loser in a pair (Model Q2; ‘2’ denotes dual learning rates) learns worse than the symmetric model (Q1) where $\alpha^+ = \alpha^-$ (Fig. 2b,c). Implicit value learning generally gives rise to a ‘symbolic distance effect’^{1,15,20}, where nearby elements are less discriminable (due to more similar value estimates) than elements with greater ordinal distance^{14,21}.

Next, we turn to a partial-feedback setting, which is the typical transitive inference scenario, with feedback being provided only for pairs of items with neighbouring values (Fig. 1b). Here, the simple RL models (Q1 and Q2) effectively learn about stimuli only at the extremes of the ordered set (for example, A and H; Extended Data Fig. 1a), since these are statistically more likely to be winners or losers than their neighbours (under uniform sampling). No value learning occurs for intermediate items (stimuli B to G), since these are equally likely to be paired with lower- and higher-valued stimuli³. However, the model can easily be adapted to performing transitive inference when extending it with a simple assumption: value updates should scale with the difference between the estimated item values, $Q(A) - Q(B)$ (for similar approaches, see refs. 14,21,22). More specifically, to the extent that A is already higher valued than B, observing the expected outcome A > B should induce weaker value updates, whereas the unexpected outcome A < B should induce stronger updates. To illustrate, observing an unknown amateur boxer win against a world champion should induce stronger changes in belief than the opposite, less surprising result (champion > amateur). When we incorporate this simple assumption into our model (Model Q1*), it learns orderly structured values, $Q(A) > Q(B) > \dots > Q(H)$, and thus accomplishes transitive inferences for all pairs of items (Fig. 2d; see also Supplementary Video for an illustration of how our Q-learning models accomplish transitive learning). We also observe a symbolic distance effect with this type of learning under partial feedback, similar to what we observed with simple RL under full feedback (Fig. 2a,d).

¹Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany. ²Max Planck UCL Centre for Computational Psychiatry and Ageing Research, Berlin, Germany. ³Human and Machine Cognition Lab, University of Tübingen, Tübingen, Germany. ⁴These authors contributed equally: Simon Ciranka, Juan Linde-Domingo. e-mail: spitzer@mpib-berlin.mpg.de

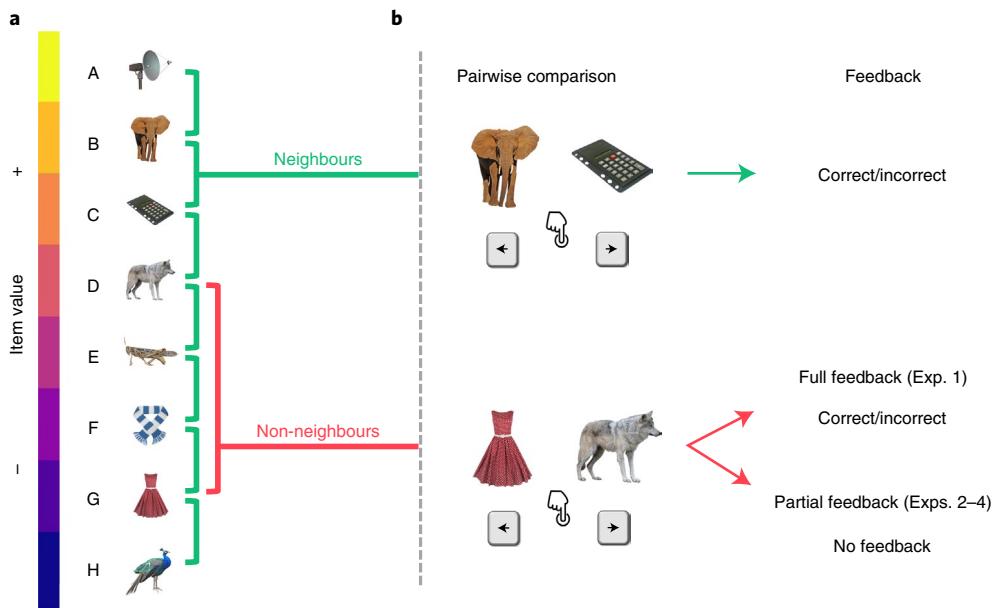


Fig. 1 | Relational learning paradigm. **a**, Example stimulus set and hidden relational value structure. **b**, Example trials for pairwise comparisons of neighbouring (top) and non-neighbouring items (bottom). The participants are asked on each trial to select the higher-valued item. Choices on neighbour trials are always given feedback. Choices on non-neighbour trials are given feedback in the full-feedback condition but not in the partial-feedback condition (see the text for the details). The stimulus images are from the Bank of Standardized Stimuli (BOSS) and licensed under CC-BY-SA 3.0 (<http://creativecommons.org/licenses/by-sa/3.0/>).

Notably, the effect of asymmetric learning rates ($\alpha^+ \neq \alpha^-$, Model Q2*) under partial feedback is strikingly different from what we observed with full feedback. Under partial feedback, optimal performance is achieved with a strongly asymmetric learning policy ($\alpha^+ \gg \alpha^-$ or $\alpha^+ \ll \alpha^-$), where only the winner (or loser) in a pair is updated (Fig. 2e,f and Supplementary Video). In other words, in a setting where hidden relational structure is inferred from only local comparisons, it is surprisingly beneficial to ignore losers (or winners) in outcome attribution. Of note, the winner/loser asymmetry outlined here differs from, and is orthogonal to, previously described asymmetries in learning from positive/negative^{23–26} or (dis-)confirmatory outcomes^{27,28}. A noteworthy aspect of our Model Q2* is that the superior, asymmetric learning policy results in a compression of the observer's latent value structure (Fig. 2f). Selective updating therefore naturally gives rise to diminishing sensitivity towards larger values, as is universally observed in psychophysics²⁹, numerical cognition^{30,31} and behavioural economics³².

Going beyond typical studies of transitive inference with deterministic outcomes, we examined whether our simulation results generalize to scenarios where relational outcomes can be variable, as is the case in many real-world domains such as sports, stock markets and social hierarchies. To this end, we added random variance to the comparison outcomes such that, for example, an item won over its lower-valued neighbour in approximately 80% of cases but lost in the other 20% (see Methods for the details). Intuitively, we allowed for the possibility that competitor A may sometimes lose against B, even if A is generally stronger. We found that our simulation results held for such probabilistic environments, just as they did for deterministic scenarios (Extended Data Fig. 2).

In the models discussed so far, the observer associates each individual item (A, B, C and so on) with an implicit value ($Q(A)$, $Q(B)$, $Q(C)$, ...; item-level learning). An alternative strategy is to more directly learn response preferences for each individual item pairing ($p_{A>B}$, $p_{B>C}$, ...; pair-level learning; Methods). For instance, in our partial-feedback setting (Fig. 1), observers might learn to choose A when comparing A and B, to choose B when comparing B and C,

and so forth, even without relying on value estimates for the individual items. In its simplest form, such memory for pairwise preferences (Model P) only allows learning of pair relations that have been directly experienced (that is, only neighbouring pairs in our partial-feedback setting; Extended Data Fig. 1b, left). However, the pair-level memory can also be extended to allow for transitive inference of more distant, never experienced relations^{8,33,34}: when asked to judge, for example, A versus C, observers might 'chain together' memories of the linking neighbour preferences ($p_{A>B}$ and $p_{B>C}$) through associative recall^{17,35} or spreading activation³⁶ to infer a transitive preference ($p_{A>C}$; Model Pi; Extended Data Fig. 1b, right). Transitive inference based on such pair-level learning gives rise to an inverse symbolic distance effect (Extended Data Fig. 1b, right), where nearby pairs are more discriminable than more distant pairs, reflecting the high dimensionality of the underlying associative memory structure. In modelling our empirical data, we allow for item-level value learning (models denoted by a Q), pair-level learning (models denoted by a P) and a combination of both, in explaining human transitive inference.

Results

We report the results of four experiments ($n = 145$) where we varied whether feedback was full or partial and whether it was probabilistic or deterministic (Methods). In all experiments, the participants were shown a pair of items (drawn from a set of eight) on each trial and were asked to make a relational choice (Fig. 1). The participants were given no prior knowledge about item values and could learn only through trial-and-error feedback.

Full feedback. In Experiment 1 (Exp. 1; $n = 17$), probabilistic choice feedback (Methods) was provided after each of 448 sequential pair comparisons ('full feedback'). Figure 3a shows the mean proportions of correctly choosing the higher-valued item, averaged over all trials in Exp. 1. Descriptively, the choice matrix is dominated by a symbolic distance effect, as predicted by implicit value learning. Fitting our item-level learning models (Q1, Q2, Q1* and Q2*),

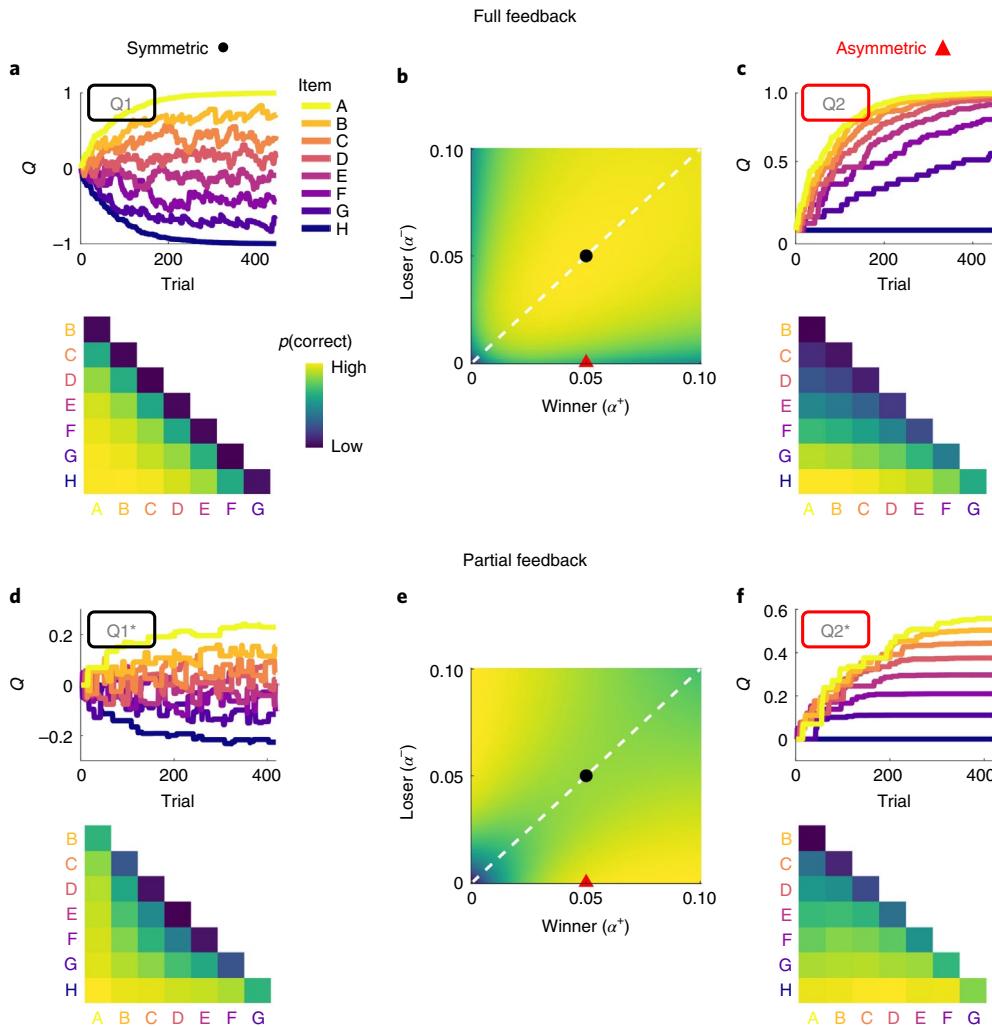


Fig. 2 | Model simulations under full and partial feedback. **a**, Item-level learning under full feedback (Exp. 1) simulated with symmetric Model Q1. Top, exemplary evolution of item values Q (a.u.) over trials. Bottom, simulated probability of making a correct choice for each item pairing (aggregated across all trials in the top panel). **b**, Simulated task performance (mean proportion correct choices on the second half of the trials) of asymmetric Model Q2 across different learning rates α^+ (winning items) and α^- (losing items). For values on the diagonal (dashed white line), Q2 is equivalent to Q1. The black dot indicates parameters used for the simulation of symmetric learning in **a**. The red triangle indicates parameters used for the simulation of asymmetric learning in **c**. **c**, Same as **a**, but using Model Q2 with asymmetric learning rates. **d**, Same as **a**, but for Model Q1* in a partial-feedback scenario (Exps. 2–4). **e,f**, Same as **b** and **c**, but using model Q2* under partial feedback. Note that asymmetric learning leads to lower performance under full feedback (**b**) but improves performance under partial feedback (**e**). Asymmetric learning results in a compressed value structure that is asymptotically stable under partial feedback (**f**) but not under full feedback (**c**).

the best fit to the data is provided by the simplest model (Q1), with a single learning rate for winners and losers (Fig. 3c; protected exceedance probability, $p_{xp}(Q1) > 0.99$; mean Bayesian information criterion (BIC), 361.79 ± 24.68 s.e.m.). In other words, participant behaviour was consistent with a symmetrical updating policy, which our simulations showed to be optimal in the full-feedback setting.

Partial feedback. In Exps. 2–4, choice feedback was provided only on neighbour pairs ('partial feedback') to study transitive inference. In these experiments, we increased the frequency at which the participants were shown neighbouring pairs relative to non-neighbouring pairs to provide more learning opportunities, since the task is inherently harder. We verified that our simulation results were invariant to this modification (Extended Data Fig. 3). Otherwise, the design of Exp. 2 ($n=31$) was identical to that of Exp. 1. Experiment 3 ($n=48$) was an online replication of Exp. 2,

where the pair items on each trial were shown side by side instead of sequentially. Experiment 4 ($n=49$) was similar to Exp. 3, but the feedback was made deterministic (100% truthful), as in previous studies of transitive inference (see Methods for details on the individual experiments).

The choice data from each of the partial-feedback experiments (Exps. 2–4, Fig. 3b) showed clear evidence for transitive inference, with above-chance performance for non-neighbouring pairs that never received feedback (mean accuracy averaged over non-neighbour trials: Exp. 2, 0.714 ± 0.028 ; Exp. 3, 0.698 ± 0.018 ; Exp. 4, 0.709 ± 0.019 ; Wilcoxon signed-rank tests against chance level (0.5): all $P < 0.001$, all $r > 0.84$; see Supplementary Table 1 for the details). Furthermore, the grand mean choice matrices showed the following descriptive characteristics: (1) a symbolic distance effect similar to that observed with full feedback, (2) an asymmetry with greater discriminability of lower-valued items and (3) relatively increased discriminability of neighbour pairs.

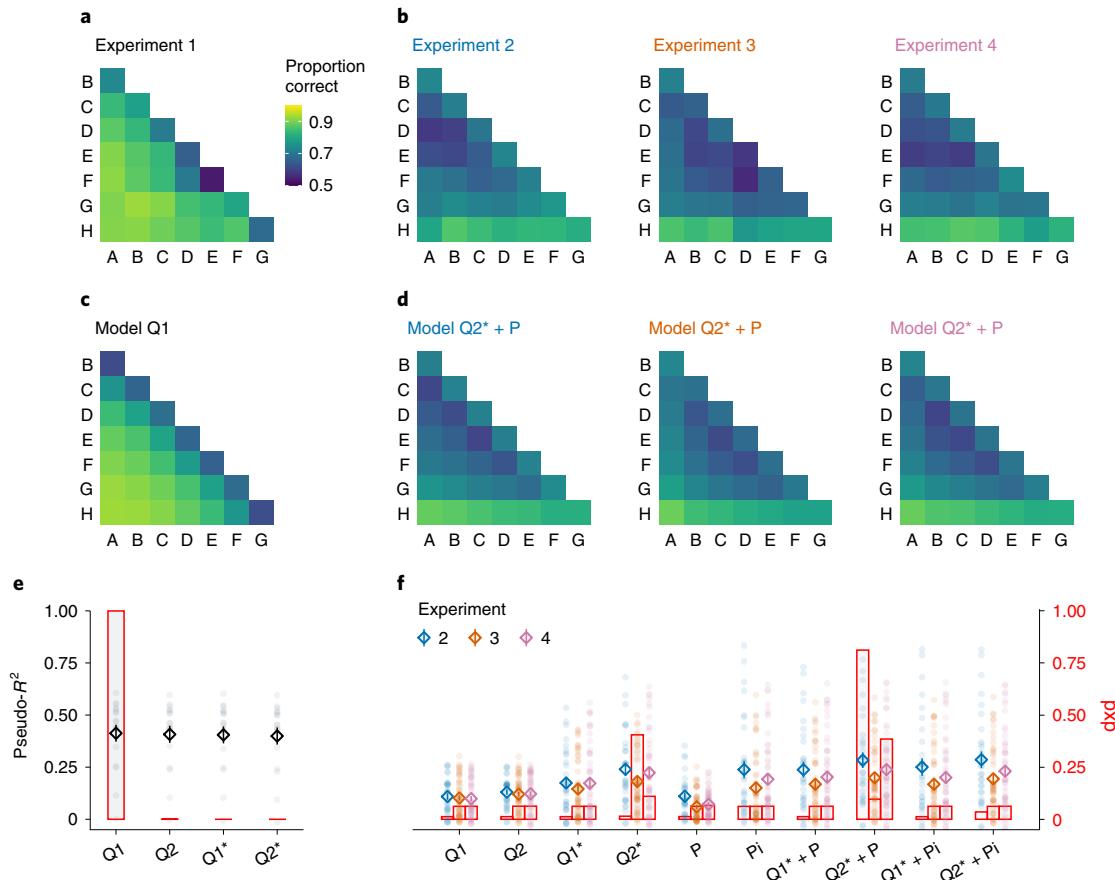


Fig. 3 | Empirical results and model fits. **a,b**, Mean proportions of correct choices observed over all trials in each experiment (with full feedback in **a** and partial feedback in **b**). **c,d**, Mean choice probabilities predicted by the best-fitting model in each experiment. **e,f**, Model comparison (for Exp. 1 in **e** and Exps. 2–4 in **f**). The markers show model fits using a pseudo- R^2 (left y axis; the diamonds and error bars show mean \pm s.e.m., and the dots show individual participants). R^2 is inversely related to BIC, with larger values indicating better fit. Intuitively, $R^2=0$ is equivalent to random chance, while $R^2=1$ corresponds to a theoretically perfect model. The overlaid red bar graphs indicate each model's probability of describing the majority of participants best (right y axis; Methods). The model space is described with the following nomenclature: Q, item-level learning; 1/2, symmetric/asymmetric; *, difference-weighted updating; P, pair-level learning; i, pair-relation-based inference.

The modelling results for the partial-feedback experiments are summarized in Fig. 3d,f (see also Extended Data Figs. 4–8). We highlight two main findings. First, the partial-feedback data were better described by asymmetric models with different learning rates for winners and losers. This held true at every level of model complexity, with our asymmetric models (Q_2 , Q_2^* , $Q_2^* + P$ and $Q_2^* + Pi$) always performing better than their symmetric counterparts (Q_1 , Q_1^* , $Q_1^* + P$ and $Q_1^* + Pi$; Wilcoxon signed-rank tests comparing BICs, Exps. 2–4 combined: all $P < 0.001$, all $r > 0.35$; see Supplementary Table 2 for details), and regardless of whether the partial feedback was probabilistic (Exps. 2 and 3) or deterministic (Exp. 4; comparison of mean BICs between asymmetric and symmetric models: all $P < 0.001$, all $r > 0.67$; see Supplementary Table 3 for the details). In other words, the participants adopted an asymmetric learning policy, which proved superior in our model simulations (Fig. 2e).

Second, behaviour in the partial-feedback scenario was not fully described by item-level value learning alone. The winning model in Exps. 2 and 4 ($Q_2^* + P$; ppx, 0.81 and 0.39; mean BIC, 609.15 ± 12.77 and 434.27 ± 6.29) incorporated pair-level learning in addition to the value estimates of the individual items. This pair-level memory (+P; ‘Models’) accounts for the increased performance for neighbouring pairs (Fig. 3b, first off-diagonals; see also Extended Data Fig. 1b). In Exp. 3, the model comparison was less clear, with model Q_2^*

showing the highest ppx (0.41) but model $Q_2^* + P$ providing a better average fit in terms of BIC (676.86 ± 19.40 versus 692.09 ± 8.61 ; Wilcoxon signed-rank test: $P < 0.001$; $z = -4.53$; $n = 48$; $r = 0.48$; 95% confidence interval (CI), 0.29 to 0.63). However, we found no evidence that pair-level memory contributed to transitive inference in our experiments. Incorporating associative recall of ‘linking’ neighbour pairs (+Pi) worsened the model fits, in terms of both ppx (all ppx < 0.07) and BIC (Exps. 2–4 combined; $Q_2^* + Pi$, 570.42 ± 15.72 ; compared with $Q_2^* + P$, 567.60 ± 15.38 ; Wilcoxon signed-rank test: $P < 0.001$; $z = -6.02$; $n = 128$; $r = 0.53$; 95% CI, 0.37 to 0.69), which is in line with the absence of an ‘inverse’ symbolic distance effect (Extended Data Fig. 1b, right) in the empirical choice data (Fig. 3b).

Figure 4a illustrates how learning of non-neighbour comparisons in our experiments evolved over time. The value compression implied by asymmetric learning of winners (Fig. 2e) predicts relatively better performance for lower-valued pairs (for example, F–H) than for higher-valued pairs (for example, A–C; Fig. 4b, right). We observed no such pattern in Exp. 1 with full feedback (Fig. 4a–c, left). In contrast, participants in Exps. 2–4 with partial feedback showed the critical pattern early on (Fig. 4a–c, right), as predicted by our asymmetric learning models (Fig. 4b, right). Turning to neighbouring pairs (Fig. 4d), which could additionally benefit from pair-level learning (+P; see above), our asymmetric model ($Q_2^* + P$) predicts

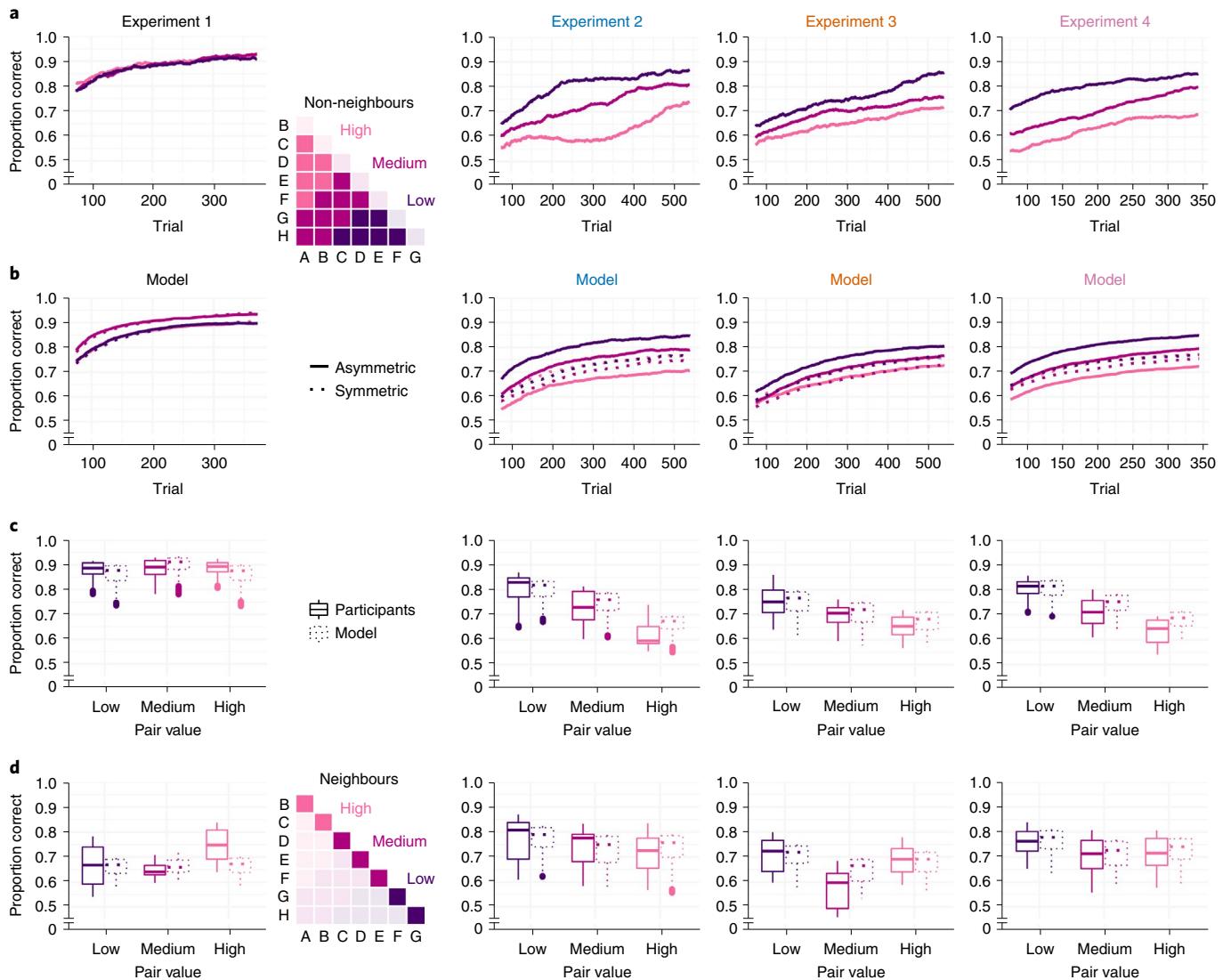


Fig. 4 | Learning curves and signatures of asymmetric learning. **a,b**, Learning of non-neighbour comparisons over time. Mean proportions of correct choices were calculated from a sliding window of 150 trials (**a**). Trajectories are shown separately for low-, medium- and high-valued pairs (see the inset matrix). In **b**, simulated learning curves using the best-fitting model in each experiment (Fig. 3c,d) are shown. The solid lines indicate the best-fitting asymmetric model; the dashed lines indicate the corresponding symmetric model with $\alpha^+ = \alpha^-$. Asymmetric learning in Exps. 2–4 is characterized by systematic performance differences (low > medium > high), as predicted by value compression (Fig. 2f) early on in each experiment. **c**, Box plot summary of the results in **a** and **b**, averaged over all trials. In each plot, the centre line indicates the median, the box limits indicate the upper and lower quartiles, the whiskers indicate 1.5× the interquartile range, and the points indicate outliers. For Exp. 1, $n=17$; for Exp. 2, $n=31$; for Exp. 3, $n=48$; and for Exp. 4, $n=49$. **d**, Same as **c**, for neighbouring item pairs.

only a modest decline in accuracy for higher-valued pairs (see also Fig. 3d), which also matched the empirical data (Fig. 4d, right).

The winner/loser asymmetries described so far might be explained by alternative learning biases, such as asymmetric learning weights for chosen versus unchosen items³⁷. Given above-chance performance, the chosen item will statistically be more likely to be the winning item. To test this alternative explanation, we repeated our modelling analyses using separate learning rates (α^+/α^-) for chosen/unchosen items instead of for the winning/losing item (Methods and equation (4)). This alternative model fit our partial-feedback data significantly worse (mean BIC collapsed across Exps. 2–4, 587.21 ± 15.91 versus 567.60 ± 15.38 ; Wilcoxon signed-rank test: $P < 0.001$; $z = -6.09$; $n = 128$; $r = 0.54$; 95% CI, 0.40 to 0.65), corroborating our interpretation that transitive inference learning was better characterized by asymmetries between winners

and losers. Previous RL studies have also highlighted potential differences in learning from positive (confirmatory) as opposed to negative (disconfirmatory) feedback^{24,25,28}. Extending our winning model to incorporate such confirmation bias (Extended Data Fig. 9) improved the overall model fit (mean BIC collapsed across Exps. 2–4, 537.91 ± 14.45 versus 567.60 ± 15.38 ; Wilcoxon signed-rank test: $P < 0.001$; $z = -8.32$; $n = 128$; $r = 0.72$; 95% CI, 0.67 to 0.81), which is consistent with previous findings in other learning contexts^{24,25,28}. However, the addition of confirmation bias left the finding of winner/loser asymmetries unchanged (Extended Data Fig. 9 and Fig. 3f), thus illustrating the robustness of our results.

We also compared our model family against two previous models of transitive inference (Supplementary Methods): a classic value-transfer model (VAT)²¹ and a more recent model based

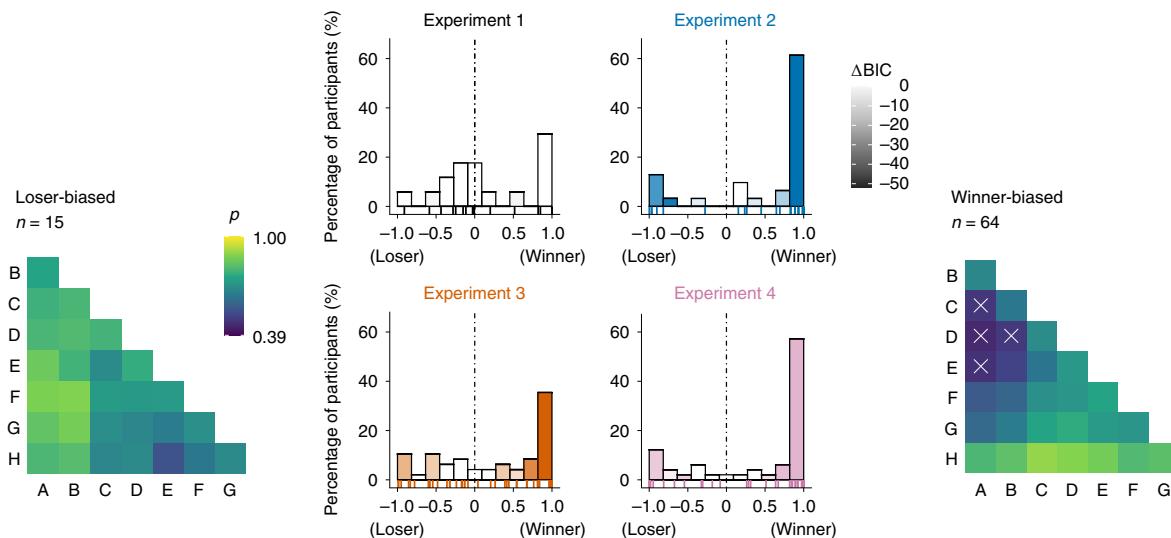


Fig. 5 | Winner/loser asymmetries in individual participants. Centre, histograms of the participants in each experiment sorted according to normalized model-estimated asymmetry: $(\alpha^+ - \alpha^-)/|(\alpha^+ + \alpha^-)|$. Saturation of the bars indicates improvement in model fit (ΔBIC ; darker colours indicate greater improvement) compared with the corresponding symmetric model (that is, Q1 in Exp. 1 and $\text{Q1}^* + \text{P}$ in Exps. 2–4). No improvements can be seen in Exp. 1, where symmetric model Q1 provided the best fit (see also Fig. 3e). The raster plots on the bottom of each panel show individual participant results. The majority of participants ($n = 79$ of 128) in the partial-feedback experiments (Exps. 2–4) showed strongly asymmetric updating either of winning or of losing items, with clear improvements in model fit. Left, mean choice behaviour of participants that were strongly biased towards losers (leftmost bars in the centre plots, Exps. 2–4). p , proportion of correct choices. Right, same as left, for participants strongly biased towards winners (rightmost bars in the centre plots). The white crosses indicate choice accuracies below chance (<0.50).

on ranking algorithms used in competitive sports such as chess (RL-ELO)²². Both VAT and RL-ELO were outperformed by our winning model Q2* + P when fitted to our partial-feedback data (Exps. 2–4 combined; mean BIC VAT, 606.59 ± 14.04 ; RL-ELO, 617.96 ± 15.07 ; Q2* + P, 567.60 ± 15.38 ; Wilcoxon signed-rank tests versus Q2* + P: both $P < 0.001$, both $r > 0.65$). This held true even when we modified VAT and RL-ELO to include pair-level learning (+P) and separate learning rates for winners and losers (mean BIC, 572.95 ± 15.17 and 576.34 ± 15.94 , respectively; both $P < 0.02$, both $r > 0.21$; see Supplementary Table 4 for detailed model comparison results). Our asymmetric Q-learning process thus explains the experimental data better than these earlier models of transitive inference.

Our model simulations (Fig. 2e) indicate two aspects of asymmetric learning that are not directly evident from the group-level results shown in Figs. 3 and 4. First, performance benefits under partial feedback emerged not only for selective updating of winners but likewise for selective updating of losers. Second, performance was highest for extreme asymmetries where the loser (or winner) in a pair was not updated at all. We examined these aspects more closely on the individual participant level (Fig. 5). Half of the participants in Exps. 2–4 ($n = 64$) were indeed characterized by extreme asymmetry towards winners (with α^- near zero). However, another subgroup ($n = 15$) showed the opposite, an extreme asymmetry towards losers (with α^+ near zero). In other words, in the partial-feedback setting, most individuals showed an extreme bias towards winners or losers, either of which proved to be an optimal policy in our model simulations (Fig. 2e and Extended Data Fig. 2, right). In contrast, we found less substantial asymmetries under full feedback (Exp. 1) when allowing the learning rates for winners and losers to vary freely (that is, using model Q2 instead of the winning model Q1). Statistical analysis confirmed that the asymmetries under full feedback (Exp. 1) were significantly lower than under partial feedback (Mann–Whitney U -test of absolute asymmetry indices collapsed over Exps. 2–4 ($n = 128$) compared

with Exp. 1 ($n = 17$): $P = 0.007$; $z = -2.69$; $r = 0.22$; 95% CI, 0.04 to 0.39; Methods).

A potentially surprising observation in the subgroup of participants in Exps. 2–4 who selectively updated winners (Fig. 5, right) is a tendency for below-chance performance for relatively high-valued non-neighbours (for example, A–D) despite each individual performing robustly above chance overall ('Participants'). A potential explanation is that participants may sometimes have confused the two pair items in working memory at the time of feedback (Fig. 1b, right). Such memory confusions would result in the items occasionally being updated with the incorrect learning rate and the incorrect sign ('Models', equation (4)). Under an updating policy that ignores losers, the losing items would then be updated only in error (and always incorrectly), resulting in a negative net learning rate for losers. Indeed, repeating our analysis while allowing for negative values of α^+ and/or α^- yielded a small but significant improvement in model fit (mean BIC, 562.52 ± 15.23 compared with 567.60 ± 15.38 ; Wilcoxon signed-rank test: $P < 0.001$; $z = -5.58$; $n = 128$; $r = 0.50$; 95% CI, 0.36 to 0.62). More specifically, in the $n = 64$ participants who selectively updated winners (that is, with a positive α^+ ; mean $= 0.063 \pm 0.007$), α^- estimates were weakly negative (mean $= -0.009 \pm 0.0016$; Wilcoxon signed-rank test against zero: $P < 0.001$; $z = 4.95$; $n = 64$; $r = 0.64$; 95% CI, 0.47 to 0.78). Memory confusions may thus explain the systematically false inferences about certain item pairs (Fig. 5, right). Together, these findings are consistent with a strongly asymmetric learning mechanism that is also prone to occasional memory errors.

To summarize our empirical findings, when transitive relations could be inferred only from local comparisons (Exps. 2–4), human learning was characterized by an asymmetric outcome attribution to either winners or losers, which proved to be surprisingly optimal in model simulations. In contrast, a symmetric attribution of relational outcomes emerged in a setting where all pair relations could be directly experienced (Exp. 1), and for which our simulations identified symmetric updating to be the most efficient.

Discussion

Reasoning about the relationships between arbitrary pairings of items is a key component of human intelligence. Through simulations, we showed how different learning regimes perform better in full- and partial-feedback contexts. Under full feedback, the best learning model used symmetric learning to update the value estimates for the winning and losing items in opposite directions, with the same magnitude. However, under partial feedback (only for neighbouring items), the best learning model used asymmetric learning to only update the value representations for either the winner or the loser. Across four experiments, we found robust evidence that human learners used the best learning rule to match their feedback context. The participants used symmetric learning under full feedback (Exp. 1) and asymmetric learning under partial feedback (Exps. 2–4). While our asymmetric models allowed for a wide range of possible learning rate combinations, a majority of the participants showed one-sided learning, where value representations were only updated for either winners or losers.

An important feature found both in our model simulations and participant behaviour is a compression of the emerging implicit value structure, which results in a systematic decrease in discriminability of higher-valued items (Fig. 2f and Fig. 4). This resembles the Weber–Fechner law in psychophysics²⁹, where sensitivity to stimulus differences diminishes with increasing magnitude (see also refs. ^{32,38}). While there exist alternative theoretical accounts for this ubiquitous phenomenon³⁹, our findings add a new perspective: compressed representations of magnitude emerge naturally from a learning policy that is optimized for inferring global relationships from local comparisons. From this perspective, subjective compression might not only reflect an efficient adaptation to the distribution of stimuli in the environment^{40–42} but could also result from learning policies that enhance transfer to novel relationships.

In other contexts, previous RL studies have discovered different types of learning asymmetries, such as between positive and negative^{24,25} or confirmatory and disconfirmatory outcomes²⁸. The one-sided learning policy highlighted here in the context of transitive inference is orthogonal to these other asymmetries but may play a similar role in leveraging a biased but advantageous learning strategy (see also refs. ^{27,43,44}). Unlike with ‘optimal’ cognitive biases reported previously^{45–49}, we did not find the benefit of the present learning asymmetries to emerge from general limitations (noise) in decision-making (Extended Data Fig. 10). We speculate that human learners may adopt the present biases more strategically, in settings where the availability of only sparse feedback presages the requirement of future inferential judgements.

Previous theories have proposed richer and more complex cognitive mechanisms for transitive inference, often with an emphasis on the key role of the hippocampus in representing relational knowledge^{15,50}. Early research appealed to the idea that individuals used spatial representations to learn ordered value sequences^{1,8,51}. More recently, various models have been proposed that use associative learning mechanisms to describe how interactions between episodic memories in the hippocampus can generalize relational knowledge from local to distant comparisons^{17,52}. In our present experiments, we found no evidence for transitive inference through such ‘associative linking’ and failed to observe its key empirical prediction (an inverse symbolic distance effect; Fig. 3b and Extended Data Fig. 1b, right). We show instead that simpler mechanisms of value learning^{21,53,54} combined with clever biases (that is, asymmetric learning rates) can be sufficient for performing transitive inference and for accurately describing human learners.

While our model explains transitive inference via learning of individual item values, our findings do not preclude the emergence of a more explicit (for example, map-like) mental model of the items’ relational structure⁵⁵ or the possibility that participants develop direct action policies for each pairing⁵⁶ after learning

progress. A related question for future work is to what extent learning from relational feedback may transfer to comparisons with new items that were not contained in the learning set.

To summarize, we report evidence for pronounced asymmetries in transitive relational learning, where observers selectively update their beliefs only about the winner (or the loser) in a pair. Although asymmetric learning yields distorted value representations, it proves beneficial for generalization to new, more distant relationships. This biased learning regime thus seems well adapted for navigating environments with relational structure on the basis of only sparse and local feedback.

Methods

Participants. The participants in Exps. 1 and 2 were recruited from a participant pool at the Max Planck Institute for Human Development. Of these, $n=20$ participated in Exp. 1 (13 female, mean age 27.15 ± 3.91 years), and $n=35$ participated in Exp. 2 (14 female, 27 ± 3.80 years). The participants in Exps. 3 and 4 were recruited online via Prolific Academic (www.prolific.co), with $n=76$ completing Exp. 3 (23 female, 24.73 ± 5.40 years) and $n=60$ completing Exp. 4 (23 female, 25.92 ± 4.54 years). The participants in Exps. 1 and 2 received compensation of €10 per hour and a bonus of €5 depending on performance. Payment in Exps. 3 and 4 was £4.87 (£1.46 bonus) and £3.75 (£1.12 bonus), respectively. We obtained written informed consent from all participants, and all experiments were approved by the ethics committee of the Max Planck Institute for Human Development.

Participants who did not reach above-chance learning levels were excluded from analysis. The threshold for inclusion was set to 60% correct judgements in the last two blocks of the experiment, which corresponds to a binomial test probability of $P < 0.01$ compared with chance level (50%). After exclusion, $n=17$ (Exp. 1), $n=31$ (Exp. 2), $n=48$ (Exp. 3) and $n=49$ (Exp. 4) participants remained for analysis.

Stimuli, task and procedure. In Exps. 1 and 2, eight pictures of everyday objects and common animals were used as stimuli (Fig. 1a). In Exps. 3 and 4, we included 12 additional pictures of objects and animals and selected for each participant a new subset of 8 images as stimuli. An additional set of 8 pictures was used for instructions and practice purposes in each experiment. All images were from the BOSS database⁵⁷, with the original white background removed.

All experiments involved learning the latent relations between the eight stimuli (A > B > C > D > E > F > G > H) through pairwise choice feedback, where the latent value structure was pseudo-randomly assigned to the pictures for each participant. On each trial, a pair of pictures was presented, and the observers were asked to choose the higher-valued stimulus (two-alternative choice with time-out). All possible stimulus pairings (7 neighbouring and 21 non-neighbouring) were randomly intermixed across trials, with randomized ordering of the elements in a pair (for example, A–B or B–A). Prior to all experiments, the participants were given written instructions and were asked to complete two brief practice blocks to become familiar with the task.

Experiment 1 (full feedback, $n=17$). On each trial in Exp. 1, two items were presented one after the other at fixation (0.5 s per item) with an inter-stimulus interval of 2–3 s (randomized). After the second item, Arabic digits ‘1’ and ‘2’ were displayed to the left and right of fixation (the positions were randomized across trials), and the participants were asked to choose the higher-valued item by pressing the corresponding arrow key (left or right) within 2 s. A written feedback message (‘great’ for correct responses, ‘incorrect’ for errors) was shown after each choice (neighbouring and non-neighbouring pairs). The items’ latent values in Exp. 1 were probabilistic (with a Gaussian distribution) and designed such that feedback was truthful on approximately 80% of neighbour trials (probabilistic feedback). Each participant performed 448 learning trials with all possible stimulus pairings ($n=56$) presented in each of eight consecutive blocks. Experiments 1 and 2 were conducted in lab, using Psychophysics Toolbox Version 3 (ref. ⁵⁸) running in MATLAB 2017a (MathWorks).

Experiment 2 (partial feedback, $n=31$). The design of Exp. 2 was nearly identical to that of Exp. 1, but choice feedback was given only after neighbouring pairs. After non-neighbouring pairs, a neutral ‘thank you’ message was displayed instead. Neighbouring pairs were presented more often (2.5 times as often as non-neighbouring pairs), resulting in 616 trials (presented in 8 blocks of 77). In Exp. 2, we additionally recorded EEG, and the participants performed a brief picture-viewing task before the experiment. These data were collected for the purpose of a different research question and are not reported here.

Experiment 3 (partial feedback, $n=48$). The basic design of Exp. 3 was identical to that of Exp. 2, except for the following changes. Both pair items were displayed simultaneously on the screen for 2.5 s, one to the left and the other to the right of a centred fixation cross. The participants were instructed to quickly select the

higher-valued item using the left or right arrow key. After neighbouring pairs, a feedback message ('win' or 'loss') was presented. After non-neighbouring pairs, no feedback message was shown. Experiments 3 and 4 were programmed in PsychoPy v.2.0.1.3 (ref. ³⁹) and conducted online (pavlovia.org), with intermittent attention checks.

Experiment 4 (partial feedback, deterministic, n = 49). The design of Exp. 4 was identical to that of Exp. 3, but feedback was always truthful (deterministic feedback). As learning expectedly proceeds faster with deterministic feedback, neighbouring pairs were presented only two times as often as non-neighbours, and we reduced the number of trials to 420 (presented in 6 blocks of 70 trials).

Models. Item-level learning. To model how observers update their value estimates about the winning item i and the losing item j after relational feedback, we assume a simple delta rule⁴⁰ (Model Q1):

$$Q_{t+1}(i) = Q_t(i) + \alpha [1 - Q_t(i)] \quad (1a)$$

$$Q_{t+1}(j) = Q_t(j) + \alpha [-1 - Q_t(j)] \quad (1b)$$

where Q_t is the estimated item value at time t , and α is the learning rate.

Transitive inference is enabled by a modified updating rule (similar to refs. ^{14,23}) based on the relative difference $d_t(i, j)$ between the value estimates for the winner i and the loser j in a pair:

$$d_t(i, j) = \eta [Q_t(i) - Q_t(j)] \quad (2)$$

where η is a scaling factor. Value updating is then moderated by the extent to which feedback is consistent (or inconsistent) with $d_t(i, j)$ (Model Q1^{*}):

$$Q_{t+1}(i) = Q_t(i) + \alpha [1 - d_t(i, j) - Q_t(i)] \quad (3a)$$

$$Q_{t+1}(j) = Q_t(j) + \alpha [-1 + d_t(i, j) - Q_t(j)] \quad (3b)$$

for the winning item i and the losing item j , respectively. Note that equation (1) is a special case of equation (3) when $\eta = 0$.

We can allow asymmetric updating of winners and losers by introducing separate learning rates, α^+ and α^- (Models Q2 and Q2^{*}):

$$Q_{t+1}(i) = Q_t(i) + \alpha^+ [1 - d_t(i, j) - Q_t(i)] \quad (4a)$$

$$Q_{t+1}(j) = Q_t(j) + \alpha^- [-1 + d_t(i, j) - Q_t(j)] \quad (4b)$$

where the winning item i is updated via α^+ , and the losing item j is updated via α^- .

To convert the value estimates from item-level learning into pairwise choice probabilities for any two items i and j , we use a logistic choice function to define the probability of choosing $i > j$ on the basis of the difference between the estimated item values:

$$CP_{item, t} = \frac{1}{1 + \exp(-(Q_t(i) - Q_t(j))/\tau_{item})} \quad (5)$$

where τ_{item} is the (inverse) temperature parameter controlling the level of decision noise in choices based on item-level learning.

Pair-level learning. For the partial-feedback scenario, we also define an alternative learning model (Model P) that learns pairwise preferences between neighbouring items (rather than the individual items' values). For each neighbouring pair $n (1\dots7)$, we can describe the preference between its members (for example, $p_{A>B}$) probabilistically in terms of a beta distribution:

$$p_h \sim Beta(U_n, L_n)$$

Following truthful feedback (for example, 'correct' when A > B was chosen), the upper value of the beta distribution is updated, increasing the preference in favour of the higher-ranking pair member:

$$U_{n,t+1} = U_{n,t} + \gamma \quad (6a)$$

whereas following untruthful feedback (only in experiments with probabilistic feedback; see Exps. 2 and 3), the lower value is updated, reducing the preference for the higher-ranking member:

$$L_{n,t+1} = L_{n,t} + \gamma \quad (6b)$$

with γ acting as a learning rate. We can thus define the learned neighbour preference at time t on the basis of the expectation of the beta distribution (Model P):

$$p_{n,t} = \frac{U_{n,t}}{U_{n,t} + L_{n,t}} \quad (7)$$

where $p_{n,t} = 0.5$ reflects indifference, and values of $p_{n,t}$ larger (or smaller) than 0.5 reflect a preference for the higher (or lower) ranking pair member. While this mechanism can learn the relations between neighbouring items under partial feedback, it fails to learn the relations between non-neighbouring items, for which there is no direct feedback signal. However, transitive inference of preferences between non-neighbouring items is possible through associative recall of those neighbour preferences that 'link' the two non-neighbour items in question. To allow for this possibility, we define the inferred preference between any two items i and j via the set M of intermediate neighbour preferences $p_{n,t}$ separating i and j (Model Pi):

$$p_{i>j,t} = \frac{\sum_{p_{n,t} \in M} (p_{n,t} - 0.5)}{|i - j|^{\lambda+1}} + 0.5 \quad (8)$$

where $|i - j|$ is the rank distance between the items' true values, and λ is a free parameter reflecting failure to retrieve linking pair preferences in the range $[0, \infty]$. If $\lambda = 0$, preferences between non-neighbours will be a lossless average of all intermediate neighbour preferences (that is, perfect memory). As λ grows, the preference between non-neighbours will shrink to indifference with increasing distance between j and i . In other words, this model performs perfect transitive inference if $\lambda = 0$ and no transitive inference as $\lambda \rightarrow \infty$. Note that for neighbour pairs (where $|i - j| = 1$), equation (8) is equivalent to equation (7).

We again use a logistic choice rule to define the probability of choosing item i over j on the basis of pair preference $p_{i>j,t}$ subject to decision noise τ_{pair} :

$$CP_{pair, t} = \frac{1}{1 + \exp(-p_{i>j,t}/\tau_{pair})} \quad (9)$$

From equations (6)–(9), we constructed alternative models incorporating basic pair-level learning (Model P, where λ is fixed at a large value) and pair-level transitive inference (Model Pi, where λ is a free parameter).

To combine item-level (equations (1)–(5)) and pair-level (equations (6)–(9)) learning, we assume that choices are triggered by whichever of the two models provides a stronger preference on a given trial. Choices are thus based on item-level learning (CP_{item}) if:

$$|CP_{item, t} - 0.5| > |CP_{pair, t} - 0.5| \quad (10a)$$

and are based on pair-level learning (CP_{pair}) if:

$$|CP_{item, t} - 0.5| < |CP_{pair, t} - 0.5| \quad (10b)$$

This effectively implements a mixture of item- and pair-level learning.

Model space. From equations (1)–(10), we constructed a nested model space (Extended Data Fig. 1c) with either one or two learning rates (1, symmetric; 2, asymmetric updating; equation (4)). One set of models allows for simple item-level RL only (Models Q1 and Q2) or additionally for item-level transitive inference (Models Q1^{*} and Q2^{*}, equations (1)–(5)). Alternative models (equations (6)–(9)) incorporated pair-level learning (Model P) and pair-level inference (Model Pi). Mixture models (equation (10)) combined item-level and pair-level learning (Q1^{*} + P, Q2^{*} + P, Q1^{*} + Pi and Q2^{*} + Pi). Technically, all models under study were derived from the most flexible model, Q2^{*} + Pi, with individual parameter restrictions (for example, $\gamma = 0$ yields Model Q2^{*}, or $\alpha^+ = \alpha^-$ yields symmetric updating).

Performance simulations. We simulated the performance of our item-level learning models (Q1, Q2, Q1^{*} and Q2^{*}) in tasks akin to those used in the human experiments, with full and partial feedback (Fig. 2 and Extended Data Figs. 1 and 2). The performance simulations were run in MATLAB R2020a (MathWorks). The models were initialized with flat priors about the item values (all $Q_t(i) = 0$ —that is, the first choice was always a random guess with $CP_1 = 0.5$). As in the human experiments, choice feedback was provided either for all pairs (full feedback) or only for neighbour pairs (partial feedback). We simulated model performance over a range of learning rates (α^+ and α^- , 0 to 0.1 in increments of 0.001). Relational difference-weighting (η) was set to either 0 (Models Q1 and Q2) or 8 (Models Q1^{*} and Q2^{*}), and decision noise (τ_{item}) was set to 0.2 and 0.04 (full and partial feedback), which resembles the noise levels estimated in our human observers in the respective experiments. Mean choice probabilities (for example, Fig. 2a, bottom) and performance levels (for example, Fig. 2b) were simulated using the same number of trials and replications (with a new trial sequence) as in the respective human experiments. Simulation results under partial feedback (Fig. 2e and Extended Data Figs. 2 and 3) were qualitatively identical when inspecting performance on non-neighbouring pairs only.

Parameter estimation and model comparison. Model parameters were estimated by minimizing the negative log-likelihood of the model given each observer's single-trial responses (from all trials in the experiment) across values of the model's free parameters (within bounds (lower;upper): $\alpha/\alpha^+/\alpha^-$ (0;0.2), η (0;10), τ_{item} (0;1), γ (0;1), λ (0;100), τ_{pair} (0;1), with a uniform prior). The best-fitting parameter

estimates are shown in Extended Data Fig. 4. Model fitting was performed in R (ref. ⁶¹). Minimization was performed using a differential evolution algorithm⁶² with 200 iterations. We then computed the BIC of each model for each participant and evaluated the models' probability of describing the majority of participants best (p_{pxp})⁶³. In Fig. 3e,f, we also provide a pseudo- R^2 computed as $R^2 = 1 - (\text{BIC}_{\text{model}}/\text{BIC}_{\text{null}})$, which quantifies goodness of fit relative to a null model of the data, with larger values indicating better fit (similar to ref. ⁶⁴). Model comparisons for Exp. 1 (full feedback) were restricted to item-level learning models, as the availability of direct feedback for every pairing would equate pair-level learning models (P and P_i) to homogenous learning of all pairs, obviating contributions from transitive inferences.

To quantify model-estimated asymmetry (Fig. 5), we computed an index of the normalized difference in learning rates, $A = (\alpha^+ - \alpha^-)/(|\alpha^+ + \alpha^-|)$, which ranges from -1 (updating of losers only) to 1 (updating of winners only), with $A=0$ indicating symmetric updating. For comparison between full- and partial-feedback experiments, we contrasted the absolute $|A|$ estimated from the winning model in Exps. 2–4 ($Q2^* + P$, see Fig. 3f) with that estimated from Model Q2 in Exp. 1.

Model and parameter recovery. To establish whether the individual models can be distinguished in model comparison, we simulated, for each participant and model, 100 experiment runs using the individuals' empirical parameter estimates under the respective model. We then fitted the generated datasets (binomial choice data) with each model and evaluated how often it provided the best fit (in terms of BIC). This way, we estimated the conditional probability that a model fits best given the true generative model ($p(\text{fit}|\text{gen})$). However, a metric more critical for evaluating our empirical results is $p(\text{gen}|\text{fit})$, which is the probability that the data was generated by a specific model, given that the model was observed as providing the best fit to the generated data⁶⁵. We compute this probability using Bayes's theorem, with a uniform prior over models ($p(\text{gen})$):

$$p(\text{gen}|\text{fit}) = \frac{p(\text{fit}|\text{gen})p(\text{gen})}{\sum_{\text{sim}=1}^{n\text{Models}} p(\text{fit}|\text{gen})_{\text{sim}}p(\text{gen})_{\text{sim}}}$$

To mimic the level of inference in our human data fitting, we examined mean $p(\text{fit}|\text{gen})$ and $p(\text{gen}|\text{fit})$ on the experiment level, on the basis of full simulations of all participants in Exp. 1 (full feedback) and Exp. 2 (partial feedback). Critically, under partial feedback (Exps. 2–4), all our models were robustly recovered with this approach (Extended Data Fig. 5).

Under full feedback (Exp. 1), human participant behaviour was best characterized by symmetric learning rates ($\alpha^+ \approx \alpha^-$), even when both learning rates were free parameters (Figs. 3e and 4). To test whether we could have detected asymmetric learning had it occurred in Exp. 1, we enforced asymmetry in the simulation by setting α^- to values near zero (by drawing from a rectified Gaussian with $\mu=0$ and $\sigma=0.01$). We likewise enforced difference-weighted updating ($\eta > 0$) when simulating Model Q2*, by setting η to similar levels as empirically observed in the partial-feedback experiments ($\mu=3$ and $\sigma=0.5$). With this, the model recovery for Exp. 1 successfully distinguished between symmetric (Q1 and Q1*) and asymmetric learning models (Q2 and Q2*; Extended Data Fig. 6). However, models with difference-weighted updating (Q1* and Q2*, equations (2) and (3)) were partly confused with Models Q1 and Q2. In other words, our empirical finding of Q1 as the winning model in Exp. 1 (Fig. 3e) does not rule out the possibility of Q1* as the generative process under full feedback.

To establish whether our inferences about model parameters (for example, Fig. 5) are valid, we simulated choices under partial feedback (Exp. 2) using our winning model ($Q2^* + P$). Choice datasets were simulated using each participant's empirical parameter estimates and iteratively varying each parameter over 20 evenly spaced values within the boundaries used in 'Parameter estimation and model comparison' (see above). We then fit the model to the simulated datasets and examined the correlations between generative and recovered parameters (Extended Data Figs. 7 and 8). All fitted parameters correlated most strongly with their generative counterparts (min 0.59, max 0.93), while correlations with other generative parameters were generally weaker (min -0.44 , max 0.43).

Statistical analyses. The behavioural and modelling results were analysed using non-parametric tests (two-sided), as detailed in the Results. In the case of multiple tests, the maximum P value (uncorrected) is reported in the main text, while the individual test results are detailed in Supplementary Tables 1–4.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data that support the findings of this study are available at <https://doi.org/10.5281/zenodo.5561411>.

Code availability

The experiment and analysis code is available at <https://doi.org/10.5281/zenodo.5561411>.

Received: 1 April 2021; Accepted: 25 November 2021;

Published online: 31 January 2022

References

- Bryant, P. E. & Trabasso, T. Transitive inferences and memory in young children. *Nature* **232**, 456–458 (1971).
- Burt, C. Experimental tests of general intelligence. *Br. J. Psychol.* **3**, 94–177 (1909).
- Jensen, G., Muñoz, F., Alkan, Y., Ferrera, V. P. & Terrace, H. S. Implicit value updating explains transitive inference performance: the betasort model. *PLoS Comput. Biol.* **11**, e1004523 (2015).
- Piaget, J. *Judgment and Reasoning in the Child* (Harcourt, Brace, 1928); <https://doi.org/10.4324/9780203207260>
- Vasconcelos, M. Transitive inference in non-human animals: an empirical and theoretical analysis. *Behav. Process.* **78**, 313–334 (2008).
- Boysen, S. T., Berntson, G. G., Shreyer, T. A. & Quigley, K. S. Processing of ordinality and transitivity by chimpanzees (*Pan troglodytes*). *J. Comp. Psychol.* **107**, 208–215 (1993).
- Gillan, D. J. Reasoning in the chimpanzee: II. Transitive inference. *J. Exp. Psychol. Anim. Behav. Process.* **7**, 150–164 (1981).
- McGonigle, B. O. & Chalmers, M. Are monkeys logical? *Nature* **267**, 694–696 (1977).
- Davis, H. Transitive inference in rats (*Rattus norvegicus*). *J. Comp. Psychol.* **106**, 342–349 (1992).
- Bond, A. B., Kamil, A. C. & Balda, R. P. Social complexity and transitive inference in corvids. *Anim. Behav.* **65**, 479–487 (2003).
- Lazareva, O. F. & Wasserman, E. A. Transitive inference in pigeons: measuring the associative values of Stimuli B and D. *Behav. Process.* **89**, 244–255 (2012).
- Wynne, C. D. L. Pigeon transitive inference: tests of simple accounts of a complex performance. *Behav. Process.* **39**, 95–112 (1997).
- Delius, J. D. & Siemann, M. Transitive responding in animals and humans: exaptation rather than adaptation? *Behav. Process.* **42**, 107–137 (1998).
- Wynne, C. D. L. Reinforcement accounts for transitive inference performance. *Anim. Learn. Behav.* **23**, 207–217 (1995).
- Dusek, J. A. & Eichenbaum, H. The hippocampus and memory for orderly stimulus relations. *Proc. Natl Acad. Sci. USA* **94**, 7109–7114 (1997).
- Garvert, M. M., Dolan, R. J. & Behrens, T. E. A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *eLife* **6**, e17086 (2017).
- Kumaran, D. & McClelland, J. L. Generalization through the recurrent interaction of episodic memories: a model of the hippocampal system. *Psychol. Rev.* **119**, 573–616 (2012).
- Smith, C. & Squire, L. R. Declarative memory, awareness, and transitive inference. *J. Neurosci.* **25**, 10138–10146 (2005).
- Frank, M. J., Rudy, J. W., Levy, W. B. & O'Reilly, R. C. When logic fails: implicit transitive inference in humans. *Mem. Cogn.* **33**, 742–750 (2005).
- Hamilton, J. M. E. & Sanford, A. J. The symbolic distance effect for alphabetic order judgements: a subjective report and reaction time analysis. *Q. J. Exp. Psychol.* **30**, 33–41 (1978).
- von Fersen, L., Wynne, C. D., Delius, J. D. & Staddon, J. E. Transitive inference formation in pigeons. *J. Exp. Psychol. Anim. Behav. Process.* **17**, 334–341 (1991).
- Kumaran, D., Banino, A., Blundell, C., Hassabis, D. & Dayan, P. Computations underlying social hierarchy learning: distinct neural mechanisms for updating and representing self-relevant information. *Neuron* **92**, 1135–1147 (2016).
- Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T. & Hutchison, K. E. Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proc. Natl Acad. Sci. USA* **104**, 16311–16316 (2007).
- Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S. & Palminteri, S. Behavioural and neural characterization of optimistic reinforcement learning. *Nat. Hum. Behav.* **1**, 0067 (2017).
- Palminteri, S., Khamassi, M., Joffily, M. & Coricelli, G. Contextual modulation of value signals in reward and punishment learning. *Nat. Commun.* **6**, 8096 (2015).
- van den Bos, W., Cohen, M. X., Kahnt, T. & Crone, E. A. Striatum–medial prefrontal cortex connectivity predicts developmental changes in reinforcement learning. *Cereb. Cortex* **22**, 1247–1255 (2012).
- Lefebvre, G., Summerfield, C. & Bogacz, R. A normative account of confirmatory biases during reinforcement learning. *Neural Comput.* https://doi.org/10.1162/neco_a_01455 (2021).
- Palminteri, S., Lefebvre, G., Kilford, E. J. & Blakemore, S.-J. Confirmation bias in human reinforcement learning: evidence from counterfactual feedback processing. *PLoS Comput. Biol.* **13**, e1005684 (2017).
- Weber, E. H. *De Pulsu, Resorptione, Auditu et Tactu: Annotationes Anatomicae et Physiologicae...* (C.F. Koehler, 1834).
- Cheyette, S. J. & Piantadosi, S. T. A unified account of numerosity perception. *Nat. Hum. Behav.* **4**, 1265–1272 (2020).

31. Nieder, A. & Miller, E. K. Coding of cognitive magnitude: compressed scaling of numerical information in the primate prefrontal cortex. *Neuron* **37**, 149–157 (2003).
32. Kahneman, D. & Tversky, A. Prospect theory: an analysis of decision under risk. *Econometrica* **47**, 263–291 (1979).
33. Eichenbaum, H. Hippocampus, cognitive processes and neural representations that underlie declarative memory. *Neuron* **44**, 109–120 (2004).
34. O'Reilly, R. C. & Rudy, J. W. Conjunctive representations in learning and memory: principles of cortical and hippocampal function. *Psychol. Rev.* **108**, 311–345 (2001).
35. Whittington, J. C. R. & Bogacz, R. Theories of error back-propagation in the brain. *Trends Cogn. Sci.* **23**, 235–250 (2019).
36. Anderson, J. R. *The Architecture of Cognition* (Harvard Univ. Press, 1983).
37. Jensen, G., Terrace, H. S. & Ferrera, V. P. Discovering implied serial order through model-free and model-based learning. *Front. Neurosci.* **13**, 878 (2019).
38. Dehaene, S. The neural basis of the Weber–Fechner law: a logarithmic mental number line. *Trends Cogn. Sci.* **7**, 145–147 (2003).
39. Pardo-Vazquez, J. L. et al. The mechanistic foundation of Weber's law. *Nat. Neurosci.* **22**, 1493–1502 (2019).
40. Bhui, R. & Gershman, S. J. Decision by sampling implements efficient coding of psychoeconomic functions. *Psychol. Rev.* **125**, 985–1001 (2018).
41. Stewart, N., Chater, N. & Brown, G. D. A. Decision by sampling. *Cogn. Psychol.* **53**, 1–26 (2006).
42. Summerfield, C. & Li, V. Perceptual suboptimality: bug or feature? *Behav. Brain Sci.* **41**, e245 (2018).
43. Gigerenzer, G. & Brighton, H. Homo heurtisticus: why biased minds make better inferences. *Top. Cogn. Sci.* **1**, 107–143 (2009).
44. Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D. & Meder, B. Generalization guides human exploration in vast decision spaces. *Nat. Hum. Behav.* **2**, 915–924 (2018).
45. Juechems, K., Balaguer, J., Spitzer, B. & Summerfield, C. Optimal utility and probability functions for agents with finite computational precision. *Proc. Natl Acad. Sci. USA* **118**, e2002232118 (2021).
46. Li, V., Herce Castañón, S., Solomon, J. A., Vandormael, H. & Summerfield, C. Robust averaging protects decisions from noise in neural computations. *PLoS Comput. Biol.* **13**, e1005723 (2017).
47. Luyckx, F., Spitzer, B., Blangero, A., Tssetsos, K. & Summerfield, C. Selective integration during sequential sampling in posterior neural signals. *Cereb. Cortex* **30**, 4454–4464 (2020).
48. Spitzer, B., Waschke, L. & Summerfield, C. Selective overweighting of larger magnitudes during noisy numerical comparison. *Nat. Hum. Behav.* **1**, 0145 (2017).
49. Tssetsos, K. et al. Economic irrationality is optimal during noisy decision making. *Proc. Natl Acad. Sci. USA* **113**, 3102–3107 (2016).
50. Eichenbaum, H. A cortical–hippocampal system for declarative memory. *Nat. Rev. Neurosci.* **1**, 41–50 (2000).
51. De Soto, C. B., London, M. & Handel, S. Social reasoning and spatial paralogic. *J. Personal. Soc. Psychol.* **2**, 513–521 (1965).
52. Whittington, J. C. R. et al. The Tolman–Eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. *Cell* **183**, 1249–1263.e23 (2020).
53. Frank, M. J., Rudy, J. W. & O'Reilly, R. C. Transitivity, flexibility, conjunctive representations, and the hippocampus. II. A computational analysis. *Hippocampus* **13**, 341–354 (2003).
54. Van Elzakker, M., O'Reilly, R. C. & Rudy, J. W. Transitivity, flexibility, conjunctive representations, and the hippocampus. I. An empirical analysis. *Hippocampus* **13**, 334–340 (2003).
55. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**, 1204–1215 (2011).
56. Hayden, B. Y. & Niv, Y. The case against economic values in the orbitofrontal cortex (or anywhere else in the brain). *Behav. Neurosci.* **135**, 192–201 (2021).
57. Brodeur, M. B., Guérard, K. & Bouras, M. Bank of Standardized Stimuli (BOSS) Phase II: 930 new normative photos. *PLoS ONE* **9**, e106953 (2014).
58. Brainard, D. H. The Psychophysics Toolbox. *Spat. Vis.* **10**, 433–436 (1997).
59. Peirce, J. et al. PsychoPy2: experiments in behavior made easy. *Behav. Res.* **51**, 195–203 (2019).
60. Rescorla, R. A. & Wagner, A. R. in *Classical Conditioning II: Current Theory and Research* 64–99 (Appleton-Century-Crofts, 1971).
61. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2020); <https://www.r-project.org/>
62. Mullen, K. M., Ardia, D., Gil, D. L., Windover, D. & Cline, J. DEoptim: an R package for global optimization by differential evolution. *J. Stat. Softw.* **40**, 1–26 (2011).
63. Rigoux, L., Stephan, K. E., Friston, K. J. & Daunizeau, J. Bayesian model selection for group studies—revisited. *NeuroImage* **84**, 971–985 (2014).
64. McFadden, D. *Conditional Logit Analysis of Qualitative Choice Behavior* (Institute of Urban and Regional Development, Univ. of California, 1973).
65. Wilson, R. C. & Collins, A. G. Ten simple rules for the computational modeling of behavioral data. *eLife* **8**, e49547 (2019).

Acknowledgements

We thank K. Tssetsos, S. Nelli, C. Summerfield and N. Schuck for helpful feedback and discussion, and S. Appelhoff and J. Wässcher for technical support. This work was supported by a Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) research grant to B.S. (DFG SP-1510/6-1). C.M.W. is supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center FKZ: 01IS18039A, and funded by the DFG under Germany's Excellence Strategy—EXC 2064/1–390727645. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

J.L.-D. and B.S. designed the experiments. J.L.-D., C.W. and I.P. performed the experiments. B.S. designed the modelling approach. S.C. and B.S. performed the simulations and analyses with contributions from C.M.W. and J.L.-D. S.C. and B.S. visualized the simulations and results. B.S., C.M.W., S.C. and J.L.-D. wrote the paper.

Funding

Open access funding provided by Max Planck Society.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41562-021-01263-w>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41562-021-01263-w>.

Correspondence and requests for materials should be addressed to Bernhard Spitzer.

Peer review information *Nature Human Behaviour* thanks Maël Lebreton and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

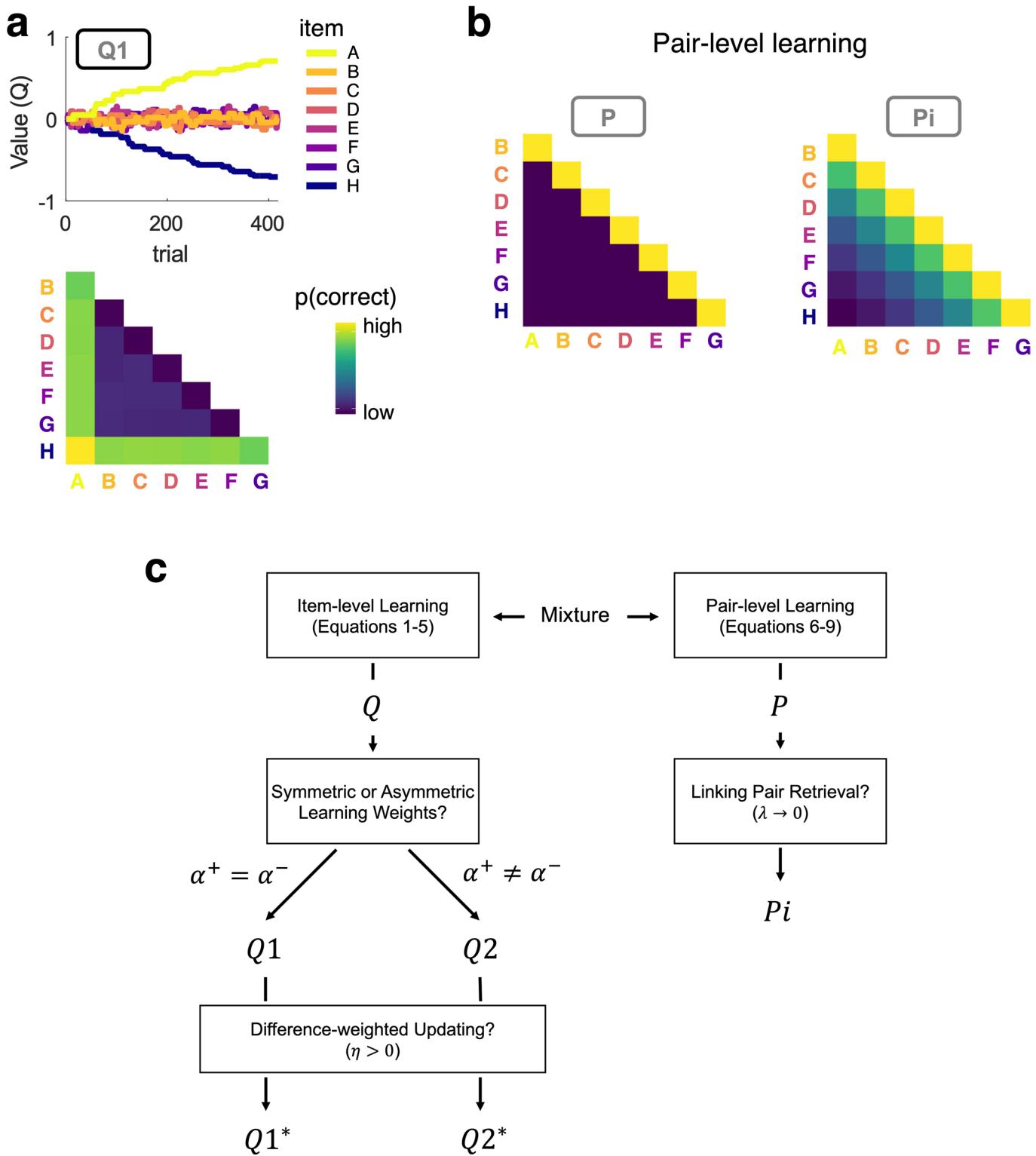
Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

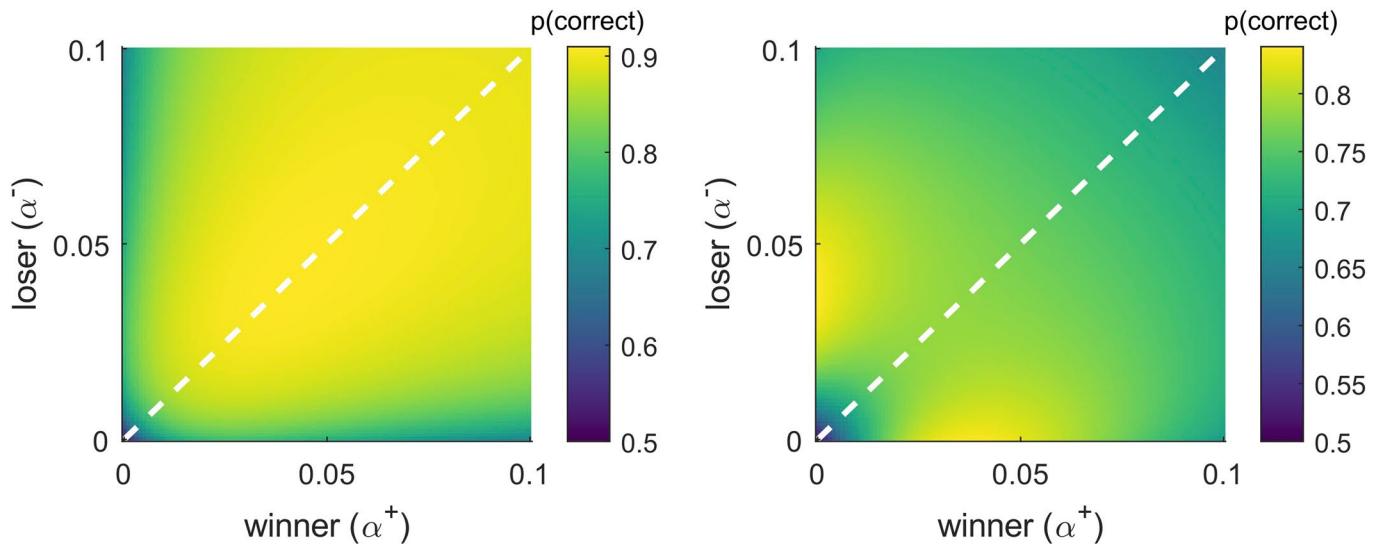


Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

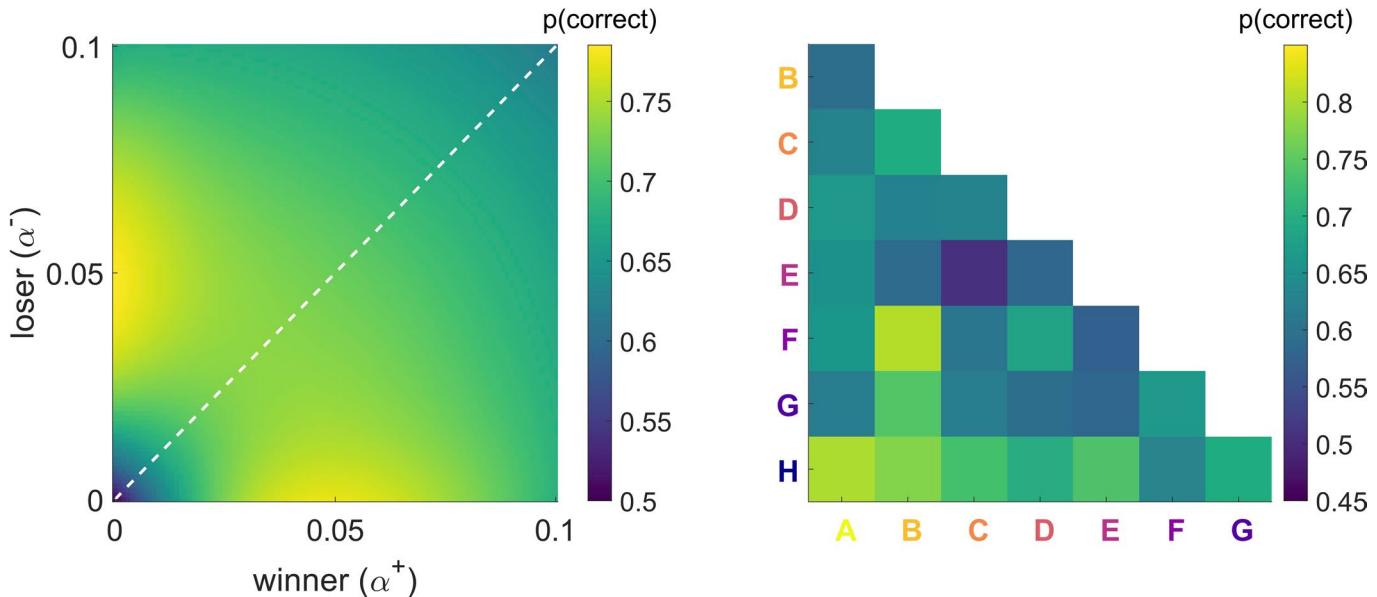
© The Author(s) 2022



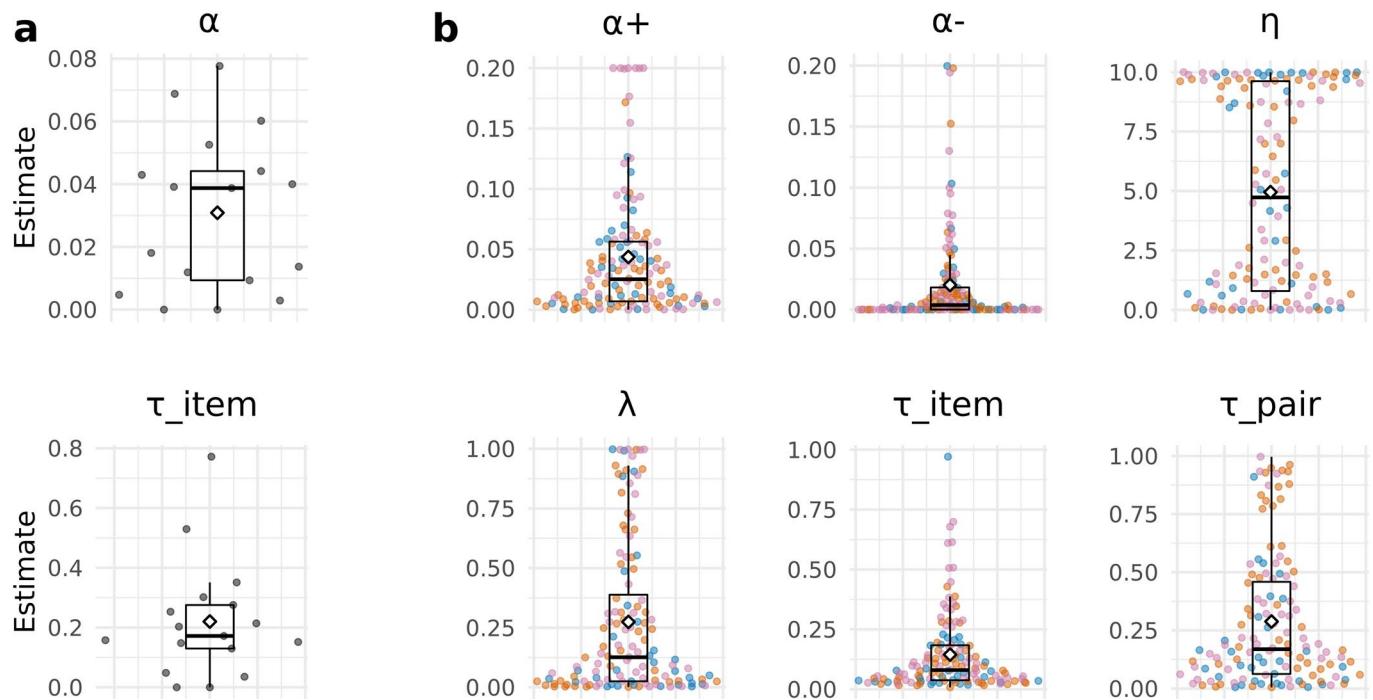
Extended Data Fig. 1 | Additional model simulations and illustration of the model space. **a**, Simulation of model Q1 under partial feedback. Same conventions as in Fig. 2. The simple Q-learning models (Q1/Q2) can only learn about the extreme items (here, A and H) under partial feedback. **b**, Choice matrices predicted by pair-level learning without (left, model P) or with associative recall of ‘linking’ pair relationships (right, model Pi). Choice behaviour was simulated with a pair-level learning rate $\gamma=1$. Associative recall in model Pi (right) was enabled by additionally lowering parameter λ to 1 (see Methods for details). **c**, Schematic overview of the model space. See Methods, Models.



Extended Data Fig. 2 | Performance simulations with probabilistic choice outcomes. Same conventions as in Fig. 2. Left, full feedback (for all pairs; cf. Figure 2b). Right, partial feedback (only for non-neighbouring pairs; cf. Figure 2e). Optimal learning under partial feedback is characterized by asymmetric updating ($\alpha^+ \neq \alpha^-$), just as was observed with deterministic outcomes (cf. Figure 2e).

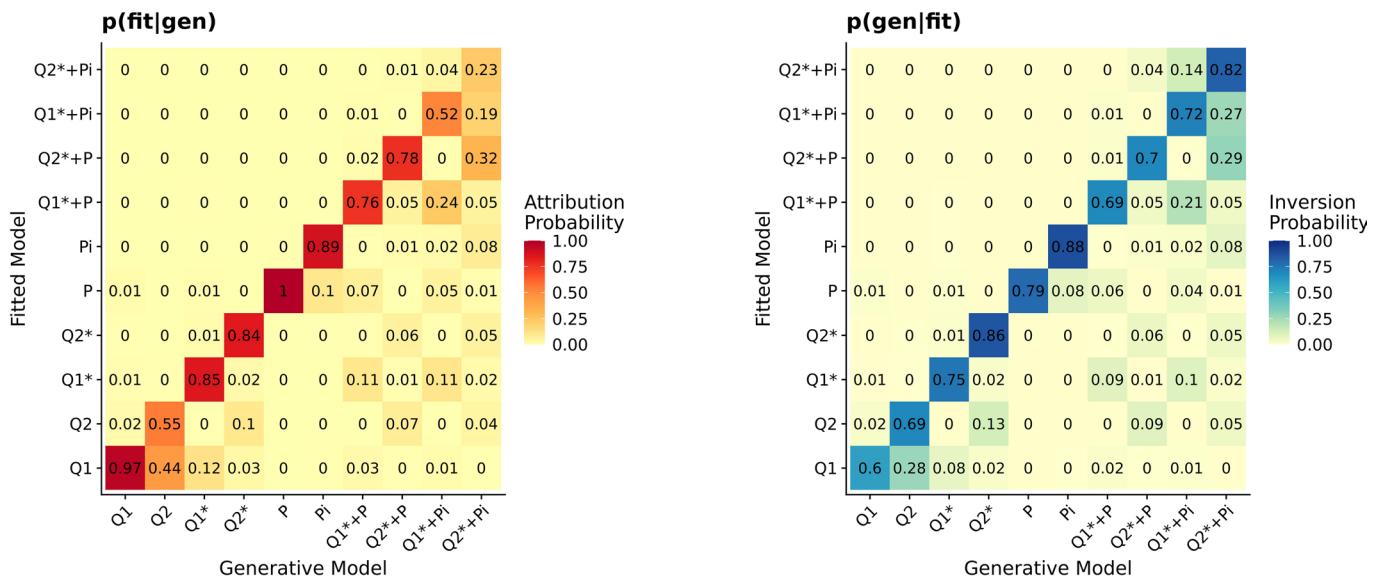


Extended Data Fig. 3 | Pilot experiment with partial feedback (n=11). The design was identical to Exp. 2, except that all item pairs (neighbours and non-neighbours) were presented equally frequently (like in Exp. 1). *Left*, Performance simulation shows a similar benefit of asymmetric updating as we observed in simulation of Exp. 2-4 (where neighbouring pairs were presented more frequently, cf. Figure 2e and S2, *right*). *Right*, Mean proportions of correct choices in the pilot experiment. The overall learning level was relatively low, with n=9 (of 20) pilot participants not meeting our inclusion threshold for above-chance performance (cf. *Methods: Participants*). The descriptive choice data of the remaining 11 pilot participants (shown in *right*) indicate a similar learning asymmetry as we observed in our main experiments.

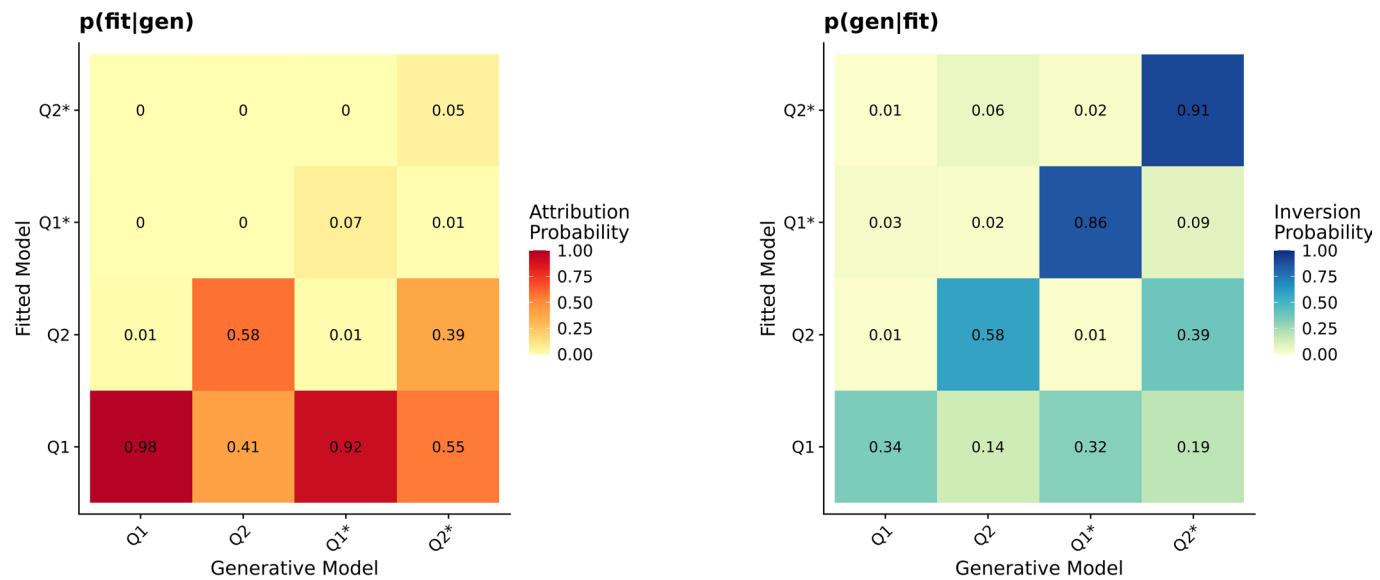


Extended Data Fig. 4 | Individual parameter estimates for the winning model in each experiment (cf. Figure 3). a, Experiment 1 (model Q1).

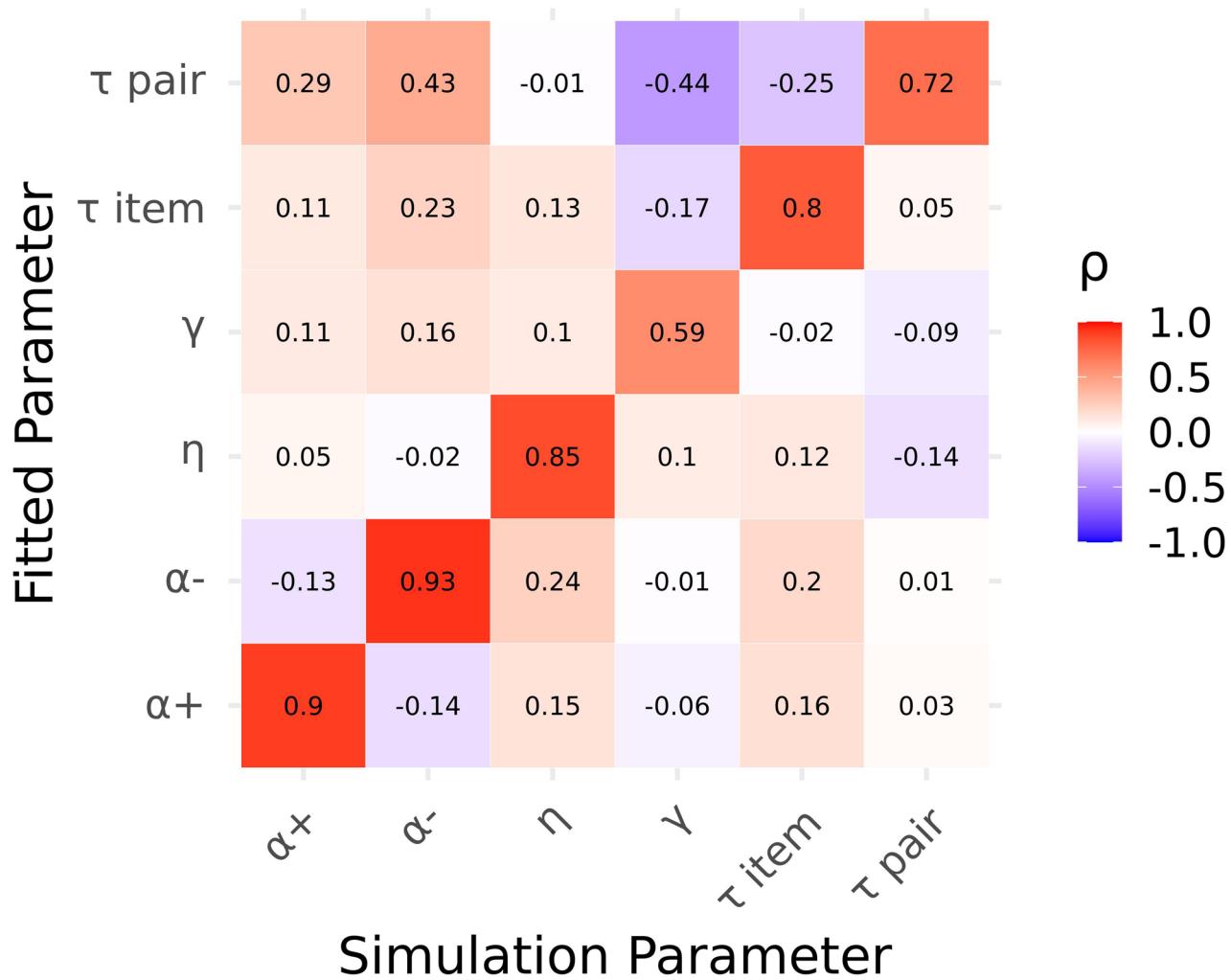
b, Experiments 2–4 (model Q2*+P). Coloured dots show individual participant estimates, where blue, orange and pink colours refer to Experiment 2, 3 and 4, respectively. The diamond shape denotes the mean across experiments. While the estimates of η showed large variability with many values close to the upper bound, we observed no improvement in model fit when increasing the upper bound further.



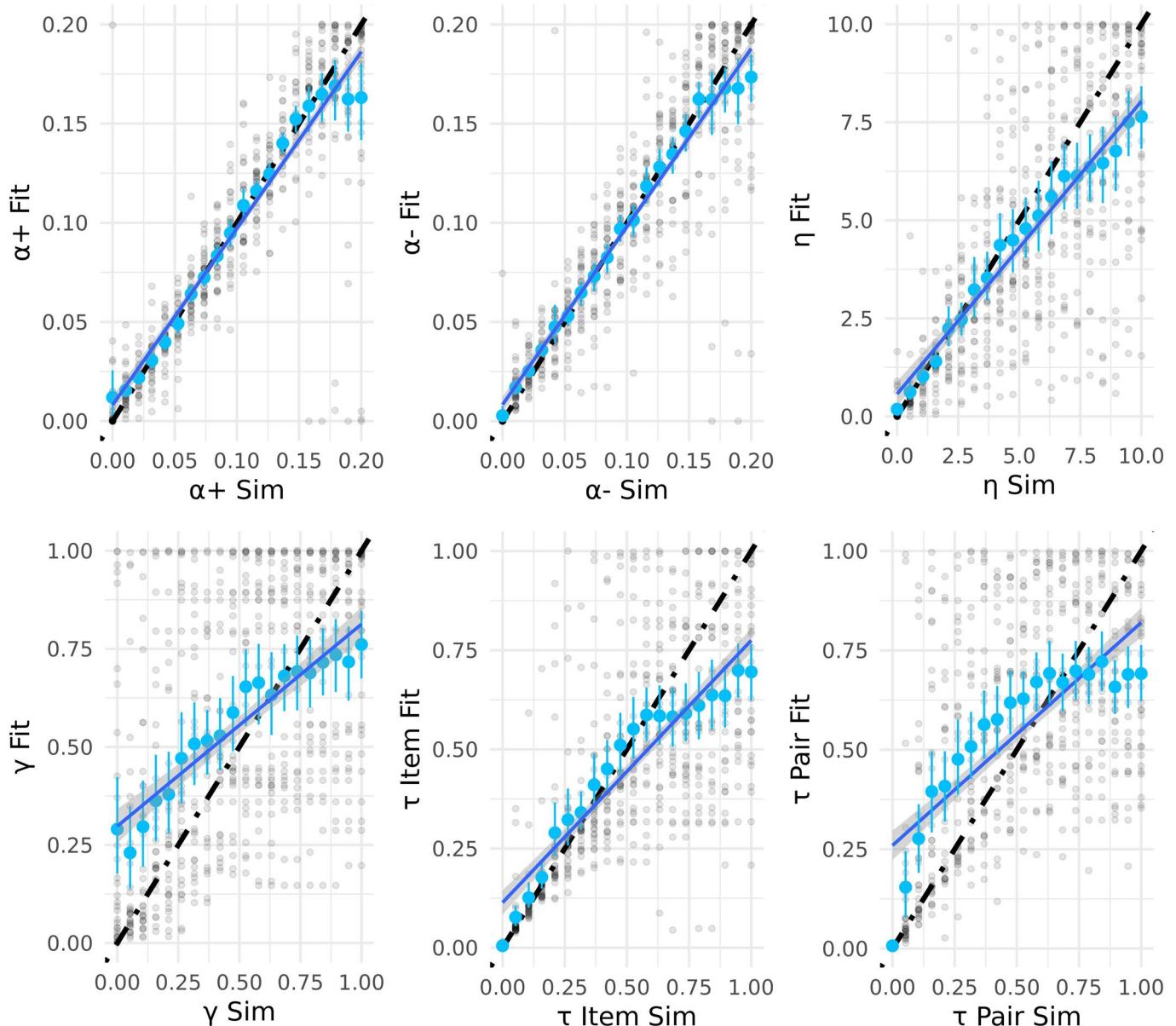
Extended Data Fig. 5 | Model recovery results, partial feedback (cf. Figure 3f). The models were generally well distinguished both in $p(\text{fit}|\text{gen})$ and in $p(\text{gen}|\text{fit})$. Of particular importance, our best asymmetric models (Q^2* and Q^2*+P , see *Results*) were well distinguished from their symmetric counterparts (Q^1* and Q^1*+P), with confusion rates no higher than 5%. Columns (left) and rows (right) may not always sum to 1 due to rounding of cell entries.



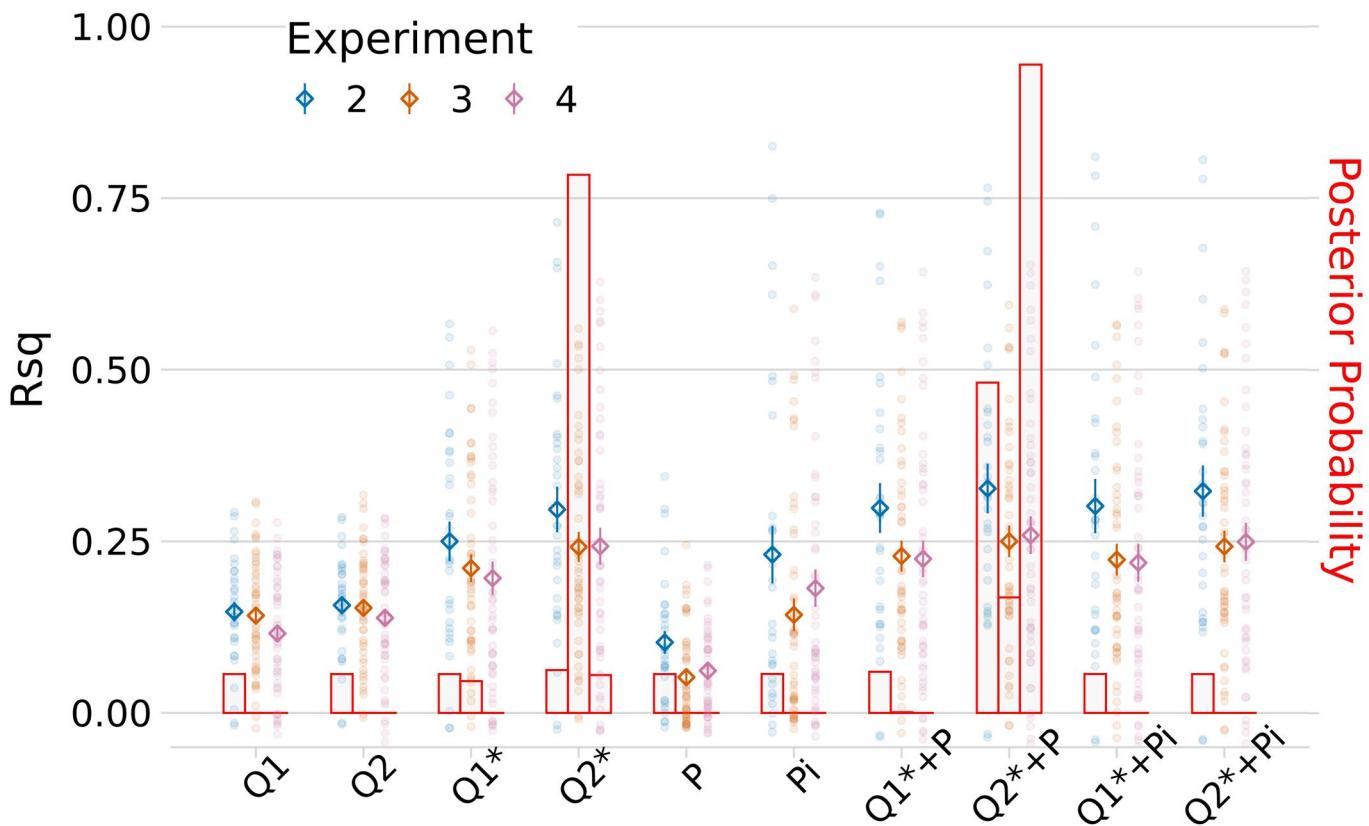
Extended Data Fig. 6 | Model recovery results for Exp. 1 with full feedback (cf. Figure 3e). Simple Q-learning (models Q1/Q2) could not be confidently distinguished from models Q1*/Q2*. However, symmetric (Q1/Q1*) and asymmetric learning (Q2/Q2*) were distinguished relatively well. See *Methods: Model- and parameter recovery* for details.



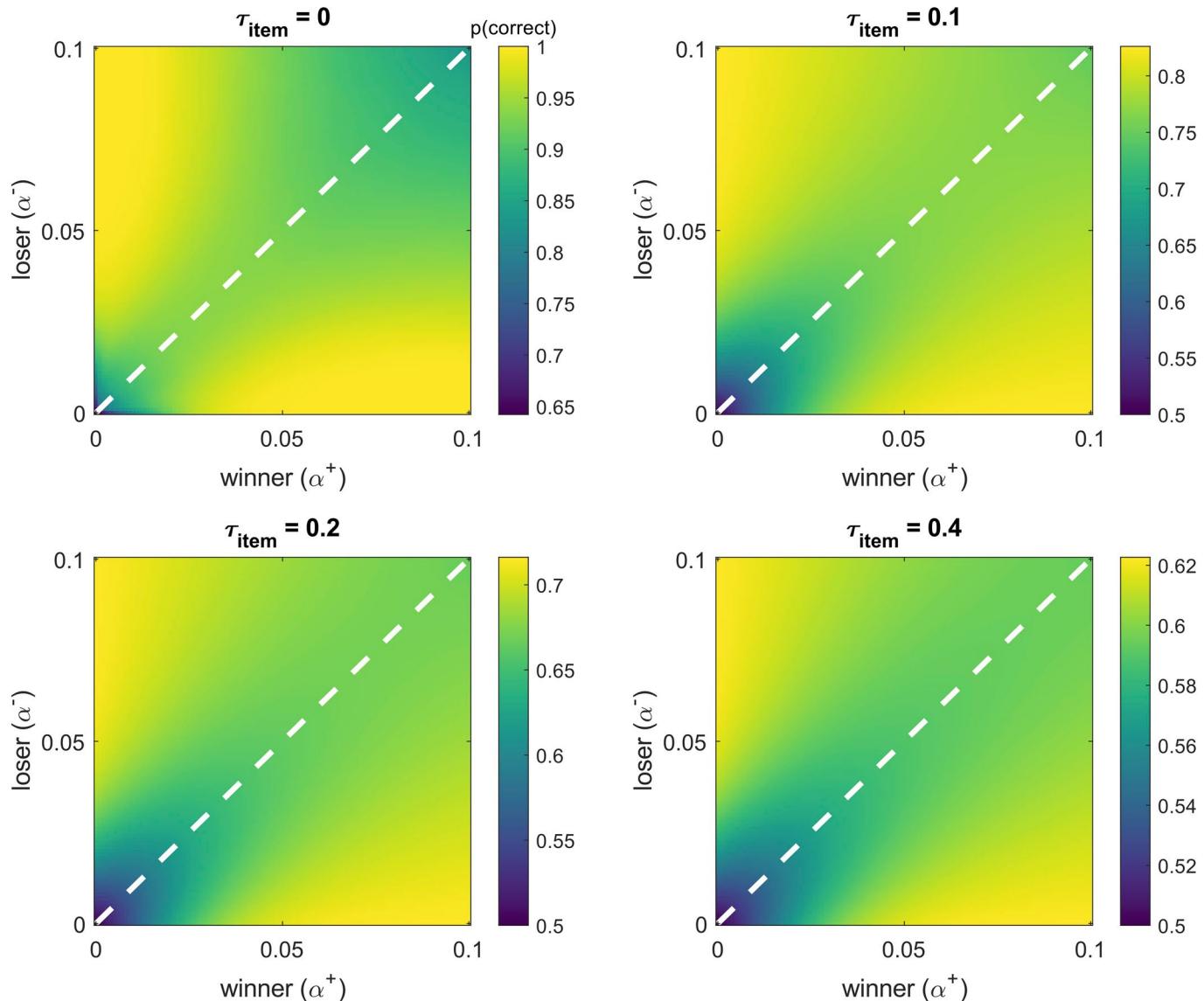
Extended Data Fig. 7 | Parameter recovery results under partial feedback for our best-fitting model (Q2*+P). All fitted parameters correlate most strongly with their generative counterparts (diagonal) while correlations with other generative parameters (off-diagonal) are generally weaker.



Extended Data Fig. 8 | Detailed parameter recovery results for the individual parameters. The parameter values used to simulate choice data are plotted on the x-axes and the parameter estimates obtained from fitting the model to the simulated data are plotted on the y-axes. Light blue: mean recovered parameter values with bootstrapped 95% confidence intervals. Dark blue line shows linear fit. Results from individual recovery runs are shown as half-transparent black dots.



Extended Data Fig. 9 | Model comparison (Exp. 2-4) analogous to Fig. 3f, but additionally allowing for differential learning from confirmatory vs. disconfirmatory choice feedback. Here, all models included an additional parameter $\omega \in (0;10)$, by which belief-confirming learning rates were modelled as: $\alpha_{\text{conf}} = \alpha * \omega$. Same conventions as in Fig. 3f. Markers show model fits using a pseudo-R-squared (Rsq, left y-axis; diamonds and error bars show mean \pm s.e.m., dots show individual participants). Overlaid red bar graphs indicate each model's probability of describing the majority of subjects best (right y-axis, ppx: protected exceedance probability). While the extra parameter ω led to general improvements in fit (note overall higher Rsq compared to Fig. 3f), the model comparison result with respect to winner/loser asymmetries was identical, both in terms of Rsq/BIC and ppx. The estimates of parameter ω in the winning models were larger than 1 (mean=4.72, SE=0.27, $p < 0.001$, $r=0.83$, Wilcoxon signed-rank test against 1, collapsed across Exp. 2-4), indicating an overall bias towards confirmatory feedback.



Extended Data Fig. 10 | Performance simulations under partial feedback (analogous to Fig. 2e) for different levels of decision noise (τ_{item}). Asymmetric learning is beneficial regardless of decision noise level and accordingly, across a wide range of overall performance levels. Note that different colour scales for each panel are used to increase interpretability (see colour bars). Simulations with probabilistic outcomes yielded a qualitatively very similar pattern.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection Exp. 1-2: Matlab R2017a (MathWorks), Psychophysics Toolbox Version 3 (Brainard, 1997); Exp. 3-4: Psychopy (<https://www.psychopy.org/>)
builder/builder.html); custom code: <https://doi.org/10.5281/zenodo.5561411>

Data analysis R (<http://www.R-project.org/>), Matlab R2020a (MathWorks); custom code: <https://doi.org/10.5281/zenodo.5561411>

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data that support the findings of this study are available at: <https://doi.org/10.5281/zenodo.5561411>

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Quantitative experimental (four experiments) and computational modeling /simulation
Research sample	Participants in Exp. 1 and 2 were healthy young adults recruited from a participant pool at the Max Planck Institute for Human Development in Berlin, Germany. Of these, n=20 participated in Experiment 1, (13 female, 7 male, mean age 27.15 ± 3.91 years) and n=35 participated in Exp. 2 (14 female, 27 ± 3.80 years). Participants in Exp. 3 and 4 were healthy young adults recruited online via Prolific Academic (https://www.prolific.co) with n=76 completing Exp. 3 (23 female, 24.73 ± 5.40 years) and n=60 completing Exp. 4 (23 female; 25.92 ± 4.54 years). Informed consent was obtained by all participants. The online samples were restricted to English speaking participants.
Sampling strategy	Participants were assigned to the different experiments by opportunity sampling. A pilot experiment with partial feedback (cf. supplementary information) suggested that learning asymmetries could be detected in a sample of n=11 participants ($p<0.05$, comparison of model BICs), which was substantially exceeded in all experiments (1-4). Sampling sizes in the online experiments (3-4) were chosen to be larger than in the in-lab experiments (1-2) due to anticipated higher drop-out rates in online testing.
Data collection	Experiments 1 and 2 were conducted in lab, using Psychophysics Toolbox Version 3 (Brainard, 1997) running in MATLAB 2017a. In Exp. 2, we additionally recorded electroencephalography for the purpose of a different research question. The experimenters present during in-lab testing were blind to the object-value associations. Experiments 3 and 4 were conducted online (on https://www.pavlovia.org), using Qualtrics (https://www.qualtrics.com) for statement of consent and basic demographics.
Timing	Start and finish of data collection: Exp. 1 from 11th January 2019 to 15th January 2019 Exp. 2 from 13th March 2019 to 5th July 2019 (n=28). A second cohort (n=7) was collected for the purpose of EEG recordings from 9th March 2020 to 4th September 2020 (with forced breaks due to COVID-19 pandemic). Exp. 3 from 1st July 2020 to 4th July 2020. Exp. 4 from 30th September 2020 to 1st October 2020.
Data exclusions	The data from participants who did not reach above-chance learning levels were excluded from analysis. The threshold for inclusion was set to 60% correct judgments in the last two blocks of the experiment, which corresponds to a binomial test probability of $p<0.01$ (uncorrected) compared to chance-level (50%). After exclusion, n=17 (Exp. 1), n=31 (Exp. 2), n=48 (Exp. 3) and n=49 (Exp. 4) participants remained for analysis.
Non-participation	No participants dropped out or declined participation in Exp. 1-2. In Exp. 3-4 (online studies), n=12 (Exp. 3) and n=8 (Exp.4) individuals signed up but did not complete the experiment.
Randomization	Participants were randomly assigned to the individual experiments (opportunity sampling). The association between stimuli and its value structure was pseudo-randomly assigned to the pictures for each participant. Stimulus pairings (8 neighbors and 20 non-neighbors) were randomly intermixed across trials, with randomized ordering of the elements in a pair (e.g., A-B or B-A).

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

- | | |
|-------------------------------------|---|
| n/a | Involved in the study |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Antibodies |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Eukaryotic cell lines |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Palaeontology and archaeology |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Animals and other organisms |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> Human research participants |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Clinical data |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Dual use research of concern |

Methods

- | | |
|-------------------------------------|---|
| n/a | Involved in the study |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> ChIP-seq |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Flow cytometry |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> MRI-based neuroimaging |

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	Young healthy adults.
Recruitment	In Exp. 1 and 2, participants were recruited from a participant pool at the Max Planck Institute for Human Development. Participants in Exp. 3 and 4 were recruited online via Prolific Academic (www.prolific.co).
Ethics oversight	All experiments were approved by the ethics committee of the Max Planck Institute for Human Development and conducted in accordance with the Human Subjects Guidelines of the Declaration of Helsinki.

Note that full information on the approval of the study protocol must also be provided in the manuscript.