Supporting Information Appendix

Social threat learning transfers to decision making in humans

Björn Lindström[1,2,3], Armita Golkar[3,4], Simon Jangard[3], Philippe N. Tobler[2], Andreas Olsson[3]

1. Department of Social Psychology, University of Amsterdam, The Netherlands
2. Laboratory for Social and Neural Systems Research, Department of Economics, University of Zürich, Switzerland
3. Section for Psychology, Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden
4. Department of Clinical Psychology, University of Amsterdam, The Netherlands

www.pnas.org/cgi/doi/10.1073/pnas.1810180116

Method

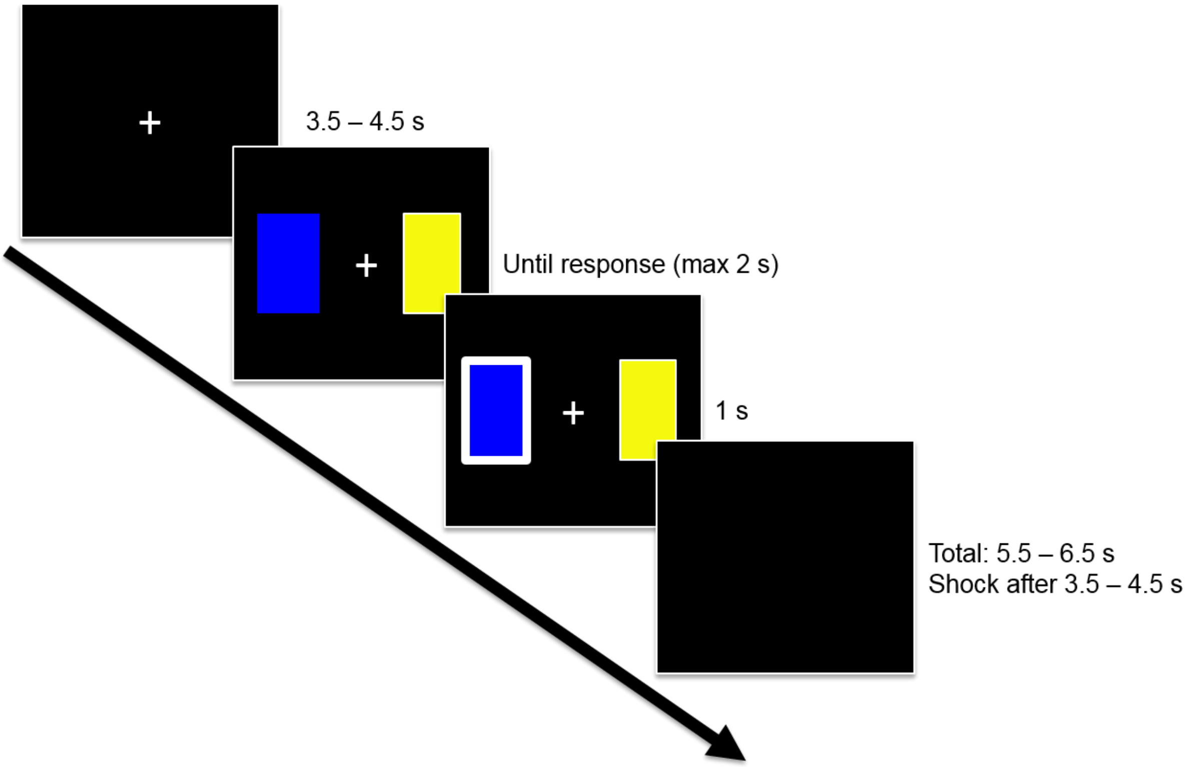

Figure S1. Trial overview for all experiments. The figure shows the trial timeline, which was the identical for all three experiments (and all control experiments). Variable durations were drawn from a uniform distribution.

Materials and stimuli

The experiments were conducted in a sound attenuated experimental chamber on a PC connected to a 19” CRT monitor. The aversive reinforcement was a monopolar 100 ms DC-pulse electric stimulation (STM200; Biopac Systems Inc, www.biopac.com) applied to the participant’s non-dominant forearm. In all experiments, the electric stimulator was attached before the conditioning block. The intensity of the electric shock stimulation was adjusted individually for each participant in a thresholding procedure, based on the standard criterion “unpleasant but not painful”. In Experiments 2 & 3, the shock thresholding procedure was conducted after the conditioning block (see below), while in Experiment 1 (and Experiment

3C), the shock thresholding procedure occurred by necessity before conditioning. The CS+ color was counter balanced across participants.

Skin conductance during the conditioning block was measured by Ag-AgCL electrodes attached to the distal phalanges of the index and middle finger of the left hand. The physiological signals were amplified and filtered between 0.05 and 5Hz. Phasic skin conductance responses (SCRs) were scored as an increase in skin conductance within 0.5 to 4.5 sec after stimulus onset (1). Responses below .02 µS were scored as zero and the raw SCR scores were z-transformed prior to analysis.

Additional information about Task and procedure

In Experiment 1, the participants were exposed to a standard protocol for Pavlovian (classical) threat/fear conditioning, in which they received mild electric shocks (US) to one (CS+, 4 out of 6 presentations) but not the other (CS-) of two stimuli (CS identity was counter-balanced across subjects in all experiments).

In Experiment 2, the participants were exposed to a standard protocol for observational threat/fear learning (1), in which they viewed a video of an individual receiving mild electric shocks (US) to one (CS+, 4 out of 6 presentations) but not the other (CS-) of two stimuli.

Experiment 3 used a modified procedure for instructed threat/fear conditioning (2), where the experimenter first instructed the participant that they could receive shocks to one (CS+), but never the other (CS-) cue. In addition, the same instruction was presented for 1 s before every cue presentation during the conditioning block (12 presentations). In order to avoid extinction and maintain the instructed threat expectations in the absence of shocks during the conditioning block, which is crucial for the comparison with observational threat learning, the participants never observed the CS cues (see below "Experiment 3B" for a control experiment where the CS cues were displayed during the conditioning block). Instead, two control cues were displayed (red and green triangles, 6 times each) during the conditioning block. This design thereby allowed the instructed threat expectation for the CS+ cue to be maintained in the absence of shocks.

During decision-making, the location of the stimuli varied randomly between the left and right position across trials to prevent spatial selection strategies. After observational threat learning, the participants in Experiment 2 were asked to rate their experience of watching the Demonstrator receiving shocks.

Power calculations and statistical analysis

Sample size for Experiment 1 was determined using a power analysis, based on the effect size (*Cohen's D* = .93) of the Transfer phase in Experiment 4 of ref. (3), which had a similar design. Forty participants gave a power of .80 to discover an equivalent effect size. We adopted the same sample size for Experiments 2-3.

Reported main and interaction effects were evaluated with "Type III" analysis of deviance (i.e., analogous to Type III Sum of Squares ANOVA) tests based on the Wald statistic. We report unstandardized parameter estimates as effect sizes for GLMMs, and 95% confidence intervals based on the normal distribution. In all analyses of average differences between groups, we included a logarithmic term for trial number (by phase, i.e., ln[1-35]) as a nuisance regressor to account for non-linear individual differences in learning curves (Including this regressor improves statistical model fit in all analyses. All results are qualitatively identical without this regressor). For consistency with these analyses, all figures depicting effects averaged across trials (e.g., Fig. 3-5 A, main text) show estimates adjusted for learning curves.

Competing Systems Model

The competing systems – model was based on two independent, basic reinforcement learning algorithms. During conditioning, the Pavlovian controller learned to predict aversive outcomes:

$$V^{i}_{Pavlovian}(t+1) = V^{i}_{Pavlovian}(t) + \alpha\left(R(t) - V^{i}_{Pavlovian}(t)\right) \qquad [1]$$

where R = [-1,0], and α is the learning rate. During decision-making, the instrumental controller learned the expected value of actions in an identical manner:

$$Q^{i}_{Instrumental}(t+1) = Q^{i}_{Instrumental}(t) + \alpha\left(R(t) - Q^{i}_{Instrumental}(t)\right) \qquad [2]$$

where R = [-1,1] (4, 5), and α is, for simplicity, the same learning rate as used by the Pavlovian controller. In contrast, the Pavlovian controller learned about the compound stimulus *ij* during decision-making because the two cues were presented together and co-terminated with the decision (Figure S1):

$$V^{i}_{Pavlovian}(t+1) = V^{i}_{Pavlovian}(t) + \alpha(R(t) - (V^{i}_{Pavlovian}(t) + V^{j}_{Pavlovian}(t)))$$
[3]

As expressed in the Rescorla-Wagner rule (6), the consequence of this is that no differential Pavlovian learning occurs during decision-making. In other words, the Pavlovian influence was constant during decision-making (assuming no Pavlovian updating at all during decision-making gives almost identical predictions). Finally, the learned cue values ($V_{Pavlovian}$ and $Q_{Instrumental}$) from both systems are multiplied to compute the probability of choice *i* at time *t*. We express competition between controllers in a standard Softmax function:

$$P(i) = \frac{e^{(1-\omega)Q^{i}_{Instrumental}+\omega V^{i}_{Pavlovian}/\beta}}{\sum_{j=1}^{n} e^{(1-\omega)Q^{j}_{Instrumental}+\omega V^{j}_{Pavlovian}/\beta}} \qquad [4]$$

where β ($0 < \beta \leq 1$) regulates how deterministically the action with the highest value will be chosen, *n* refers to cues *A* & *B*, and *ω* is the relative weight on the Pavlovian system.

Model simulations

The simulations of the competing systems - model were conducted using the same reinforcement probabilities and number of trials as in the experiment, but independent of the experimental time-series (i.e., simulated participants). We refer to these as the a priori predictions of the model. These were derived from the mean performance across 100 model runs for each parameter combination ($0.01 \leq \alpha \leq 1$, $0.01 \leq \beta \leq 1$, in steps of 0.1. *ω* was set to 0.5), and are displayed in Figure 2C of the main text.

We also we investigated the predictions of the competing systems – model in more detail. We first asked how the Pavlovian weight normatively should be set to promote adaptive behavior. To address this, we simulated the model to find the weight that maximized decision-making performance (i.e., minimized punishment) for different probabilities of environmental change (*C* parameter). The simulations showed that if the environment on average is stable, a positive weight (with the optimum at *ω* = .5) on the Pavlovian controller is adaptive (Figure S2). As we describe in the main text, *ω* was not different from 0.5 in either Exp. 1 or 2, suggesting that behavior well-adjusted under the assumption of a stable environment. What happens if Pavlovian weight is positive and the environment in fact does change? As we describe in the main text (Figure 2), this results in strictly maladaptive decision-making.

To assess how the generality of prediction of maladaptive decision-making after an environmental change, we plot the average difference in decision-making performance between No Change and Change groups (for *ω* = .5) across the parameter space for the learning rate α and temperature β (Figure S4). The figure shows that the prediction that changing the environment will lead to maladaptive decision making is a general feature of the model. Similarly, we plot (Figure S5) predicted difference in behavior if the environment is unstable also during decision—making, as in the Reversal phase of our experimental task (see Figure 2, main text). As seen, the reversal of the Pavlovian influence generalizes across the parameter space.

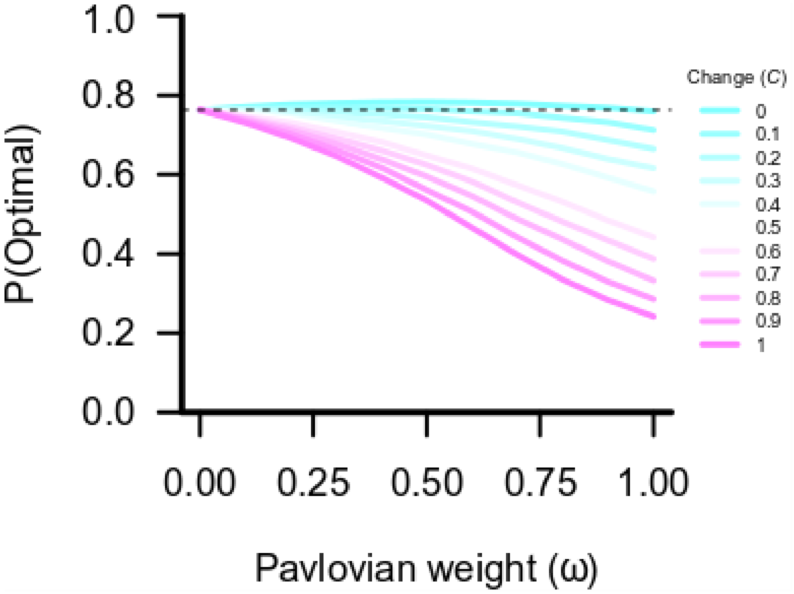

Figure S2. Consequences of Pavlovian weighting for adaptive behavior (A) Decision-making performance as a function of the weight, *ω,* on the Pavlovian controller and the probability (parameter *C*) that the environment changes from fear conditioning to decision-making (averaged across 100 runs for parameter combination of α and β). The horizontal dotted line indicates for reference the average performance if the Pavlovian influence is null (*ω* = 0). For a stable environment, *ω* = 0.5 is optimal.

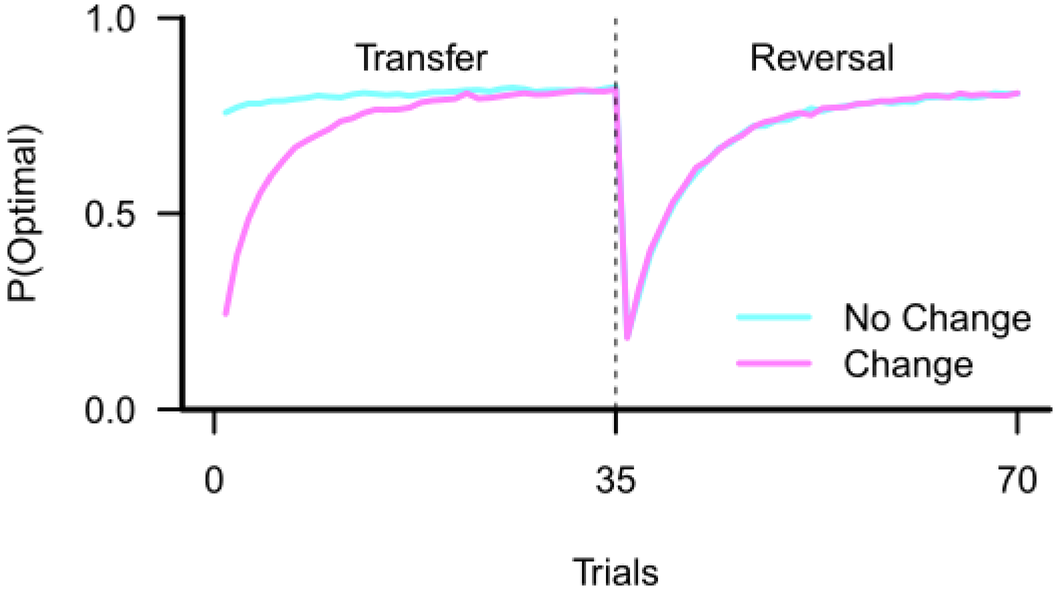

Figure S3. A one-system model where a single controller system learns during both conditioning and decision-making does not predict a difference between groups after the reversal. A priori asymptotic model predictions (by simulating across the range of the model parameters and taking the mean as expectation, see Methods) for the *No-change* (turquoise line) and *Change* (magenta line) condition derived from a one system – model where the same expected values were updated in both condition and decision-making (i.e., a basic Q-learning model). See Table S1-S2 for quantitative evaluation of this model.

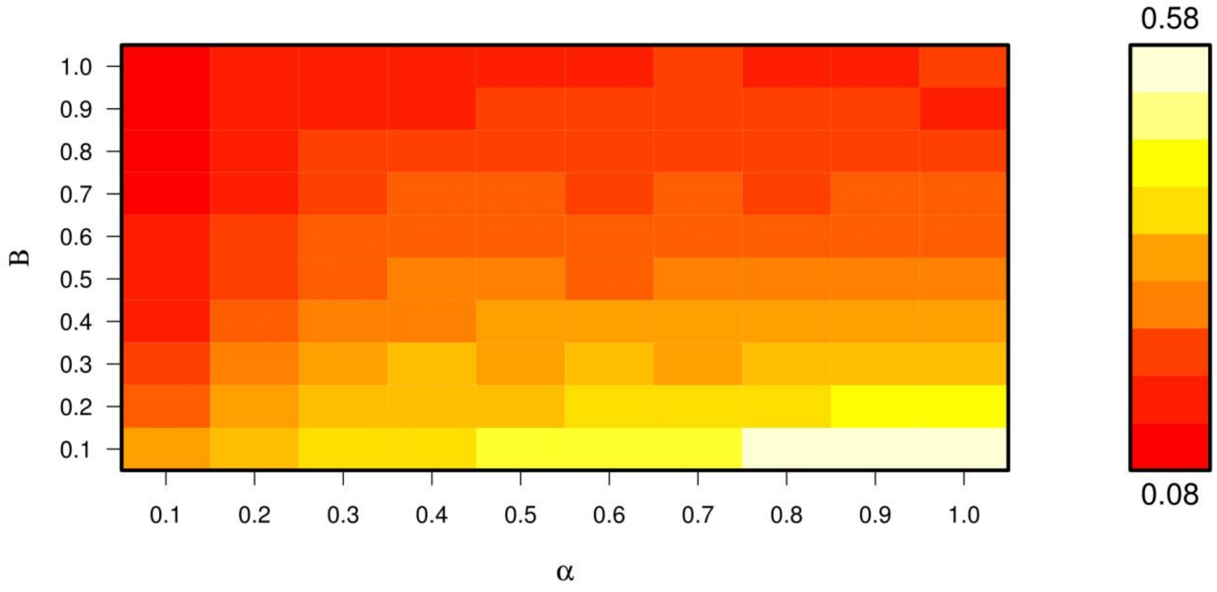

Figure S4. Simulation of the competing systems - model across the parameter space of the prediction that an environmental change results in maladaptive decision-making (i.e., P(Optimal) No Change > P(Optimal) Change) during Transfer (for $\omega = 0.5$). Empirical values

were within the range of predictions (Exp. 1 = 0.11, Exp. 2 = 0.23). Each cell is the average of 100 simulation runs.

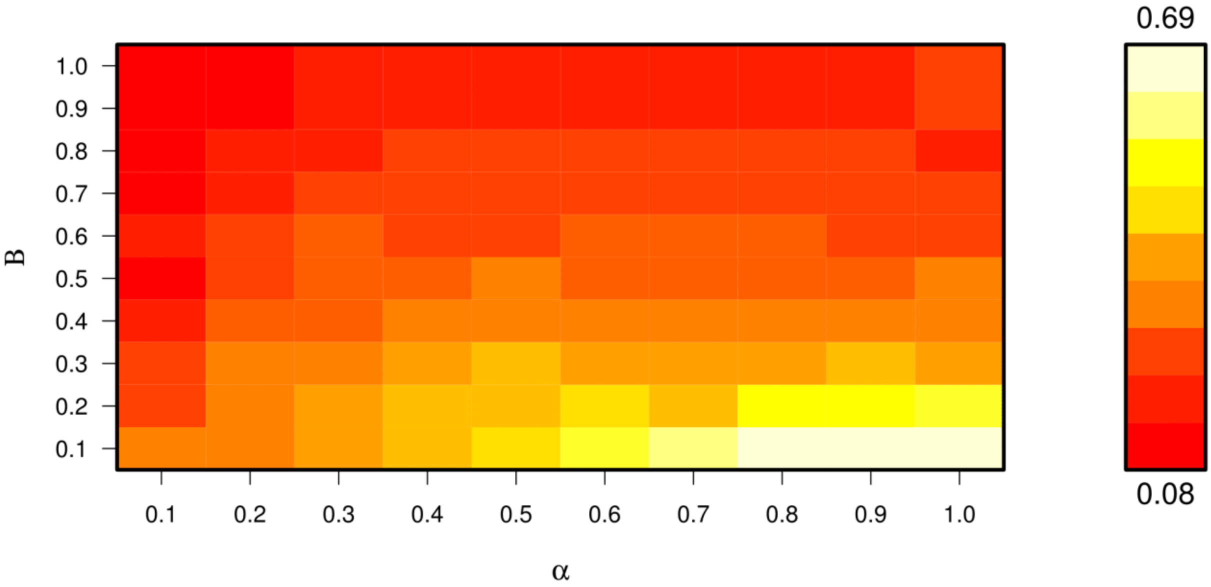

Figure S5. Simulation of the competing systems - model across the across parameter space of the prediction that the difference between the groups would reverse in the Reversal phase (i.e., P(Optimal) Change > P(Optimal) No Change) (for ω = 0.5). Empirical values were within the range of predictions (Exp. 1 = 0.13, Exp. 2 = 0.24). Each cell is the average of 100 simulation runs.

Parameter estimation

Parameter estimation was conducted using the maximum-likelihood approach, which finds the set of parameters that maximize the probability of the participant´s trial-by-trial choices given the model. Optimization was done by to minimizing the negative log-likelihood, -*L*, computed by:

$$-L = -\sum_{t=1}^{T} ln\,(P_{choice}(t)),$$

where *T* denotes the total number of trials. Parameters were independently fitted to each subject using the BFGS optimization method. To avoid local minima in parameter fitting, optimization was initiated with 10 randomly selected start values. Model implementations and parameter fitting was done in R (R Development Core Team, 2012).

Model comparison

Model comparison was primarily based on the Akaike Information Criterion (AIC), a measure of goodness of fit of a model that penalizes complexity (7):

$$BIC = -2\ln(L) + 2k$$

where *–ln(L)* is the negative log-likelihood and k is the number of model parameters. A smaller AIC hence indicates a better model fit.

We summed the AIC across participants, and calculated AIC weights (*wAIC*) from the summed AIC, to derive a group-level measure of model fit. The *wAIC* can be interpreted as the probability that a given model has the lowest AIC, and thus best fit, in the set of candidate models (8).

Results

	Competing systems	**One system 1**	**One System 2**
AIC (sum)	2373.1	2451.77	2394.81
wAIC	0.99998	0	0.00002
α	0.46 (0.35)	0.4 (0.29)	0.39 (0.33)
β	0.28 (0.35)	0.5 (0.32)	0.44 (0.34)
ω	0.5 (0.39)	-	-
α_{Direct}	-	-	0.3 (0.32)

Table S1. Model comparison in Experiment 1. Model comparison between the competing systems – model and two version of the model involving only one system (i.e., equation 4 only included instrumental Q-values) showed that the competing systems – model explained the data best. The One system models either had the same learning rate, α, in the conditioning and decision-making blocks (One system 1, cf. Figure S2), or different learning rates in the two blocks (One system 2). The table displays summed (across participants) Akaike Information Criterion (AIC) (smaller values indicate better model fit), and mean (standard deviation) parameter estimates for the three models. Akaike Information Criterion weights (*wAIC*) can be interpreted as the probability that the model has the lowest AIC in the candidate set, and thereby provides the best explanation of the data.

	Competing systems	**One system 1**	**One System 2**
AIC (sum)	2486.58	2496.81	2507.15
wAIC	0.994	0.006	0.00003
α	0.5 (0.34)	0.41 (0.28)	0.37 (0.31)
β	0.26 (0.29)	0.46 (0.32)	0.43 (0.33)
ω	0.5 (0.35)	-	-
$\alpha_{\text{Observational}}$	-	-	0.18 (0.24)

Table S2. Model comparison in Experiment 2. Model comparison between the competing systems – model and two version of the model involving only one system (i.e., equation 4 only included the instrumental Q-values) showed that the competing systems – model explained the data best. The One system - models either had the same learning rate, α, in the conditioning and decision-making blocks (One system 1, cf. Figure S2), or different learning rates in the two blocks (One system 2). The table displays summed (across participants) Akaike Information Criterion (AIC) (smaller values indicate better model fit), and mean (standard deviation) parameter estimates for the three models. Akaike Information Criterion weights (*wAIC*) can be interpreted as the probability that the model has the lowest AIC in the candidate set, and thereby provides the best explanation of the data.

Conditioned autonomic arousal responses does not predict Pavlovian transfer

We tested if individual differences in conditioned autonomic skin conductance responses (SCR) predicted the magnitude of Pavlovian transfer (focusing on the Transfer phase of the decision-making block). This was not the case. We assess individual differences in the differential (CS+>CS-) SCR response, in interaction with Group (Change/No Change) and Experiment (Exp. 1/Exp. 2 [we excluded Exp. 3 due to the absence of a transfer effect in that experiment). If higher SCR during conditioning predicts the transfer magnitude, we would expect a SCR*Group interaction (because the slopes should have different signs). Neither this interaction (χ^2 (1) = 0.19, p = .67), the main effect (χ^2 (1= 0.27, p = 0.6), nor interaction with Experiment (χ^2 (1) = 0.15, p = 0.7) were significant. In other words, in our paradigm, individual differences in conditioned responses, as indexed by SCR, did not predict the magnitude of transfer, neither in Pavlovian, nor in observational conditioning. Although at first glance puzzling, it is known from non-human animals that physiological reactions thought to indicate conditioned threat responses typically do not predict avoidance behaviors (9). Thus, the relationship between affective state, physiological correlate, and instrumental behavior might not be as straightforward as the first glance suggests.

Empathy with the Demonstrator and the unpleasantness of watching the Demonstrator receiving shocks predicts Pavlovian transfer in Experiment 2

In Experiment 2 (observational threat learning), the participants were prompted for their impressions of the Demonstrator directly after the conditioning block. We asked (i) how unpleasant it was to see the Demonstrator receiving electric shocks, (ii) how unpleasant they thought it would be for themselves to receive electric shocks, (iii) how unpleasant they thought it was for the Demonstrator to receive electric shocks, (iv) how natural the Demonstrator seemed, (v) how expressive the Demonstrator was, (vi) how much empathy they felt with the Demonstrator, (vii) how much they identified with the Demonstrator. We entered all these questions as predictors of the (arcsine transformed) proportion of correct responses during the Transfer phase, in interaction with Group (Change/No Change) in a linear model. To reduce the regression model, we used (two-ways) step-wise model selection using the AIC. All variables except *i* & *v* were left in the final model, but only two (in addition to Group, which naturally explained most variance) reached conventional significance (Group**i* (F(1,29) = 5.19, p = .03) and Group**vi* (F(1,29) = 4.56, p = .04)). In other words, the unpleasantness of watching the demonstrator receiving shocks and empathy with the demonstrator predicted a stronger influence of observational fear learning on behavior. Intriguingly, empathy with the demonstrator has previously been shown to moderate observational fear learning (10). Although both relationships are meaningful, their robustness, given the large number of predictors, and importance are unknown and requires future investigation.

Comparison with a baseline control experiment without a conditioning block

Because both the Change and No Change groups were preceded by a conditioning block, the results reported in the main text does not allow determining the directionality of the Pavlovian transfer effect (all comparisons are relative the other group, e.g., No Change > Change). To enable determining the influence of the Pavlovian transfer effects in absolute terms, we performed a baseline control experiment (n = 25, Experiment 4). Experiment 4 (18 women, mean age = 27) had exactly the same decision-making block as Experiments 1-3, but no conditioning block preceding decision-making. This allows a direct estimation of the absolute effect of the Pavlovian transfer observed in Experiments 1-2.

Because the influence of the conditioning block on decision-making did not reliably differ between Experiment 1 and Experiment 2 (see main text), we pooled the data from these experiments to maximize estimation precision. We first compared decision-making in the Transfer phase against the neutral baseline experiment, and found that the *Change* group had significantly reduced performance relative to this baseline (β = -.59, SE = .27, z = -2.22, p = .026). In contrast, the *No change* group had facilitated performance (β = 0.88, SE = 0.27, z = 3.33, p = .0008), Figure S6. Thus, in line with previous PIT findings (11), transfer of Pavlovian associations was maladaptive when misaligned with instrumental decision-making, but beneficial when aligned (the same pattern was evident when analyzing the experiments separately, although the effects were stronger in Experiment 2). In the Reversal phase, none of the groups were significantly different from the baseline experiment, although both effects had the expected signs (i.e., reversed relative to the Transfer phase) and similar magnitudes: No Change groups versus baseline (β = -.4, SE = .28, z = -1.43, p = 0.15), and Change groups versus baseline (β = 0.42, SE = .28, z = 1.48, p = .14) (see Figure S6).

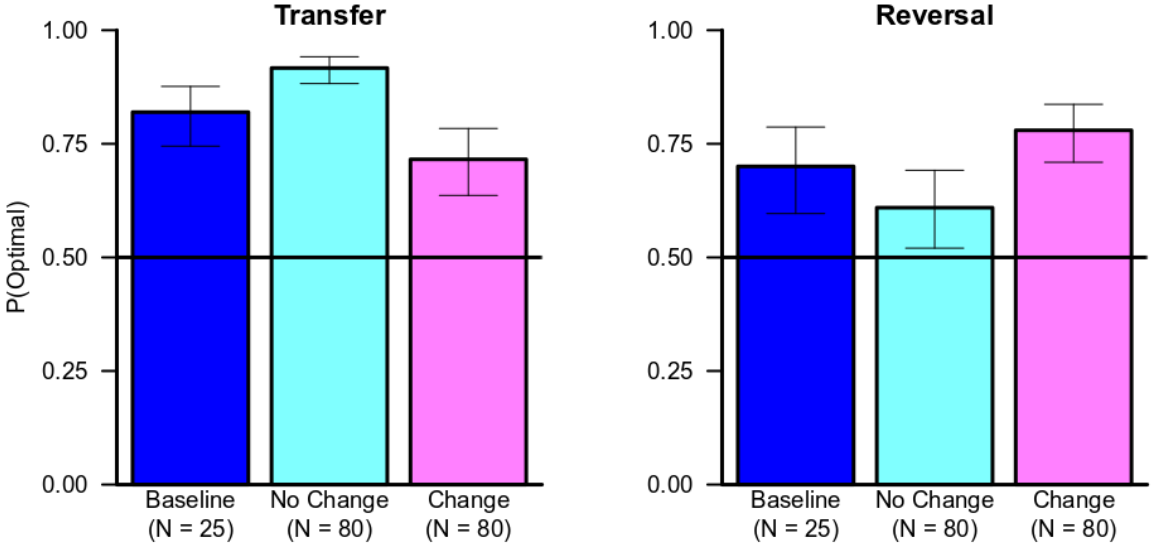

Figure S6. Estimating the directionality of Pavlovian transfer by comparison against a baseline experiment (Experiment 4, n = 25) without preceding Pavlovian conditioning. Probability of selecting the optimal action for *Baseline, No Change* and *Change* groups during the decision-making block. Experiment 1 & 2 are pooled. Error bars are 95 % parameter CI.

Does the competing systems – model predict the same pattern? To address this, we conducted simulations of the baseline condition (by setting the Pavlovian weight, ω, to 0).

As depicted in Figure S7, this was the case: the model predictions accurately captured the qualitative empirical pattern (c.f., Figure S6). As described in the main text, model predictions were generated by simulating the model across the parameter range, and using the average as the expectation. In other words, the simulations were not informed by the empirical results. These results demonstrate that the competing systems – model captures many important features of how Pavlovian associations transfer to decision-making.

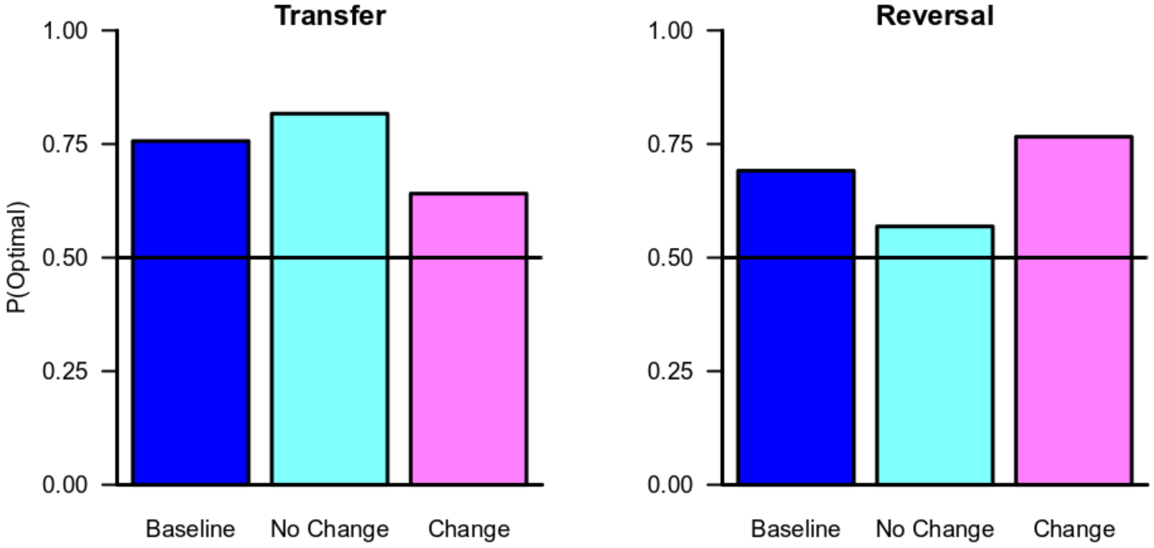

Figure S7. Competing systems – model predictions of the baseline condition. The figure shows the model predicted probability of optimal responses for the neutral baseline condition (ω = 0) relative to the *No Change* and *Change* conditions (ω = 0.5) (c.f., Figure S6).

Additional model comparison in Experiment 3.

The model comparison in the main text is based on a fixed effects model comparison method (i.e., group AIC weights), which can be sensitive to outliers. To ascertain the robustness of the conclusions that different computational mechanisms underlie the transfer from observational and instructed threat learning, we in addition used Bayesian random effects model comparison, which allows for model heterogeneity between participants (12). To this end, we used both the AIC (Figure S8) and the BIC (Figure S9) as approximations to model evidence, and computed the exceedance probability for each model. We used both AIC and BIC as these two criteria penalize model complexity differently (13). The exceedance probability expresses the probability that a given model is the most common model among the candidate models in the population (12). These additional model comparisons provided converging results with those reported in the main text, and strengthen our conclusion that different models best describe the transfer of observational versus instructed threat learning. See following sections "*Additional details about the Prior model"* and *"Difference in predictions from the Competing Systems* - *and Prior* - *models"* for additional information.

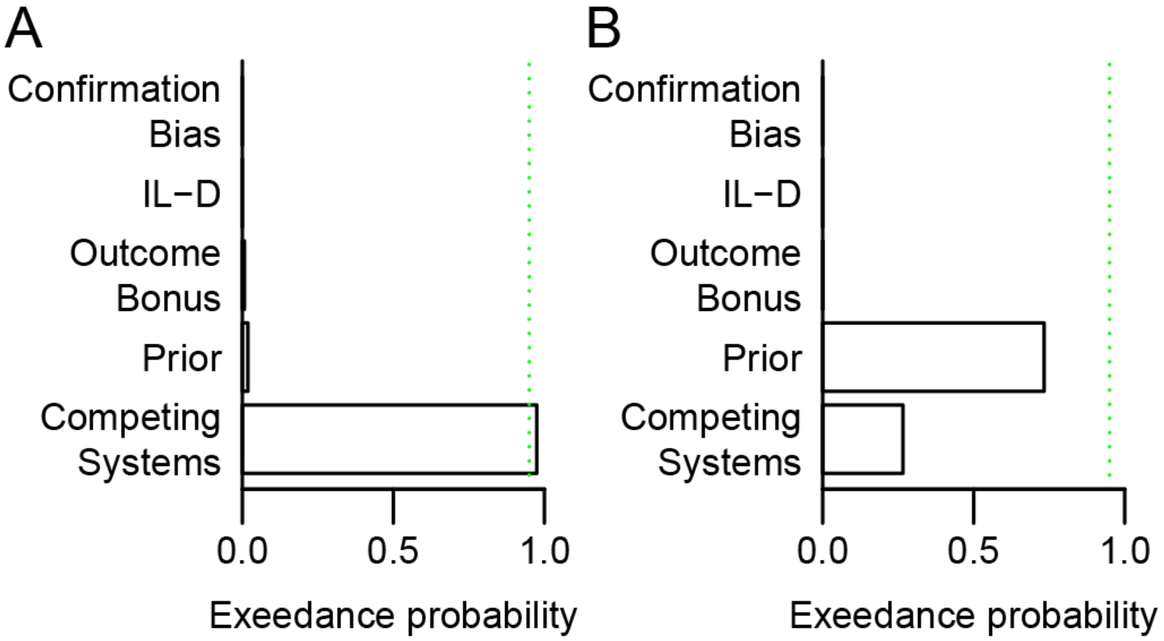

Figure S8. Bayesian random effects model comparison based on AIC as approximation to model evidence. (A) Experiment 2 - Observational threat learning. (B) Experiment 3 – Instructed threat learning. The exceedance probability expresses the probability that a model is the most common model among the candidate models in the population. The green dotted line denotes probability = 0.95.

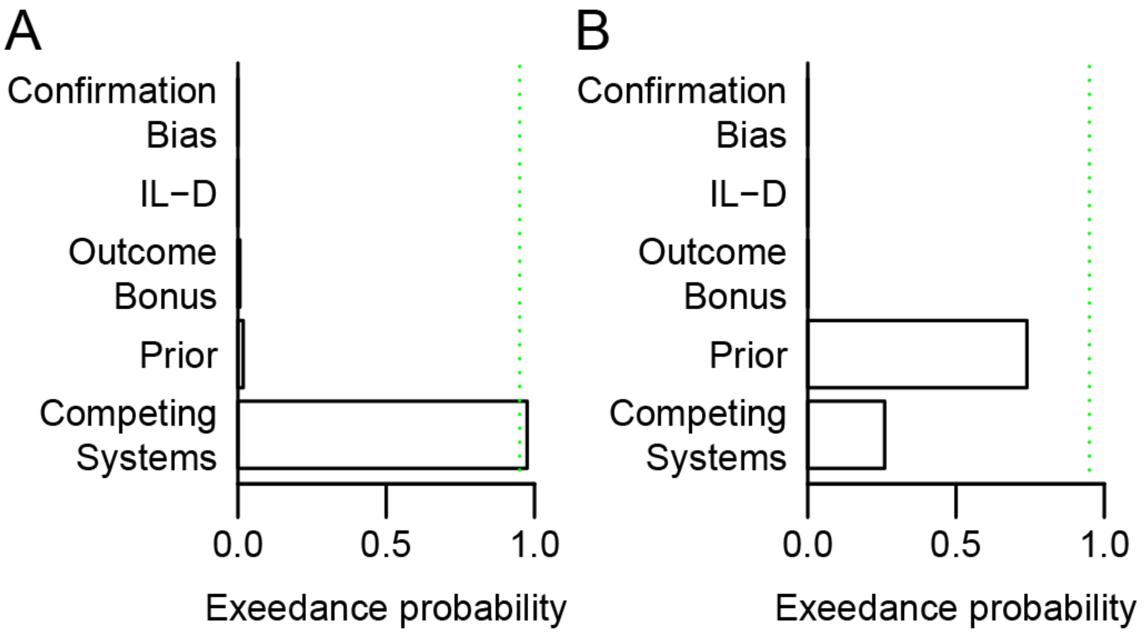

Figure S9. Bayesian random effects model comparison based on BIC as approximation to model evidence. (A) Experiment 2 - Observational threat learning. (B) Experiment 3 – Instructed threat learning. The exceedance probability expresses the probability that a model is the most common model among the candidate models in the population. The green dotted line denotes probability = 0.95.

Additional details about the Prior - model

The Prior – model, which best described the data in Experiment 3, states that instruction (or advice, as in its original applications (14, 15)) determines the initial instrumental action (i.e., Q) values at the outset of the decision-making block (see Figure S10). This differs from the Competing Systems - model (and most other models we consider), where the instrumental action values were initialized to 0. Specifically, in the Prior - model, the action values associated with the CS+ and CS- cues differed on trial 1: Q_{CS+} (t = 1) = -ρN and Q_{CS-} (t = 1) = ρN, where ρ ($0 \leq \rho \leq 1$) is a free parameter and N the number of trials (= 70) in the decision-making block (14, 15). Importantly, the Prior – model only had one system (i.e., instrumental), which allows eventually overcoming the influence of misleading threat instructions. The Prior – model thereby predicts the largest influence of threat instructions on decision-making at the outset of the decision-making block, and the smallest at the end of the decision-making block.

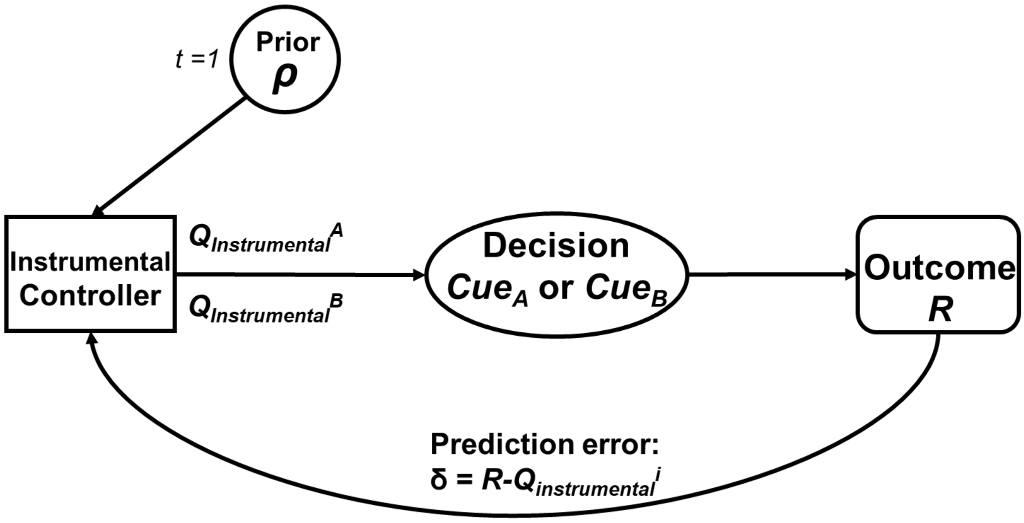

Figure S10. The structure of the Prior – model. An instrumental controller learns about the value of actions from their outcomes during decision-making. Instruction functions as a prior on instrumental action values at the outset of decision-making (*t = 1*).

Difference in predictions from the Competing Systems - and Prior - models

To understand which features of the data distinguished between the prediction of the prior and competing systems models, we compared the fitted model predictions (Figure S11). As expected from the formulation of the Prior – model, we found that the models make divergent predictions at the beginning of the experiment, where the prior model accounts better for the initial difference between the groups than the competing systems - model (cf. Figure 1). The average absolute difference between the model predictions was indeed largest at the outset of the experiment (black dotted line, Figure S11).

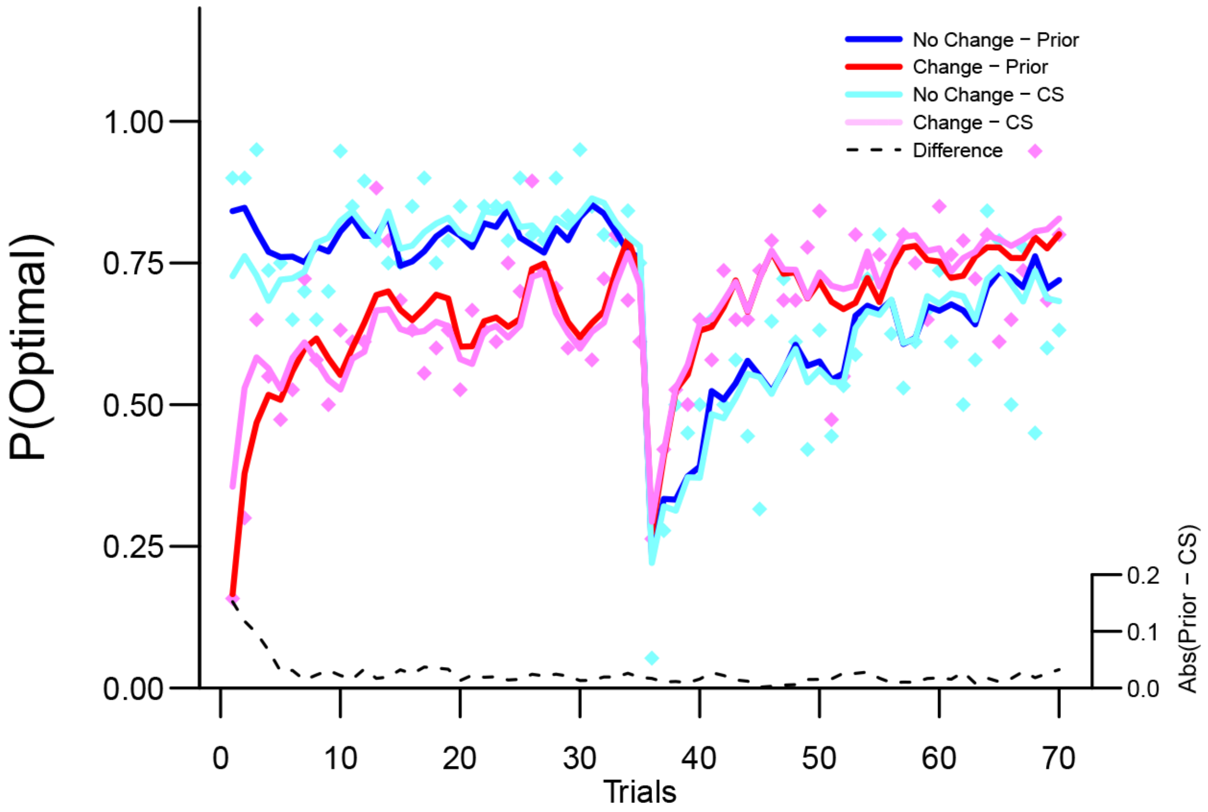

Figure S11. Comparison of predicted behavior for the Prior and Competing Systems models in Experiment 3. Average trial-by-trial fitted model predictions (colored lines) for the two models, overlaid on the experimental data (dots). The black dotted line shows the average absolute difference in predictions per trial for the prior vs competing systems – models. Prior = Prior model. CS = Competing Systems – model.

Experiment 3B

In addition to the novel instructed threat learning paradigm we present in the main text (Experiment 3), we had originally conducted a more traditional variant of an instructed threat learning experiment (Experiment 3B, N = 40, 18 women, mean age = 26). In this standard paradigm (2) participants were instructed that they would receive shocks to one cue (CS+) but not the other (CS-) (2). In contrast to Experiment 3, the CS cues were then presented without reinforcement for the same total number of times as in Experiments 1-2, potentially leading to stronger expectation violation and lower inferred reliability of the environment in this experiment compared to the observational threat learning experiment. The instruction resulted in a clear differential SCR response during the conditioning block (t(39) = 9.12, p < .0001), indicating successful instructed threat learning. However, there was no evidence for transfer to decision-making (β = 0.27, SE = 0.27, z = 1.0, p = 0.32, 95 % CI [-.79, 0.26]), or an interaction between Phase and Group ($\chi^2(1) = 0.06$, p = .81) in Experiment 3B (Figure S12). Although these result are intriguing, it is possible that participants learned that the environment could not be trusted, due to the omission of the expected shocks, which may have resulted in an extinction of the instructed threat expectancy. The version of Experiment 3 presented in the main text circumvents this concern, and shows that that instructed threat learning can transfer to decision-making.

Although no transfer effect was visible when averaged across trials, computational modelling can allow estimating more subtle behavioral effects than direct comparison of means. We therefore estimated the same set of social learning models for Experiment 3B as reported for Experiment 3 in the main text. Replicating the finding that instructed threat learning is best understood as a prior on the instrumental system, the Prior – model provided the best explanation of the data: *wAIC* (Prior) = 1. As the transfer effect in Experiment 3B was weak relative to Experiment 3, this should be reflected in the magnitude of the estimated ρ (i.e., Prior) parameter. As expected, the average parameter value of ρ was significantly lower in Experiment 3B (M = 0.33) than in Experiment 3 (M = 0.75), t(63.17) = -3, p = .003. This finding implies that instructed threat learning had a weak influence on action values in Experiment 3B, which was easily overruled following unexpected outcomes. We directly tested this implication by comparing the *No Change* and *Change* groups in Experiment 3B on the very first decision-making trial. The model predicts that instruction should lead to an initial difference in the instrumental expected values. In line with the prediction, groups initially differed in their choices (logistic regression: β = -1.65, SE = 0.70, z = -2.35, p = 0.019, see Fig. S12B). These results together with Experiment 3 provide converging evidence that instructed threat learning influences decision-making as a prior on the instrumental system.

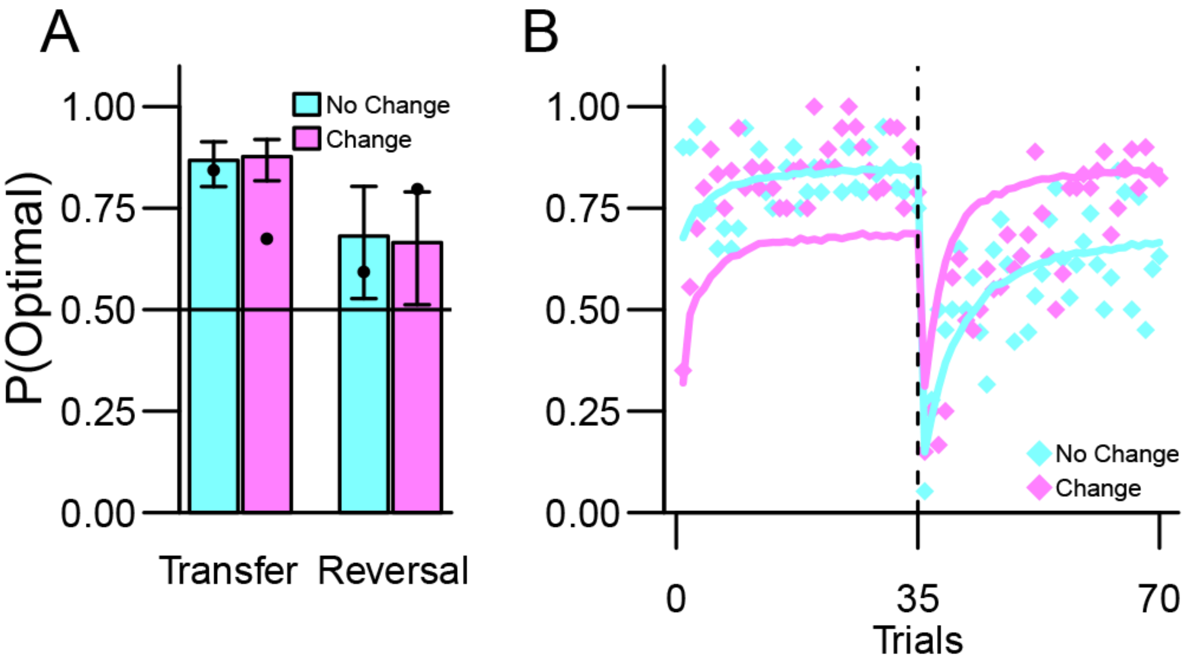

Figure S12. Experiment 3B (N = 40). Probability of selecting the optimal action for *No Change* and *Change* groups during the decision-making block in Experiment 3B. Black points indicate the *a priori* predictions from the competing systems – model. Error bars are 95 % parameter CI from the GLMM model. **(B)** In contrast to Experiments 1-3, there was no difference between the groups in fraction of optimal responses during neither the Transfer phase (trials 1-35, see Fig. 1), nor the Reversal phase (trials 36-70). The solid lines show the *a priori* predictions from the competing systems – model.

Experiment 3C

In Experiment 3C (N = 40, 25 women, mean age = 24.9), we investigated how a learning environment that provided both instruction of threat and actual experience of shock influenced decision-making. As in Experiments 3, participants were instructed that they would receive shocks to one cue (CS+) but not the other (CS-) at the start of the conditioning block, and (as text on screen) before every CS presentation. In contrast to Experiments 3, participants in Experiment 3C in addition experienced shocks in accordance to these instructions (following the same conditioning schedule as in Experiment 1). This design prevented violation of expectations arising from instructions in a different manner than Experiment 3. Moreover, it allowed us to estimate how the combination of instructed threat learning and fully-expected direct shock experience transfers to decision-making. As expected, Experiment 3C replicated both the Transfer (β = -0.88, SE = 0.38, z = -2.31, p = .02, 95% CI [-1.63, -0.14]), and reversal (Group*Reversal: χ^2 (1) = 8.1, p = .004) of Pavlovian threat learning to decision-making (Figure S13). Next, we compared Experiments 3 and 3C to discern if transfer was stronger when instruction and direct experience were combined (as in Experiment 3C) than for instruction alone (Experiment 3, main text). This was not the case (simple interaction, Group*Experiment: $\beta = 0.428$, $SE = 0.24$, $z = 1.8$, $p = 0.08$). If anything (based on the non-significant interaction estimate), transfer was stronger following instructed threat learning in the absence of shocks. Furthermore, there was no evidence for a difference between experiments in the Group*Reversal interaction (χ^2 (1) = 0.3, p = .59). In other words, the combination of instruction and direct experience (Experiment 3C) did not result in more potent transfer to decision-making than instruction alone (Experiment 3).

For completeness, we fitted the Competing Systems – and Prior – models to the data from Experiment 3C. The Prior – model explained the data best (*wAIC* (Prior) ~ 1), suggesting that threat instructions constituted the most important influence on decision-making in Experiment 3C. This result resembles previous reports of equivalent shock expectancies and SCRs after threat instruction on its own, and after combined threat instruction and direct shock experience (16).

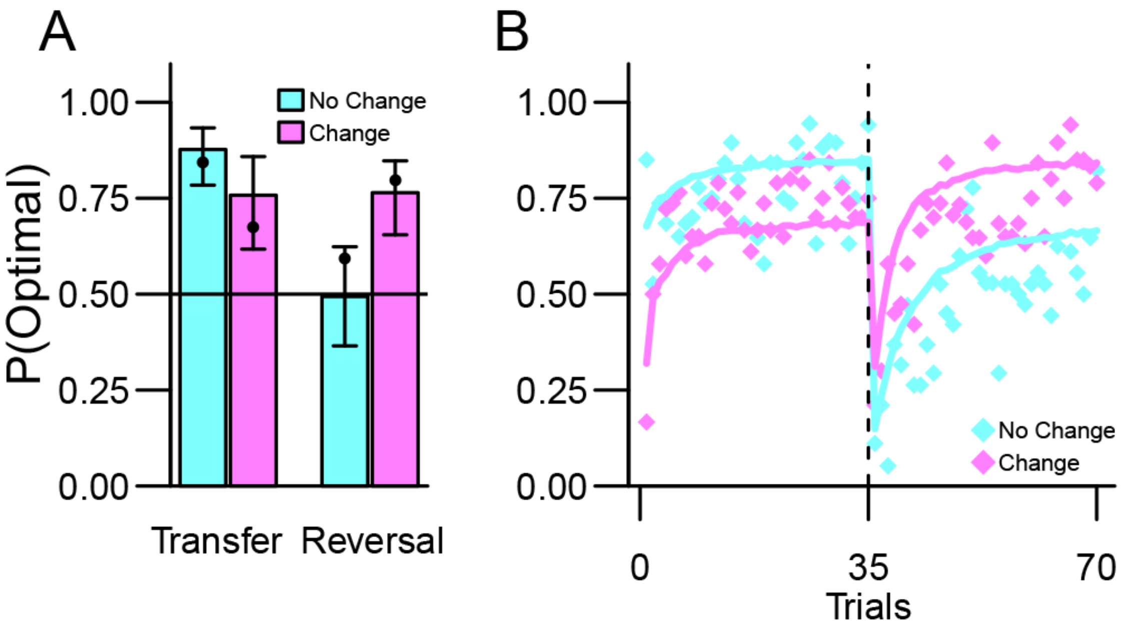

Figure S13. Experiment 3C (N = 40). Probability of selecting the optimal action (i.e., CS with the lowest probability of shock) for *No Change* and *Change* groups during the decision-making block in Experiment 3C. Black points indicate the *a priori* predictions from the competing systems – model. Error bars are 95 % parameter CI from the GLMM. **(B)** Higher fraction of optimal action selection during the Transfer phase by *No Change* group (trials 1-35, see Fig. 1) was reversed during the Reversal phase (trials 36-70). The solid lines show the *a priori* predictions from the competing systems – model.

Supplementary References

1. Haaker J, Golkar A, Selbing I, Olsson A (2017) Assessment of social transmission of threats in humans using observational fear conditioning. *Nat Protoc* 12(7):1378–1386.

2. Olsson A, Phelps EA (2004) Learned fear of "unseen" faces after Pavlovian, observational, and instructed fear. *Psychol Sci* 15(12):822–8.

3. Lindström B, Golkar A, Olsson A (2015) A Clash of Values: Fear-Relevant Stimuli Can Enhance or Corrupt Adaptive Behavior Through Competition Between Pavlovian and Instrumental Valuation Systems. *Emotion* 15(5). doi:10.1037/emo0000075.

4. Huys QJM, Dayan P (2009) A Bayesian formulation of behavioral control. *Cognition* 113(3):314–28.

5. Kim H, Shimojo S, O'Doherty JP (2006) Is avoiding an aversive outcome rewarding? Neural substrates of avoidance learning in the human brain. *PLoS Biol* 4(8):e233.

6. Rescorla RA, Wagner AR (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Class Cond II Curr Res Theory* 21(6):64–99.

7. Daw ND (2011) Trial-by-trial data analysis using computational models. *Decision Making, Affect, and Learning: Attention and Performance XXIII,* eds Delgado MR, Phelps E, Robbins TW (Oxford University Press, New York), pp 1–26.

8. Wagenmakers E-J, Farrell S (2004) AIC model selection using Akaike weights. *Psychon Bull Rev* 11(1):192–6.

9. LeDoux JE, Moscarello J, Sears R, Campese V (2017) The birth, death and resurrection of avoidance: a reconceptualization of a troubled paradigm. *Mol Psychiatry* 22(1):24–36.

10. Olsson A, et al. (2016) Vicarious Fear Learning Depends on Empathic Appraisals and Trait Empathy. *Psychol Sci* 27(1):25–33.

11. Huys QJM, et al. (2011) Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. *PLoS Comput Biol* 7(4):e1002028.

12. Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ (2009) Bayesian model selection for group studies. *Neuroimage* 46(4):1004–1017.

13. Burnham KP, Anderson DR (2002) Model Selection and Multimodel Inference - A Practical Information-Theoretic Approach. *Springer*. Available at: http://www.springer.com/statistics/statistical+theory+and+methods/book/978-0-387-95364-9 [Accessed September 20, 2013].

14. Biele G, Rieskamp J, Krugel LK, Heekeren HR (2011) The neural basis of following advice. *PLoS Biol* 9(6):e1001089.

15. Biele G, Rieskamp J, Gonzalez R (2009) Computational Models for the Combination of Advice and Individual Learning. *Cogn Sci* 33(2):206–242.

16. Raes AK, De Houwer J, De Schryver M, Brass M, Kalisch R (2014) Do CS-US Pairings Actually Matter? A Within-Subject Comparison of Instructed Fear Conditioning with and without Actual CS-US Pairings. *PLoS One* 9(1):e84888.