Social threat learning transfers to decision making in humans

Björn Lindström[a,b,c,1], Armita Golkar[c,d], Simon Jangard[c], Philippe N. Tobler[b], and Andreas Olsson[c]

[a]Department of Social Psychology, University of Amsterdam, 1018 WT Amsterdam, The Netherlands; [b]Laboratory for Social and Neural Systems Research, Department of Economics, University of Zürich, 8001 Zürich, Switzerland; [c]Section for Psychology, Department of Clinical Neuroscience, Karolinska Institutet, 171 77 Stockholm, Sweden; and [d]Department of Clinical Psychology, University of Amsterdam, 1018 WT Amsterdam, The Netherlands

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved January 15, 2019 (received for review June 18, 2018)

In today's world, mass-media and online social networks present us with unprecedented exposure to second-hand, vicarious experiences and thereby the chance of forming associations between previously innocuous events (e.g., being in a subway station) and aversive outcomes (e.g., footage or verbal reports from a violent terrorist attack) without direct experience. Such social threat, or fear, learning can have dramatic consequences, as manifested in acute stress symptoms and maladaptive fears. However, most research has so far focused on socially acquired threat responses that are expressed as increased arousal rather than active behavior. In three experiments (n = 120), we examined the effect of indirect experiences on behaviors by establishing a link between social threat learning and instrumental decision making. We contrasted learning from direct experience (i.e., Pavlovian conditioning) (experiment 1) against two common forms of social threat learning—social observation (experiment 2) and verbal instruction (experiment 3)—and how this learning transferred to subsequent instrumental decision making using behavioral experiments and computational modeling. We found that both types of social threat learning transfer to decision making in a strong and surprisingly inflexible manner. Notably, computational modeling indicated that the transfer of observational and instructed threat learning involved different computational mechanisms. Our results demonstrate the strong influence of others' expressions of fear on one's own decisions and have important implications for understanding both healthy and pathological human behaviors resulting from the indirect exposure to threatening events.

social learning | decision making | Pavlovian instrumental transfer | fear | reinforcement learning

Pavlovian threat conditioning (1), based on the formation of predictive associations between personally experienced stimuli, has long been the standard model of how humans acquire threat associations and learn to avoid threatening and fear-inducing outcomes (2). However, in today's ultra-social world, mass-media and online social networks expose us to unprecedented quantities of second-hand violence (3) and thereby the chance of indirectly forming threat associations between previously innocuous situations (e.g., a subway station) and unpleasant outcomes (e.g., newsreels including cell phone videos of bloody and screaming people after a terrorist attack) without direct experience. Such media-based exposure to violence and trauma is now known to be associated with significant negative psychological outcomes (3). For example, a recent study showed that extended exposure to television newsreels from the Boston Marathon bombings resulted in higher levels of acute stress and PTSD-like symptoms than actually having personally experienced the bombings (4).

This sensitivity to second-hand forms of threat suggests that modern media piggyback on a well-established capacity for social learning, an adaption that allows the organism to avoid the dangers, such as predation, associated with individual learning (5–7). This capacity may cut across species. For example, rhesus monkeys who observed conspecifics behaving fearfully toward snakes readily acquired fear of snakes (8). Although rapid progress has been made toward understanding the mechanisms underlying social threat learning processes (9, 10), these studies have typically focused on passive emotional arousal responses (as in Pavlovian conditioning), and the relevance of such learning for human instrumental decision making is therefore unknown. In real life, threat associations that are socially acquired are likely to shape behavior: we might, for example, decide to avoid a certain subway station because we saw it paired with threat reactions on the news.

Here, we describe this crucial link between social threat learning and instrumental decision making and provide a mechanistic framework for understanding how different types of social information (e.g., video vs. verbal) influence behavior. To experimentally model the social threat learning, we used well-validated paradigms based on threat conditioning (11, 12) rather than more naturalistic material (e.g., actual news reels), as this allowed for a high degree of experimental control while still capturing the "one-to-many" transmission characteristic of mass-media and social networks.

We contrasted the influence of threat learning from direct aversive experience (Pavlovian threat conditioning, experiment 1) against two common forms of social threat learning in human culture—social observation (observational threat learning, experiment 2) and verbal instruction (instructed threat learning, experiment 3)—on subsequent decision making using behavioral experiments and computational modeling. Together, these types of learning correspond to the three archetypal pathways to human fears and phobias as described in the clinical literature (13–15) and reliably induce conditioned physiological and neural threat responses in humans (11).

Significance

In today's world, indirect exposure to threatening situations is more common than ever, as illustrated by footage of terror and disaster in social media. How do such social threat learning experiences shape our decisions? We found that learning about threats from both observation and verbal information strongly influenced decision making. As with learning from our own experience, this influence could be either adaptive or maladaptive depending on whether the social information provided accurate expectations about the environment. Our findings can help explain both adaptive and pathological behaviors resulting from the indirect exposure to threatening events.

Author contributions: B.L., A.G., P.N.T., and A.O. designed research; B.L. and S.J. performed research; B.L. contributed new reagents/analytic tools; B.L. analyzed data; and B.L., A.G., P.N.T., and A.O. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

[1]To whom correspondence should be addressed. Email: bjorn.r.lindstrom@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1810180116/-/DCSupplemental.

Published online February 13, 2019.

Despite their apparent similarity, the mechanisms underlying the two types of social threat learning—observational and instructed—might be very different (11). While learning via observation of conspecifics is phylogenetically widespread and conserved (9), linguistically mediated learning is unique to humans. Indeed, instruction can flexibly modify learned associations. For example, verbal instructions about changed threat contingencies can cause rapid updating of conditioned physiological responses (16, 17), which is reflected by prefrontal cortex activity (while, in contrast, the amygdala tracks direct experience) (17). Misleading advice, however, can result in long-lasting maladaptive influences on reward-based decision making, which are thought to be underpinned by prefrontal and hippocampal influences on the striatum (18). Whether such putative differences in the mechanisms underlying the two forms of social threat learning are expressed as different influences on decision making is currently unknown.

Computational learning theory provides a framework for predicting how different learning mechanisms should affect decision making. An influential account posits that behavior is controlled by at least two interacting valuation systems (henceforth "controllers") with distinct properties (19–22). The Pavlovian controller assigns value to a limited number of behaviors, such as avoidance, in response to biologically relevant stimuli (e.g., pain) (20, 22, 23). In classical conditioning, the Pavlovian controller learns to predict the outcome [the unconditioned stimulus (US), e.g., an electric shock] from a cue [the conditioned stimulus (CS)]. Characteristic conditioned responses (such as increased arousal) reflect these predictions (19). In contrast, the instrumental controller flexibly assigns value to arbitrary actions based on their reinforcement history to guide adaptive selection of the most appropriate action (21). The interaction of these systems is revealed, for example, in the phenomenon of Pavlovian Instrumental Transfer (PIT) (24–27), where the presentation of a Pavlovian CS comes to bias instrumental behavior. Importantly, if the response tendency of the Pavlovian controller (e.g., avoidance) and the optimal instrumental action are a mismatch, this interaction can cause maladaptive consequences (20, 28).

Notably, recent studies have demonstrated strong similarities in the neural and computational mechanisms involved in Pavlovian and observational threat learning (10), suggesting that these forms of learning are based on the same computational system (9, 11). By extension, biases arising from the Pavlovian controller might transfer to decisions involving observationally conditioned cues. We therefore predicted that observational threat learning, in analogy to direct Pavlovian conditioning (29), would transfer to later instrumental decision making. In contrast, instructed threat conditioning, which by definition entails an explicit "model" of the environment (30, 31), is thought to be based on mechanisms that are distinct from basic Pavlovian computations (11, 31–33). We therefore expected that threatening information learned via verbal instruction would be more flexibly adjusted to match changing contingencies, as observed in learning about both threats and rewards (17, 34, 35). We tested these predictions in three main experiments, and several control experiments, using a novel version of a PIT task (27, 29).

Results

In our experiments, participants first underwent a conditioning block (with two cues, yellow and blue shapes serving as CSs) and subsequently an instrumental decision-making block, involving the same two cues as the conditioning block (Fig. 1). The decisions were probabilistically punished with electric shocks. For half of the participants, the environment was stable: the same cue was most likely to be paired with shocks in both parts of the experiment (No Change groups). For the other half, the environment changed: the cue most likely to be paired with shocks differed between the two parts of the experiment (Change groups) (Fig. 1*A*). Our experiments build on the logic of PIT paradigms, but differ in several important ways from previous studies. These studies typically either measured the Pavlovian influence in the absence of external reinforcement (i.e., in extinction) (25–27, 36) or arranged the task so that specific actions (e.g., withholding a response to get a reward) revealed the Pavlovian influence indirectly (22). In contrast, manipulating whether Pavlovian associations were reliable or outdated allowed us to use a standard probabilistic two-choice decision-making task with tangible reinforcement (electric shocks) that provided a direct and general measure of the maladaptiveness/adaptiveness of Pavlovian transfer. To derive clear predictions for when and how learned Pavlovian associations should influence decision-making, we first developed a simple formal model (see *SI Appendix* available online for details), building on previous models of PIT (20, 28, 36),

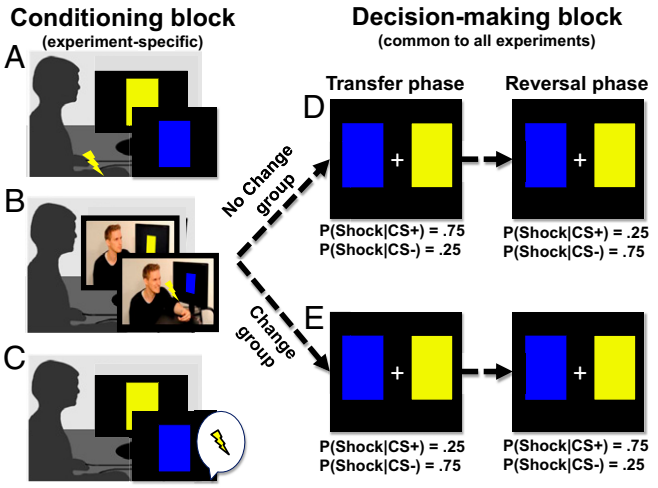

Fig. 1. Experimental design. We conducted three independent experiments, in which a conditioning block (*A*–*C*) was followed by a decision making block (*D* and *E*). The conditioning block of the experiments consisted of (*A*) Pavlovian (classical) threat conditioning (experiment 1), (*B*) observational threat learning (experiment 2), or (*C*) instructed threat learning (experiment 3). In the conditioning block, one stimulus (CS+), but not another (CS−), was paired with shocks for the participant (experiment 1), for another person (experiment 2), or verbally associated with the possibility of shock (experiment 3) (*Methods*). Next, participants made 70 decisions between the same two cues (CS+, CS−), which were both probabilistically punished with electric shocks to the participants (see *SI Appendix*, Fig. S1 for a trial timeline). In each experiment, the participants were divided into two subgroups. The shock probabilities differed between the (*D*) No Change groups, and (*E*) the Change groups. In the No Change groups, the CS+ from the conditioning block had a higher probability of being followed by shocks in the Transfer phase of the decision making block, while in the Change groups, the CS+ had a lower probability of being followed by shocks. After 35 trials, the outcome contingences reversed (Reversal phase). CS, conditioned stimulus.

Our competing systems model assumes that the Pavlovian and instrumental controllers learn independently (Fig. 2) (20, 36). The Pavlovian controller learns about the predictive value of cues (e.g., CSs) during conditioning, and the instrumental controller learns about the expected value of choices (e.g., between the same CSs) during decision making. Finally, the learned cue values ($V_{\text{Pavlovian}}$ and $Q_{\text{Instrumental}}$) from both systems compete to determine choices because the decision cues have both Pavlovian and instrumental values. The relative influence of the Pavlovian controller is determined by a weight (ω, $0 \leq \omega \leq 1$) (20). If this weight is positive, the Pavlovian value associated with a cue will bias the probability of choosing it. The model predicts that changing the environment, as for the Change groups in our experimental task, will result in impaired decision making (Fig. 2*C*), as the Pavlovian associations then are outdated, relative to a

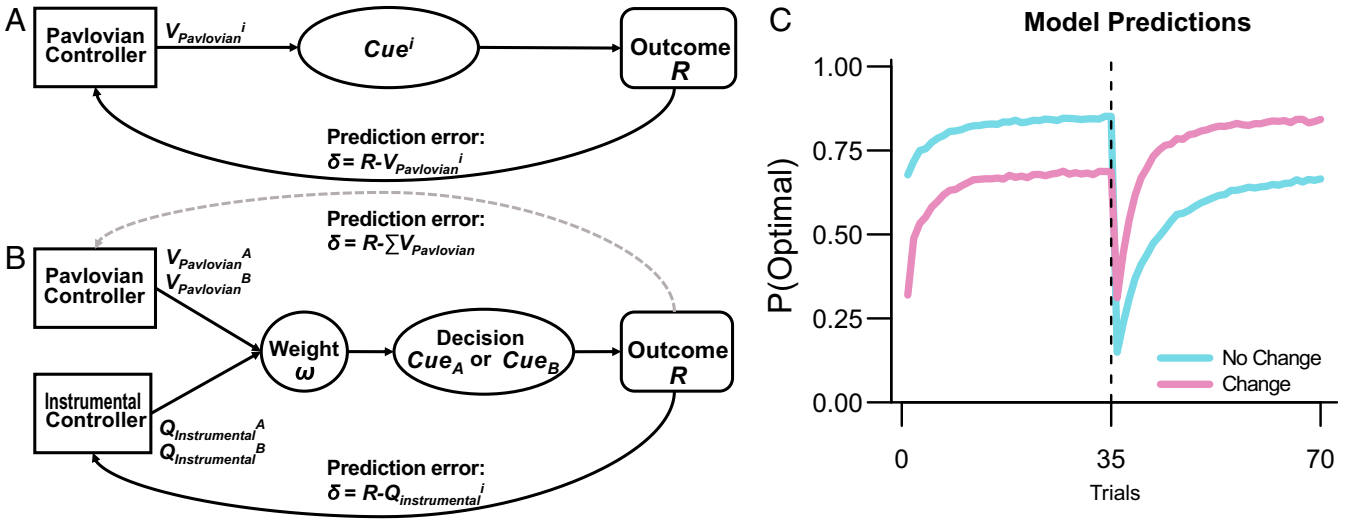

Fig. 2. Model framework for the influence of Pavlovian control on decision making. The structure of the competing systems model, involving a Pavlovian and one instrumental controller. The controllers were implemented as simple, independent reinforcement learning algorithms (see *SI Appendix* for details). (*A*) The Pavlovian controller learns to predict outcomes from cues during conditioning. (*B*) During decision making, the Pavlovian and instrumental controller compete for control of behavior (see *SI Appendix* for details) because the decision cues have both Pavlovian and instrumental values. Only the instrumental values are differentially updated based on the outcome of the decision. The Pavlovian values are updated together, which in practice results in an unchanged Pavlovian influence during decision making (denoted by the dashed gray line). (*C*) Predictions of the competing systems model. A priori asymptotic model predictions (*SI Appendix*) for the No Change and Change conditions. The Pavlovian weight ω was fixed to 0.5 (equal influence of both controllers on decision making). The dashed vertical line denotes the reversal point. If Pavlovian associations transfer from the conditioning to the decision-making block, the model predicts that the performance in Change groups will be reduced relative to the No Change group. After the reversal of the outcome probabilities (Reversal phase), the model predicts that this group difference should be reversed.

stable environment (see *SI Appendix* for data and simulations assessing the influence of Pavlovian transfer relative to a control condition without conditioning preceding decision making).

Furthermore, because our implementation of the competing systems model predicts that the magnitude of Pavlovian transfer should be largely unchanged throughout the decision-making block (*SI Appendix*), we included an additional experimental manipulation that provided a strong test of this prediction: after the first 35 decision trials (Transfer phase), the outcome probabilities were suddenly reversed (Reversal phase). If the Pavlovian influence is constant, as predicted by the model, a reversal of the contingencies should lead to a reversal of this influence (see Fig. 2*C* for model-based predictions and *SI Appendix*, Figs. S4 and S5). In contrast, if the underlying computational mechanisms are more flexible—which would be the case if, for example, the same controller system were learned during both conditioning and decision making—the reversal should attenuate the difference between the groups (*SI Appendix*, Fig. S3, shows how a one-system model predicts a diminishing difference between groups during the transfer phase and no difference in the reversal phase). Finally, the competing systems model also provides a framework for interpreting differences between the two types of social learning (observation vs. instruction). To this end, we quantitatively compare the competing systems model with previously established models of social influences on decision making, which are based on non-Pavlovian mechanisms (18, 34, 37).

Pavlovian Threat Conditioning Transfers to Decision making. The first experiment ($n = 40$) verified that Pavlovian threat conditioning transfers to decision making, which is required for evaluating the predictions about the transfer of social threat learning. The participants first underwent Pavlovian (classical) threat conditioning, where one (CS+), but not the other (CS−), stimulus was followed by electric shocks. This procedure is the quintessential example of Pavlovian learning in the aversive domain. Skin-conductance responses (SCR), a standard measure of autonomic arousal, demonstrated successful threat conditioning [CS+ > CS−, one sample t test: $t(38) = 2.33$, $P = 0.025$].

To test the core prediction that Pavlovian threat learning would transfer to decision making, we first focused on the performance during the Transfer phase of the decision-making block. As outlined (cf. Fig. 2*C*), our model (building on previous PIT research) predicts that transfer of Pavlovian threat conditioning should lead to impaired decision making (i.e., more shocks) if the environment changes between the conditioning and decision-making blocks. Indeed, the Change group had a much decreased probability of making the optimal (with lowest probability of shock) decision relative to the No Change group [random-effects logistic regression: $\beta = -1.28$, SE = 0.34, $z = -3.81$, $P = 0.0001$, 95% CI (−1.94, −0.62) (Fig. 3*A*)]. These results are predicted by previous work on PIT (25, 36) and thereby validate our experimental model. In *SI Appendix*, we in addition compare Pavlovian transfer to a control experiment without a conditioning block preceding the decision making and find that Pavlovian transfer has both beneficial (No Change group) and maladaptive (Change group) effects on decision making relative to this baseline. In other words, the relative difference between the Change and No Change groups reflects the combined effect of these influences.

Next, we tested the strong model prediction that the difference between the Change and No Change groups would be reversed when the outcome probabilities were reversed after the first 35 decision-making trials (Fig. 2*C*). As shown in Fig. 3*A*, the data (bullet points) confirmed this prediction: the difference between the No Change and Change groups reversed after the contingency reversal [Group*Reversal interaction: $\chi^2(1) = 6.86$, $P = 0.009$, simple effects contrast Change > No Change in Transfer phase: $\beta = 0.62$, SE = 0.34, $z = 1.86$, $P = 0.06$, 95% CI (−0.033, 1.29)]. The magnitude of the transfer was not explained by individual differences in conditioned threat responses (*SI Appendix*).

To verify that these results cannot be accounted for simply by a difference in the expected values at the outset of the

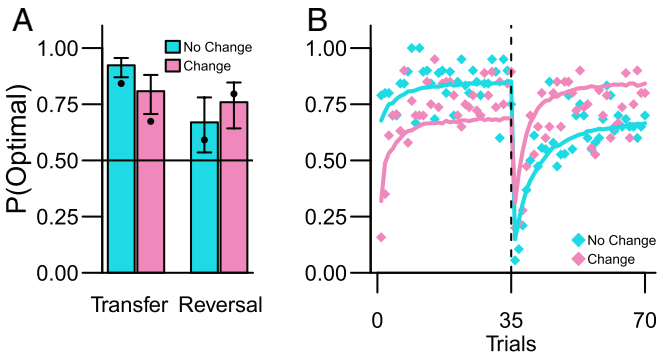

Fig. 3. Pavlovian threat conditioning transfers to decision making (experiment 1). (*A*) Probability of selecting the optimal action (i.e., CS with the lowest probability of shock) for No Change and Change groups during the decision-making block. Bullet points indicate the a priori predictions from the competing systems model (*SI Appendix*). 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, Fig. 1) was reversed during the Reversal phase (trials 36–70). The solid colored lines show the a priori predictions from the competing systems model (*SI Appendix*).

decision-making phase, we next turned to quantitative model comparison. We contrasted the competing systems model with the alternative models that postulated only one system that learned during both conditioning and decision making (cf. *SI Appendix*, Fig. S3). However, these models provided an inferior quantitative account of the data, corroborating the involvement of Pavlovian mechanisms (*SI Appendix*, Table S1).

Taken together, these results show that Pavlovian threat learning transfers to decision making in line with theoretical predictions and previous research. The results of experiment 1 thereby provide a baseline for evaluating Pavlovian transfer of social threat learning to decision making.

Observational Threat Learning Transfers to Decision making. Having established that Pavlovian threat conditioning transfers to decision making as expected by PIT theory, we next compared the two paradigmatic types of social threat learning—observational (experiment 2) and instructed learning (experiment 3)—against this Pavlovian baseline.

In experiment 2 ($n = 40$), we tested our hypothesis that observational threat learning would transfer to, and thereby bias, decision making in the same way as Pavlovian learning. Experiment 2 was identical to experiment 1 apart from one key difference: instead of directly experiencing shocks during conditioning, participants underwent a standard observational threat learning procedure (12), where a video depicted how one (CS+), but not the other (CS−), stimulus was followed by electric shocks to another person, the demonstrator (Fig. 1*A*). Mean SCRs across the conditioning block were higher during presentation of the social CS+ than CS−, confirming social threat learning [one-sample t test: $t(38) = 2.36$, $P = 0.02$, 95% CI (0.054, 0.7)].

We next tested our central prediction that observationally learned threat associations would transfer to decision making. In analogy to the analyses for experiment 1 (above), we first contrasted decision-making performance between the No Change and Change groups. Mirroring the results from experiment 1, transfer from observational threat learning to decision making resulted in markedly impaired decision making when the environment had changed, relative to if it was stable [$\beta = -1.73$, SE = 0.32, $z = -5.38$, $P < 0.0001$, 95% CI (−2.36, −1.1) (Fig. 4*A*)]. Accordingly, the Change group received 20.4% more shocks during the Transfer phase than the No Change group (two-sample Wilcox test: $W = 295.5$, $P = 0.009$) (see also *SI Appendix* for comparison with a control experiment without a conditioning phase preceding decision making). As for experiment 1, the magnitude of the transfer was not explained by individual differences in conditioned defensive responses (*SI Appendix*). However, individual differences in empathy with the demonstrator, and the subjective unpleasantness of observing the demonstrator receive electric shocks, slightly moderated transfer strength (*SI Appendix*).

Finally, we tested the prediction that the effect of transfer would be reversed after reversing the outcome contingencies during decision making (cf. Fig. 2*C*) as in experiment 1. The data agreed with this prediction [Group*Reversal interaction: $\chi^2(1) = 18.54$, $P < 0.0001$, simple effects contrast of Change > No Change in the Reversal phase: $\beta = 1.09$, SE = 0.40, $z = 2.72$, $P = 0.007$, 95% CI (0.31, 1.88) (Fig. 4*B*)].

Moreover, in line with our hypothesis that Pavlovian and observational threat learning are driven by the same Pavlovian controller, there were no reliable differences in the size of the transfer effect [Group*Experiment interaction in Transfer phase: $\chi^2(1) = 1.39$, $P = 0.24$] or its reversal [Group*Reversal*Experiment: $\chi^2(1) = 1.72$, $P = 0.19$] between experiments 1 and 2. Quantitative estimation of the Pavlovian weight (ω) parameter showed that it was highly similar in both experiments [experiment 1: $M = 0.49$, experiment 2: $M = 0.5$, $t(76.13)$, $P = 0.96$] and in neither was different from 0.5 (smallest $P = 0.49$). As for experiment 1, we also used a quantitative model comparison to verify that these results are not well described by a one-system model (*SI Appendix*, Table S2). Together, these results show how observationally acquired threat associations can transfer to, and thereby bias, decision making.

Instructed Threat Learning Transfers to Decision making. In our third experiment ($n = 40$), we tested how the second paradigmatic type of social threat learning—instructed threat learning—influenced subsequent decision making. We predicted that this type of social threat learning would not transfer to decision making because it is likely not based on the Pavlovian controller (11, 32).

At the start of the experiment, and before each trial, participants were instructed that one (CS+), but not the other (CS−), cue would be followed by shocks. To avoid extinction of the instructed threat expectancy, which is crucial for comparison with experiment 2, participants were never exposed to the CS cues during the conditioning block. Instead, they saw two control stimuli (*Methods* and *SI Appendix* for control experiment 3B that

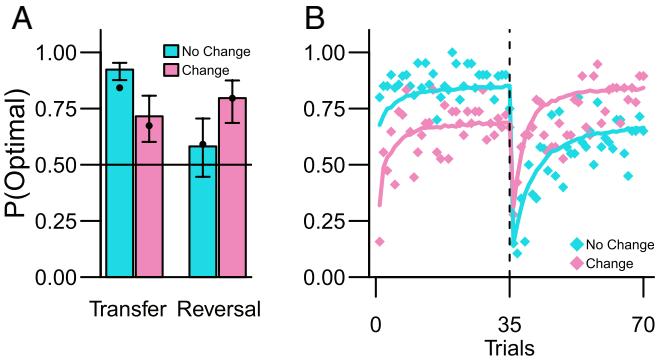

Fig. 4. Observational threat learning transfers to decision making (experiment 2). (*A*) Probability of selecting the optimal action for No Change and Change groups during the decision-making block (i.e., CS with the lowest probability of shock). Bullet points indicate the a priori predictions from the competing systems model (*SI Appendix*). Error bars are 95% parameter CI from the GLMM. (*B*) As in experiment 1, the higher fraction of optimal action selection during the Transfer phase (trials 1–35, Fig. 1) was reversed during the Reversal phase (trials 36–70). The solid colored lines show the a priori predictions from the competing systems model (*SI Appendix*).

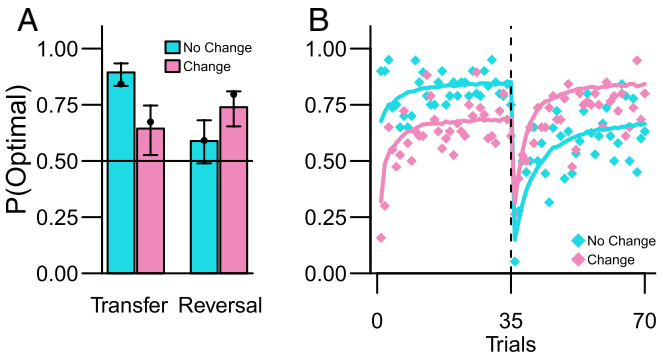

Fig. 5. Instructed threat learning transfers to decision making (experiment 3). (*A*) Probability of selecting the optimal action for No Change and Change groups during the decision-making block. Bullet points indicate the a priori predictions from the competing systems model (*SI Appendix*). Error bars are 95% parameter CI from the GLMM model. (*B*) The higher fraction of optimal action selection during the Transfer phase (trials 1–35, Fig. 1) was reversed during the Reversal phase (trials 36–70). The solid colored lines show the a priori predictions from the competing systems model (*SI Appendix*).

indicates that this design feature is important). In the subsequent decision-making block, participants made choices between the instructed CS cues as in experiments 1–2. In contrast to our prediction, we observed a clear transfer [$\beta = -1.56$, SE = 0.32, $z = -4.85$, $P < 0.0001$, 95% CI (−2.19, −0.93)] and reversal [Group*Reversal: $\chi^2(1) = 19.6$, $P < 0.0001$] of instructed threat learning to decision making (Fig. 5). The magnitude of these effects was comparable to experiment 2 [Group*Reversal*Experiment: $\chi^2(1) = 0.22$, $P = 0.64$]. In an additional control experiment, we evaluated the effect of combining threat instruction and experienced shocks (experiment 3C) and found results comparable to experiment 3 (*SI Appendix*).

Computational Modeling Indicates That Observational and Instructed Threat Learning Rest on Distinct Computational Mechanisms. Our account predicts that the influence of observational threat learning on decision making emerges from the competition between competing Pavlovian and instrumental controllers (Fig. 2), while instructed threat learning might rest on other mechanisms (17). The overt similarity in how observational and instructed threat learning affects decision making might, in other words, reflect the contribution of distinct underlying computational mechanisms.

To test this, we compared the quantitative fit of the competing systems model to a set of prominent previous models of how social advice influences monetary decision making (as there are no previous studies of transfer of social threat learning to decision making). These models assume that advice either leads to a more favorable processing of advice-congruent outcomes [outcome bonus (34, 37), instructed learning D (18), and confirmation bias (35)] or acts as a prior at the outset of decision making [prior model (34, 37)]. In line with our predictions and preceding results (e.g., Figs. 4 and 5), quantitative model comparison showed that the competing systems model explained the transfer of observational threat learning to decision making (experiment 2) better than these alternative models (Fig. 6*A*). In contrast, the influence of instructed threat learning on decision making was best explained by a model where instruction functions as a prior on instrumental action values (Fig. 6*B*; see *SI Appendix* for converging results using Bayesian model comparison and additional details about, and analysis of, the prior model) (37). In the prior model, the threat instruction directly (adjusted by a free parameter) determines the instrumental action values at the outset of the decision-making block, so that the cues have a differential value in the absence of direct experience. In turn, this promotes avoidance of the CS+ cue. This result was replicated in experiment 3B (*SI Appendix*). Together, these results indicate that, despite similar effects on average behavior, the two forms of social threat learning rest on distinct underlying computational mechanisms.

Discussion

We investigated if, and how, socially acquired threat associations would transfer to instrumental decision making. We demonstrate that associations acquired both by social observation (e.g., through video) and by instruction (e.g., through spoken language) strongly transfers to decision making. Notably, this transfer led to maladaptive decisions when socially acquired associations were outdated rather than valid. Our findings present documentation of how social threat learning transfers to human decision making and present the extension of the concept of PIT to social learning.

While the two types of social threat learning had similar effects on decision making, our modeling results indicate that this similarity arose from two distinct computational mechanisms. The transfer of observational threat learning to decision making could, in the same way as Pavlovian conditioning, be explained by competition between a Pavlovian and an instrumental controller system (Fig. 2). In contrast, model comparison suggested that instructed threat functioned as a prior on the instrumental controller in a similar manner to verbal advice (34, 35, 37). This difference in underlying mechanisms might be crucial for understanding how the two types of social threat learning influence behavior and, in extension, how this influence can be counteracted. For example, whereas the competing systems model suggests that Pavlovian counterconditioning or extinction is needed for overcoming the bias elicited by observational threat learning, the prior model implies that sufficient divergent instrumental experience is enough for overcoming the influence of instruction on decision making. Evaluating such predictions is an important goal for future research.

Our findings confirm that social threat learning generalizes beyond the passive expression of emotional responses, as has

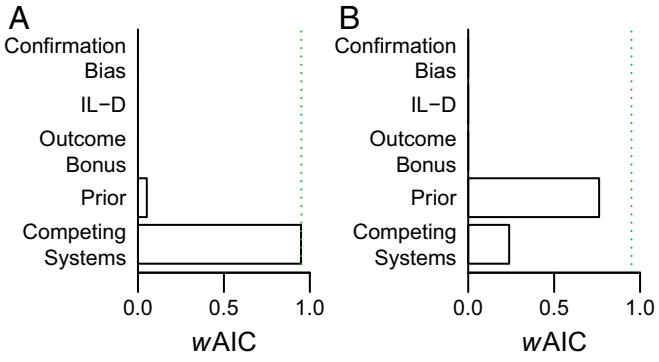

Fig. 6. Model comparison demonstrates that observational and instructed threat learning rest on distinct computational mechanisms. (*A*) Observational threat learning (experiment 2). Model comparison showed that the competing systems model explained the results of experiment 2 better than the alternative social learning models (derived from the literature), which did not posit a competition between systems. (*B*) Instructed threat learning (experiment 3). In contrast to observational threat learning (experiment 2), model comparison showed that a model where instruction functions as a prior on decision making explained the results of experiment 3 better than alternative social learning models (*SI Appendix*). Akaike Information Criterion (AIC) weights (*wAIC*) can be interpreted as the probability that the model provides the best explanation of the data in the candidate set (see *SI Appendix* for results based on Bayesian random effects model comparison). The dotted green line denotes *wAIC* = 0.95. IL-D, instructed learning D model.

typically been the focus of previous research on both observational and instructed threat learning, and thereby provide clear evidence that these paradigms have broad explanatory value for understanding the social transmission of fear and avoidance (12). Furthermore, given the central role of the amygdala in both observational and direct threat conditioning across species (10, 11), characterizing the role of this brain region in the transfer of social learning to decision making is an important goal for future research. However, because instructed threat learning also involves the amygdala (38, 39), it is likely that different patterns of connectivity in overlapping neural networks (10), or multivariate response patterns (33), underpin the difference between observational and instructed threat learning.

In summary, our findings show that threats acquired through observation and instruction can transfer to, and thereby bias, decision making. Furthermore, we show that the transfer of observational and instructed threat learning on decision making rests on different underlying computational mechanisms. We hope that these findings will be valuable for understanding both adaptive and maladaptive human behavior and for informing policy, as well as therapeutic, applications designed to prevent the transmission of human fears.

Methods

Participants. One-hundred twenty-one participants [69 females, mean age = 26 y (experiment 1 = 26.7, experiment 2 = 25.9, experiment 3 = 25)] took part in the study and provided written consent. One participant was excluded due to a technical failure. Participants received two movie vouchers for their participation. In addition, 25 participants took part in a control experiment where decision making was not preceded by a conditioning block and 80 participants in control experiments 3B–C (see *SI Appendix* for details). All procedures were approved by the local ethics committee at Karolinska Institutet.

Task and Procedure. The experiments consisted of two blocks (Fig. 1). The content of the first block (conditioning) was the only difference between experiments 1–3 (see *SI Appendix* for details). In the subsequent decision making block (Fig. 1), which was identical for all experiments, the participants made 70 choices between the same two stimuli (i.e., CS+/CS−) that were presented during the conditioning block (see *SI Appendix*, Fig. S1, for an overview of the trial time line). The participants were asked to learn to avoid shocks, but were not given any information about the shock probabilities. The decision-making block consisted of two phases: a Transfer phase and a Reversal phase. The Transfer phase tested the direct effect of conditioning on decision making. We use the No Change condition as baseline for estimating the Pavlovian influence in the Change condition. Therefore, all comparisons reported in the main text are relative (in the *SI Appendix*, in addition we compare the results to the baseline control experiment). In the Reversal phase, which occurred unsignaled after half the experiment, the contingencies between choice stimuli and shocks reversed (Fig. 1 *D* and *E*). The participants were not informed about the contingency reversal.

Statistical Analyses. Logistic Generalized Linear Mixed Models (GLMMs) with by-subject random intercept and by-subject random slopes for all fixed effects were used to analyze the choice data (see *SI Appendix* for additional details). Data are available at https://osf.io/xfe72/.

ACKNOWLEDGMENTS. We thank Marc Guitart-Masip, Ida Selbing, and Philip Pärnamets for comments on an earlier version of the manuscript and Tove Hensler, Diana Diaz Del Castillo, and Erik Enström for help with data collection. This research was supported by Knut and Alice Wallenberg Foundation Grant KAW 2014.0237; Riksbankens Jubileumsfond Grant P11-1017; and Independent Starting Grant 284366 (Emotional Learning in Social Interaction project) from the European Research Council (to A.O.). B.L. was supported by Forte Grant COFAS2: 2014-2785 FOIP; A.G. was supported by Swedish Research Council Grant 2015-00312; and P.N.T. was supported by Swiss National Science Foundation Grants 100014_165884 and 100019_176016.

1. LeDoux JE (2017) Semantics, surplus meaning, and the science of fear. *Trends Cogn Sci* 21:303–306.
2. Beckers T, Krypotos A-M, Boddez Y, Effting M, Kindt M (2013) What's wrong with fear conditioning? *Biol Psychol* 92:90–96.
3. Hopwood TL, Schutte NS (2017) Psychological outcomes in reaction to media exposure to disasters and large-scale violence: A meta-analysis. *Psychol Violence* 7:316–327.
4. Holman EA, Garfin DR, Silver RC (2014) Media's role in broadcasting acute stress following the Boston Marathon bombings. *Proc Natl Acad Sci USA* 111:93–98.
5. Laland KN (2004) Social learning strategies. *Learn Behav* 32:4–14.
6. Lindström B, Selbing I, Olsson A (2016) Co-evolution of social learning and evolutionary preparedness in dangerous environments. *PLoS One* 11:e0160245.
7. Lindström B, Olsson A (2015) Mechanisms of social avoidance learning can explain the emergence of adaptive and arbitrary behavioral traditions in humans. *J Exp Psychol Gen* 144:688–703.
8. Cook M, Mineka S (1989) Observational conditioning of fear to fear-relevant versus fear-irrelevant stimuli in rhesus monkeys. *J Abnorm Psychol* 98:448–459.
9. Debiec J, Olsson A (2017) Social fear learning: From animal models to human function. *Trends Cogn Sci* 21:546–555.
10. Lindström B, Haaker J, Olsson A (2018) A common neural network differentially mediates direct and social fear learning. *Neuroimage* 167:121–129.
11. Olsson A, Phelps EA (2007) Social learning of fear. *Nat Neurosci* 10:1095–1102.
12. Haaker J, Golkar A, Selbing I, Olsson A (2017) Assessment of social transmission of threats in humans using observational fear conditioning. *Nat Protoc* 12:1378–1386.
13. Rachman S (1977) The conditioning theory of fear-acquisition: A critical examination. *Behav Res Ther* 15:375–387.
14. Askew C, Field AP (2008) The vicarious learning pathway to fear 40 years on. *Clin Psychol Rev* 28:1249–1265.
15. Cameron G, Roche B, Schlund MW, Dymond S (2016) Learned, instructed and observed pathways to fear and avoidance. *J Behav Ther Exp Psychiatry* 50:106–112.
16. Atlas LY, Phelps EA (2018) Prepared stimuli enhance aversive learning without weakening the impact of verbal instructions. *Learn Mem* 25:100–104.
17. Atlas LY, Doll BB, Li J, Daw ND, Phelps EA (2016) Instructed knowledge shapes feedback-driven aversive learning in striatum and orbitofrontal cortex, but not the amygdala. *eLife* 5:12964–12977.
18. Doll BB, Jacobs WJ, Sanfey AG, Frank MJ (2009) Instructional control of reinforcement learning: A behavioral and neurocomputational investigation. *Brain Res* 1299:74–94.
19. Bach DR, Dayan P (2017) Algorithms for survival: A comparative perspective on emotions. *Nat Rev Neurosci* 18:311–319.
20. Dayan P, Niv Y, Seymour B, Daw ND (2006) The misbehavior of value and the discipline of the will. *Neural Netw* 19:1153–1160.
21. Rangel A, Camerer C, Montague PR (2008) A framework for studying the neurobiology of value-based decision making. *Nat Rev Neurosci* 9:545–556.
22. Guitart-Masip M, Duzel E, Dolan R, Dayan P (2014) Action versus valence in decision making. *Trends Cogn Sci* 18:194–202.
23. Clark JJ, Hollon NG, Phillips PE (2012) Pavlovian valuation systems in learning and decision making. *Curr Opin Neurobiol* 22:1054–1061.
24. Rescorla RA, Solomon RL (1967) Two-process learning theory: Relationships between Pavlovian conditioning and instrumental learning. *Psychol Rev* 74:151–182.
25. Lewis AH, Niznikiewicz MA, Delamater AR, Delgado MR (2013) Avoidance-based human Pavlovian-to-instrumental transfer. *Eur J Neurosci* 38:3740–3748.
26. Talmi D, Seymour B, Dayan P, Dolan RJ (2008) Human Pavlovian-instrumental transfer. *J Neurosci* 28:360–368.
27. Cartoni E, Balleine B, Baldassarre G (2016) Appetitive Pavlovian-instrumental transfer: A review. *Neurosci Biobehav Rev* 71:829–848.
28. 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* 15:668–676.
29. Geurts DEM, Huys QJM, den Ouden HEM, Cools R (2013) Aversive Pavlovian control of instrumental behavior in humans. *J Cogn Neurosci* 25:1428–1441.
30. Huang T-R, Hazy TE, Herd SA, O'Reilly RC (2013) Assembling old tricks for new tasks: A neural model of instructional learning and control. *J Cogn Neurosci* 25:843–851.
31. Koban L, Jepma M, Geuter S, Wager TD (2017) What's in a word? How instructions, suggestions, and social information change pain and emotion. *Neurosci Biobehav Rev* 81:29–42.
32. Olsson A, Phelps EA (2004) Learned fear of "unseen" faces after Pavlovian, observational, and instructed fear. *Psychol Sci* 15:822–828.
33. Braem S, et al. (2017) Pattern analyses reveal separate experience-based fear memories in the human right amygdala. *J Neurosci* 37:8116–8130.
34. Biele G, Rieskamp J, Gonzalez R (2009) Computational models for the combination of advice and individual learning. *Cogn Sci* 33:206–242.
35. Li J, Delgado MR, Phelps EA (2011) How instructed knowledge modulates the neural systems of reward learning. *Proc Natl Acad Sci USA* 108:55–60.
36. Huys QJM, et al. (2011) Disentangling the roles of approach, activation and valence in instrumental and Pavlovian responding. *PLOS Comput Biol* 7:e1002028.
37. Biele G, Rieskamp J, Krugel LK, Heekeren HR (2011) The neural basis of following advice. *PLoS Biol* 9:e1001089.
38. Phelps EA, et al. (2001) Activation of the left amygdala to a cognitive representation of fear. *Nat Neurosci* 4:437–441.
39. Funayama ES, Grillon C, Davis M, Phelps EA (2001) A double dissociation in the affective modulation of startle in humans: Effects of unilateral temporal lobectomy. *J Cogn Neurosci* 13:721–729.