tapraid5/emo-emo/emo-emo/emo00315/emo3036d15z | xppws | S=1 | 4/7/15 | 4:24 | Art: 2014-0538

Emotion

1528-3542/15/$12.00 http://dx.doi.org/10.1037/emo0000075

A Clash of Values: Fear-Relevant Stimuli Can Enhance or Corrupt Adaptive Behavior Through Competition Between Pavlovian and Instrumental Valuation Systems

AQ: 1

AQ: au

Björn Lindström, Armita Golkar, and Andreas Olsson
Karolinska Institutet

AQ: 2

Humans and nonhuman primates preferentially learn to fear and avoid archetypical fear-relevant stimuli. Yet how these learning biases influence adaptive behavior, the basic mechanistic underpinnings of these biases, and how they interact with learning experiences during the life span of an individual remain unknown. To study this, we investigated how 4 classes of fear-relevant stimuli (snakes, threatening in-group faces, racial out-group faces, and guns) influenced adaptive behavior. We showed that stimulus-driven biases have a dramatic influence that can either promote or corrupt adaptive behavior depending on how a bias relates to the environment. We quantified and compared the effects of different fear-relevant stimuli on instrumental behavior using a computational reinforcement learning model that formalized the idea that the bias reflects competition between an instrumental and a Pavlovian valuation system. These results were further clarified by 2 independent rating studies showing that perceived danger of the stimuli corresponded well with their influence on adaptive behavior.

AQ: 3

Keywords: fear, preparedness, reinforcement learning, Pavlovian, instrumental learning

Supplemental materials: http://dx.doi.org/10.1037/emo0000075.supp

An erroneous decision in a dangerous environment might be an individual's last decision, providing strong selective pressure for optimizing adaptive behavior to avoid dangerous consequences. However, adaptive behavior is notoriously difficult to achieve and is error prone in the stochastic natural environment, which means that evolving optimal and maximally flexible behavioral systems is costly or even impossible (Haselton & Nettle, 2006; Johnson, Blumstein, Fowler, & Haselton, 2013). Past research has shown that behavior, despite these constraints, can be optimality approximated in many circumstances, partly as a result of ecologically valid shortcuts or biases resulting in adaptive behavior (Johnson et al., 2013). For example, many animal species have both innate mechanisms for recognizing and avoiding evolutionarily old predators (Smith, 1975; Veen, Richardson, Blaakmeer, & Komdeur, 2000) and learning mechanisms promoting avoidance of evolutionarily novel threats (Ferrari, Gonzalo, Messier, & Chivers, 2007). In humans, similar biases have been demonstrated in classical (stimulus–stimulus) conditioning in response to fear-relevant stimuli (Davey, 1995; Öhman & Mineka, 2001; Seligman, 1971). Both fear-relevant stimuli with a long evolutionary history (*phylogenetic*), such as snakes and threatening faces, and evolutionarily novel fear-relevant stimuli (*ontogenetic*), such as out-group faces (Navarrete et al., 2009; Olsson, Ebert, Banaji, & Phelps, 2005) and guns (Davey, 1995; Öhman & Mineka, 2001), are more easily and persistently associated with aversive stimuli than control stimuli (LeDoux, 2012; Öhman & Mineka, 2001). The development and the expression of these learning biases are thought to reflect a complex interaction of evolutionary and cultural factors (Davey, 1995). Although fear learning biases might have advantages, such biases can also be maladaptive for modern humans and their societies by supporting irrational fears and xenophobia (Öhman, 2005). Indeed, fears and phobias toward animals, such as snakes and spiders, are common in the general population and also in parts of the world where dangerous species are scarce or nonexistent (Öhman & Mineka, 2001; Seligman, 1971). However, because the vast majority of research investigating fear learning biases in humans has been based on classical conditioning, it remains unknown how fear-relevant stimuli affect adaptive instrumental behavior.

We investigated the consequences of a range of fear-relevant stimuli (snakes, threatening faces, out-group faces, guns) for adaptive behavior using a novel experimental model and computational analyses. A dominant perspective on the neurobiology of decision making posits that behavior can be understood as arising from the competition between different valuation systems. These independently assign value to stimuli and actions and can, therefore, have competing interests (Dayan, Niv, Seymour, & Daw, 2006; Huys et al., 2011; Rangel, Camerer, & Montague, 2008). Current evidence

AQ: 16 Björn Lindström, Armita Golkar, and Andreas Olsson, Division of Psychology, Department of Clinical Neuroscience, Karolinska Institutet.

This research was supported by a grant from the Swedish Research Council (Vetenskapsrådet; 421–2010–2084) and an Independent Starting Grant (Emotional Learning in Social Interaction project; 284366) from the European Research Council to Andreas Olsson. We thank Ida Selbing for valuable suggestions and help with the model fitting and Paulina Wiktor, Sofie Ährlund-Richter, Jonas Sjöström, and Simon Jangard Nielsen for assistance with data collection.

Correspondence concerning this article should be addressed to Björn Lindström, Karolinska Institutet, Nobels Väg 9, 17177, Stockholm, Sweden. E-mail: bjorn.lindstrom@ki.se

tapraid5/emo-emo/emo-emo/emo00315/emo3036d15z | xppws | S=1 | 4/7/15 | 4:24 | Art: 2014-0538

indicates that the human brain is composed of at least two such valuation systems: (a) a *Pavlovian* system that assigns value to a limited number of actions, such as avoidance, in response to biologically relevant stimuli, with relevance acquired through learning, innate predispositions, or a combination of the two (Dayan et al., 2006; Dayan & Seymour, 2007; Rangel et al., 2008) and (b) an *instrumental* system, which flexibly assigns value to actions on the basis of their reinforcement history together with current goals (Dayan et al., 2006; Rangel et al., 2008). A number of studies have investigated the interaction between these systems in humans (Geurts, Huys, den Ouden, & Cools, 2013; Guitart-Masip, Duzel, Dolan, & Dayan, 2014; Huys et al., 2011), but to date, very few, if any, have investigated the effect of fear-relevant stimuli on this interaction.

On the basis of this neurobiological perspective, we hypothesized that fear-relevant stimuli should trigger the Pavlovian valuation system and thereby bias adaptive instrumental behavior toward avoidance (LeDoux, 2012). The consequence of this bias should be directly contingent on the relationship between the fear-relevant stimulus and the environment: We hypothesized that fear-relevant stimuli would enhance adaptive behavior when they were reliable predictors of danger and corrupt adaptive behavior when they were unreliable predictors of danger. Thus, in environments where fear-relevant stimuli, on average, actually are dangerous, a stimulus-driven avoidance bias should be beneficial (Foster & Kokko, 2009; Johnson et al., 2013) and thus ecologically rational (Houston, McNamara, & Steer, 2007). However, if based on outdated, malignant, or otherwise unreliable information, such stimulus-driven biases might result in outcomes that are maladaptive on both the individual and the societal level, as is the case with phobias and racial biases. We conducted four instrumental avoidance learning experiments ($N = 156$) in which we varied the reliability of fear-relevant stimuli as predictors of danger to (a) test this hypothesis and (b) characterize the influence of four classes of fear-relevant stimuli—snakes, threatening in-group faces, social out-group (other race) faces, and guns—that have all been associated with learning biases in classical conditioning. These stimuli were also rated for perceived threat and danger by an additional 94 participants. To address the second issue (characterization of the influence of the stimuli) and provide insight into the underlying computations, we used a simple reinforcement learning (RL) model, implementing the idea of two competing valuation systems, to estimate the Pavlovian value of the different types of fear-relevant stimuli.

Method

The four experiments had identical experimental designs and are therefore, unless otherwise noted, described together throughout the Method and Results sections.

Participants

All participants were recruited at the Karolinska Institutet (Stockholm, Sweden) campus and provided written consent. Participants received two movie vouchers as compensation. All procedures were approved by the ethics committee at the Karolinska Institutet. Power calculations using G*Power 3.1.9.2 indicated that a power of .80 to detect the predicted between-participants and within-participant interactions, given a large predicted effect size (Cohen's $f \approx .45$–$.50$) would require ~40 participants. The first experiment (threatening vs. friendly faces) therefore involved 42 (22 female) participants (mean age = 26 years). Given the large effect size in this experiment, the second experiment (snakes vs. flowers) involved 32 (21 female) participants (mean age = 22 years), and the third experiment involved 34 (18 female) participants (mean age = 25 years). In both Experiments 2 and 3, we aimed for at least 30 participants. Any additional participants were included on the basis of economic and time constraints. In the fourth experiment (out-group vs. in-group faces), we predicted larger individual differences between participants and, therefore, a smaller average effect size. This experiment therefore included 48 (34 female) participants (mean age = 26 years). No participant took part in more than one experiment. The first ratings sample involved 59 (39 female) participants, and the second ratings sample involved 35 (26 female) participants. The rating studies were conducted in an undergraduate lecture setting (with first- and second-semester psychology students, respectively). The stopping rule was thereby determined by the number of volunteers attending the lectures.

Materials and Stimuli

All experiments were conducted in a sound-attenuated experimental chamber on a PC connected to a 19-in. CRT monitor. The aversive reinforcement was a monopolar 100-ms DC-pulse electric stimulation (STM200; BIOPAC Systems, Inc., Goleta, CA) applied to a participant's nondominant forearm. The intensity of the electric shock stimulation was adjusted individually for each participant in a work-up procedure on the basis of the criterion "unpleasant but not painful."

All stimuli were adjusted to fit a 462 pixel × 462 pixel frame. The stimuli for Experiment 1 were selected from the Karolinska Directed Emotional Faces (Lundqvist, Flykt, & Öhman, 1998) and consisted of two photos of the same model with happy and threatening facial expressions (model number BM17). The stimuli for Experiment 2 were selected from a stimuli set in which low-level features of the images (e.g., contrast, spatial frequency) were equalized (Wiens, Peira, Golkar, & Öhman, 2008) The stimuli for Experiment 4 were two male faces selected from the NimStim set (Actors 23 and 40) with equally neutral expressions according to the NimStim ratings (Tottenham et al., 2009). The stimuli used in Experiment 3 were grayscale images of a gun and the flower image used in Experiment 2. The specific gun exemplar was selected to be comparable to the most distinctly phylogenetic stimulus—the snake—on psychological dimensions likely to be relevant for behavior to facilitate comparison between stimulus categories. A rating study involving an independent sample of participants rated the stimuli used in Experiments 1–4 together with four images of different guns. The dimensions were threat and danger, rated on a seven-point scale ranging from 1 (*not at all*) to 7 (*very*). Finally, we asked participants how many individual negative experiences they had had in relation to what the images depicted, rated using a three-point scale (response options: *none*, *few*, *many*). The gun stimulus used for Experiment 3 was statistically comparable with the snake stimulus used in Experiment 2 for both rated threat and rated danger ($ps > .05$). In the second ratings study, participants were asked to estimate how much negatively valenced exposure

AQ: 4

AQ: 5

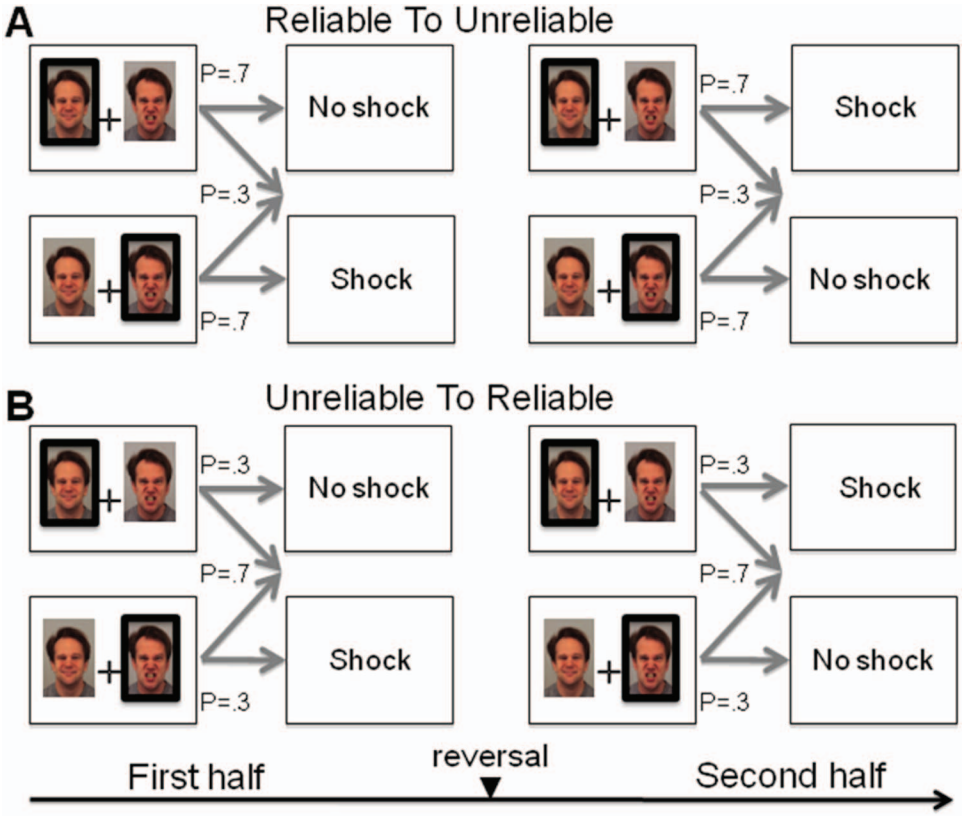
tapraid5/emo-emo/emo-emo/emo00315/emo3036d15z | xppws | S=1 | 4/7/15 | 4:24 | Art: 2014-0538

they had experienced from the different fear-relevant stimuli through (a) individual experience, (b) family and friends, (c) news media, and (d) popular culture. For this purpose, they used a 5-point scale encompassing 0 (*none*), 1–2 (*few*), 3–5 (*some*) 6–10 (*many*), and >10 (*very many*).

Task and Procedure

All four experiments had identical experimental designs, differing only in the stimuli presented (see Figure 1). The experimental task was a two-choice decision task with probabilistic aversive reinforcement (electric shocks). The participants were instructed to try to learn to avoid shocks and told that one choice stimulus might be better than the other for doing so. They were not informed about the contingency reversal. Participants performed six practice trials (no reinforcement) and then the experimental session (70 trials). They used the left and right arrow keys on a computer keyboard to indicate their choices. The location of the choice stimuli varied randomly between the left and right position across trials to prevent use of spatial-selection strategies. After the experimental session, participants in Experiment 2 completed the Snake Anxiety Questionnaire (SNAQ; Klorman, Weerts, Hastings, Melamed, & Lang, 1974), and participants in Experiment 4 completed the Modern Racism Scale (MRS) questionnaire (Akrami, Ekehammar, & Araya, 2000) and the Race Implicit Association Test (IAT). Unfortunately, we could not identify any validated questionnaires targeting fear/negative valuation of either threatening faces or guns. Participants in Experiment 3 rated the stimuli material (gun and flower) on perceived danger and threat and answered one question about previous personal negative experiences with the objects the images depicted.

As shown in Figure 1, the task had a mixed, 2 × 2 design, with one between-participants factor (predictive reliability: reliable to unreliable, unreliable to reliable) and one within-participant factor (phase: first half, second half). After the first half of the experiment (35 trials), the stimulus–reinforcement contingency reversed. For the reliable-to-unreliable groups, the fear-relevant stimulus was the more reliable predictor of shock (p[shock] = .70) in the first half and the less reliable predictor of shock (p[shock] = .30) in the second half, whereas the reverse was



ONLINE COLOR

Figure 1. Experimental design and procedures: Four experiments contrasted conditions in which the fear-relevant stimulus (here illustrated with a threatening face) was a reliable (p = .70) or unreliable (p = .30) predictor of danger (electric shock). Participants were randomly assigned to the reliable-to-unreliable or the unreliable-to-reliable group. After the first half of the experiment (35 trials), the stimulus–reinforcement contingences were reversed. These images are used with permission from the Karolinska Directed Emotional Faces stimuli set (Lundqvist, Flykt, & Öhman, 1998). The Karolinska Directed Emotional Faces - KDEF, CD ROM from Department of Clinical Neuroscence, Division of Psychology, Karolinska Institutet, ISBN 91-630-7164-9. See the online article for the color version of this figure.

tapraid5/emo-emo/emo-emo/emo00315/emo3036d15z | xppws | S=1 | 4/7/15 | 4:24 | Art: 2014-0538

true for the unreliable-to-reliable groups. This design allowed us to directly test our prediction about the interaction between the fear-relevant stimuli and the structure of the environment, control for the order effect of presentation, and disentangle transient and sustained effects of fear-relevant stimuli on adaptive behavior.

Statistical Analyses

Logistic generalized linear mixed models (GLMMs) with by-participant random intercepts and by-participant random slopes for the first-half/second-half factor (Baayen, Davidson, & Bates, 2008) were used to analyze the choice data. Logistic regression is the preferred statistical method when the dependent variable is binary (optimal = 1, suboptimal = 0; Jaeger, 2008). Reported main and interaction effects were evaluated with Type II analyses of deviance (analogous to Type II sum of squares analyses of variance [ANOVAs]) tests based on the Wald statistic, in which the goodness of fit of nested models were compared against a chi-square distribution (Fox & Weisberg, 2011). Degrees of freedom for the Wald test is the difference in the number of parameters between the compared models. For the GLMM analysis, we report the simple effect size (i.e., unstandardized beta estimate) in Table S1 in the online supplemental materials (OSMs). Presently, no consensus methodology for effect sizes in GLMM exists, but the simple effect size has the desirable property of being easily interpreted (Baguley, 2009). For 95% confidence intervals for all simple effects, see also Table 1 in the OSMs. For continuous variables, Cohen's d was used when appropriate. The computational RL models were fit to the data for each participant using maximum-likelihood estimation, and their respective goodnesses of fit were compared using the Akaike information criterion (AIC), which penalizes model complexity (see the OSMs for details). To account for individual variations in model fit, the AIC values were submitted to paired-samples t tests for model comparison (Fareri, Chang, & Delgado, 2012; Lindström, Selbing, Molapour, & Olsson, 2014; Suzuki et al., 2012). AQ: 6

Results

As shown in Figure 2, the predicted Pavlovian influence on adaptive behavior was confirmed by a pattern of crossover interactions in all four experiments (see Table S1 in the OSMs for the simple effect size estimates). These crossover interactions suggest that fear-relevant stimuli enhanced adaptive behavior when they were reliable predictors of danger (reliable conditions) but corrupted adaptive behavior when they were unreliable predictors of danger (unreliable conditions). This influence was predictably reversed after the contingency reversal, revealing the inflexible nature of the Pavlovian influence. Statistically, hierarchical logistic regressions of the trial-by-trial choices showed the predicted Phase (first half, second half) × Predictive Reliability (reliable to unreliable, unreliable to reliable) interaction for threatening versus happy facial expressions, $\chi^2(1) = 7.67$, $p < .01$; snake versus flower stimuli F2 AQ: 7

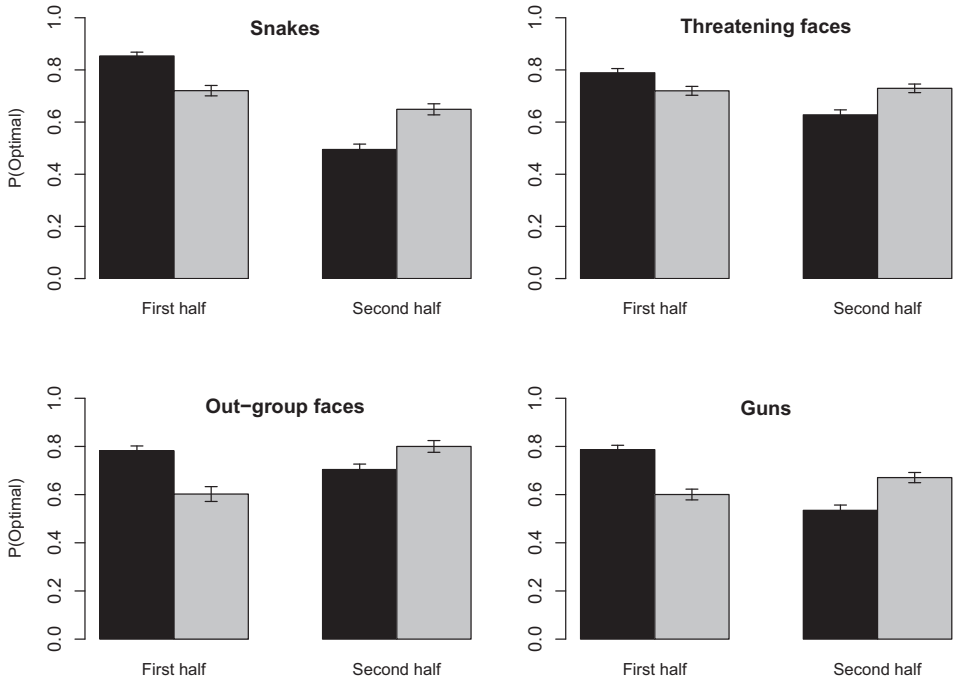


Figure 2. Probabilities of optimal behavior for the four classes of fear-relevant stimuli: snakes (interaction: $p <$ .01), threatening faces (interaction: $p < .01$), out-group faces for participants high in racial bias (interaction: $p <$ .05), and guns (interaction: $p < .01$). Error bars represent standard errors of the mean. See Figure S1 in the online supplemental materials for trial-by-trial plots. Black bars = reliable to unreliable, gray bars = unreliable to reliable.

Fn1 $\chi^2[1] = 9.39$, $p < .01$); and gun versus flower[1] stimuli, $\chi^2(1) = 12.21$, $p < .001$. For racial out-group faces, the pattern was slightly more complex. There was no overall effect of the experimental manipulation, but when a mean split of explicit racial bias was included in the analysis, the predicted interaction was observed for individuals with explicit racial bias above ($n = 20$), but not below ($n = 28$) the sample mean: Phase × Predictive Reliability × Racial Bias interaction, $\chi^2(1) = 3.91$, $p < .05$; see the Analysis of the Pavlovian Bias Elicited by Fear-Relevant Stimuli section for additional analyses of individual differences in Pavlovian valuation). Implicit racial bias (IAT) did not exhibit this pattern ($p = .18$; cf. Lindström et al., 2014). In contrast to individual differences in negative attitudes toward out-group members, individual differences in snake fear did not affect behavior ($p = .50$).

In conclusion, four independent experiments with, in total, 156 participants showed the predicted interaction between fear-relevant stimuli and the environment. The combined probability under the null hypothesis of observing this pattern of results (combining the p values of the four interaction terms using Fisher's method) was 2.849598e-07 (≈ 0.0000002849).

Characterizing the Pavlovian Influence of Fear-Relevant Stimuli

On the basis of the evidence for competition between valuation systems for behavioral output in decision making (Dayan et al., 2006; Huys et al., 2011; Rangel et al., 2008), we constructed a simple RL model to estimate the magnitude of the Pavlovian influence of the different fear-relevant stimuli (for further details, see the Computational Modeling section in the OSMs). Instrumental learning was modeled with the standard Q-learning (i.e., the Rescorla–Wagner) algorithm, where the expected value of each action (action value) depends on the reinforcement history tied to AQ:8 that action:

$$Q_{Fear-relevant}(t+1) = Q_{Fear-relevant}(t) + \alpha * (R_{Fear-relevant}(t) - Q_{Fear-relevant}(t)) \quad (1)$$

where a learning rate, α, determines how much the subsequent expected value at time point ($t + 1$) is affected by the prediction error, calculated as the difference between the actual value, $R_{\text{fear-relevant}(t)}$, and the action value, $Q_{\text{fear-relevant}}(t)$, of choosing the fear-relevant stimulus at time step t. We assumed that the fear-relevant stimulus (e.g., the snake) elicited a negative value that down-weighed the instrumental value of the action at the time of AQ: 9 choice and thereby reduced its probability:

$$P_{Fear-relevant(t)} = \frac{e^{(Q_{fear-relevant}(t) - Pavlovian_i)\beta}}{e^{(Q_{fear-relevant}(t) - Pavlovian_i)\beta} + e^{(Q_{neutral}(t)\beta}} \quad (2)$$

where $p_{\text{Fear-relevant}}$ is the probability of choosing the fear-relevant stimulus on trial (t); $Pavlovian_i$ is the Pavlovian influence, where the subscript i refers to the specific fear-relevant stimulus; and β is a noise parameter. This implementation, which we refer to as the *stimulus-bias* model, is similar to existing RL models of interaction between experimentally conditioned stimuli and instrumental learning (Dayan et al., 2006; Huys et al., 2011), and it parsimoniously accounts for how fear-relevant stimuli can enhance (when they are reliable predictors of danger) or corrupt (when they are unreliable predictors of danger) adaptive instrumental behavior (see Figure 2; Rangel et al., 2008). The qualitative pattern of predictions holds for a wide range of parameter values, because any positive value of the Pavlovian parameter will result in competition between the Pavlovian and instrumental valuation systems (see Figure 3). F3

We quantitatively confirmed that the Pavlovian parameter improved the fit to the data relative to a no bias model without this parameter, $t(152) = -2.28$, $p = .024$ (see Table S2 in the OSMs). Further, we sought to determine the stability of the Pavlovian influence on behavior—that is, whether the magnitude of the Pavlovian influence was modulated by the consequences of choosing the fear-relevant stimulus (*changing bias*; see the Alternative Models section and Table S2 in the OSMs). For example, the Pavlovian influence might have decreased if choosing fear-relevant stimulus did not result in an aversive shock. However, this model gave an impaired account of the data relative to the stimulus-bias model, $t(152) = -7.30$, $p < .001$, indicating that the Pavlovian bias was relatively constant within the time scale of the experiment.

To further test the stimulus-bias account of how fear-relevant stimuli influence adaptive behavior, we contrasted it with an alternative hypothesis derived from the literature on prepared learning in classical conditioning (Öhman & Mineka, 2001). A core idea in this literature is that the learning bias associated with fear-relevant stimuli (resistance to extinction and, often, enhanced acquisition of conditioned fear) does not reflect any intrinsic or prepotent valuation difference between stimuli but, rather, the expression of specific associations—that is, an enhanced capacity of fear-relevant stimuli to be associated with aversive events (Öhman & Mineka, 2001). We formulated two model versions of this learning-bias hypothesis (for equations, see the alternative models in the OSMs). Both models provided an impaired fit of the data relative to the stimulus-bias model—Learning Bias 1, $t(152) = -3.42$, $p < .001$; Learning Bias 2, $t(152) = -2.46$, $p = .01$ (see Table S2 in the OSMs)—suggesting that fear-relevant stimuli primarily influenced behavior through biasing action values during the decision process rather than the learning process.

Analysis of the Pavlovian Bias Elicited by Fear-Relevant Stimuli

Having established that the stimulus-bias model gives a satisfactory account of the data, we next analyzed the estimated Pavlovian parameter value from this model. The magnitude of this parameter constitutes a direct measure of the Pavlovian bias on behavior. A one-way ANOVA showed a main effect of fear-relevant stimulus type, $F(3, 147) = 3.16$, $p = .027$.[2] Simple- Fn2 effects analysis showed that the estimated Pavlovian parameter value was higher for guns than for out-group faces (β = 0.146,

[1] The snake and gun stimuli, representing the extremes of a hypothetical phylogenetic–ontogenetic continuum, were selected to be comparable in perceived threat and danger (see the Method section). To facilitate comparison, we used the same control stimulus in both experiments.

[2] Because statistical analysis of parameter estimates suffers from two sources of error—that is, from both the parameter fitting procedure and statistical error (Fareri et al., 2012)—we removed two extreme outliers (with standardized residuals exceeding 3 standard deviations; Baayen et al., 2008).

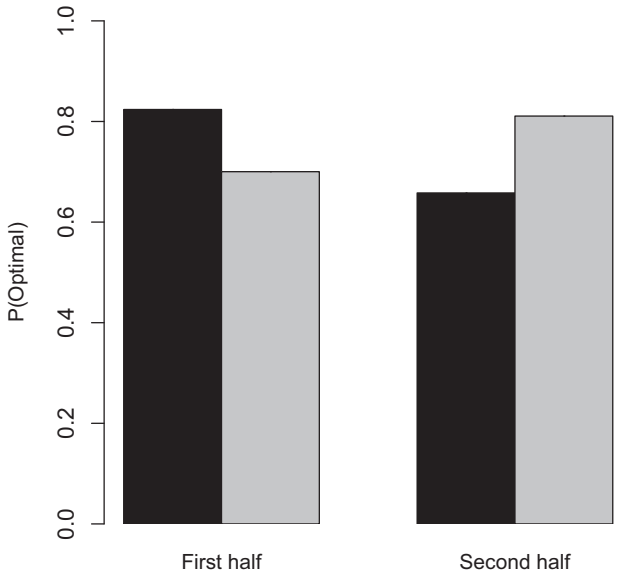
tapraid5/emo-emo/emo-emo/emo00315/emo3036d15z | xppws | S=1 | 4/7/15 | 4:24 | Art: 2014-0538

AQ:10, 11 $SE = 0.057$), $t(147) = 2.550$, $p = .012$, 95% confidence interval (CI) [0.034, 0.258], and marginally higher for guns relative to threatening faces ($\beta = 0.104$, $SE = 0.060$), $t(147) = 1.737$, $p = .084$, 95% CI [−0.014, 0.222]. The Pavlovian value of snakes was also higher than the value for out-group faces ($\beta = 0.1390$, $SE = 0.0578$), $t(147) = 2.405$, $p = .017$, 95% CI [0.026, 0.252], but not that for threatening faces ($\beta = 0.096$, $SE = 0.060$), $t(147) = 1.603$, $p = .111$, 95% CI [−0.022, 0.214]. Threatening faces and out-group faces did not significantly differ in the estimated Pavlovian value ($p = .43$), nor did snakes and guns ($p = .90$).

The direct analysis of the choice data indicated that the effect of out-group faces varied with a measure of a participant's explicit racial bias. Consistent with this pattern, racial bias was positively correlated with the estimated Pavlovian value of out-group faces AQ: 12 ($r = .30$, $p = .037$, 95% CI [.017, .538]). In contrast, self-reported snake fear (SNAQ) was not reliably correlated with the estimated Pavlovian value of snakes ($r = .15$, $p = .40$, 95% CI [−.210, .474]). The average level of snake fear in the sample was low ($M = 7.0$, $SD = 5.0$) relative to the range of the scale (0–20; Klorman et al., 1974), indicating that snakes affect voluntary instrumental behavior in the absence of direct fearfulness. Unfortunately, we could not identify any validated scales in Swedish targeting fear/negative attitudes toward threatening faces or guns, so the relationship between individual differences in attitudes and Fn3 the Pavlovian parameter remains unknown for these stimuli.[3] Analysis of the other model parameters revealed no systematic differences (see the Additional analysis of parameter estimates section in the OSMs).

Ratings of Threat, Danger, and Sources of Negative Exposure

The analysis of the estimated Pavlovian parameter showed that the fear-relevant stimuli differed in their impact on adaptive behavior (see Figure 4). It is clear that these differences cannot correspond to differences in phylogenetic status, because the Pavlovian value did not differ between snakes and guns. In an effort to clarify other possible causes for the differences in the Pavlovian influence, we performed two rating studies with two independent samples (the first rating sample was also used to select the gun stimulus for Experiment 4 [see the Method section]). The goal was to provide independent estimates of perceived threat and dangerousness for the fear-relevant stimuli used in Experiments 1–4 and estimates of their typical associated learning histories. Our use of independent samples had the benefit that the ratings of the stimuli were unrelated to the experimental manipulations. At the same time, comparisons between the ratings and the experimental results could only be qualitative in nature.



AQ:17, 18 *Figure 3.* Simulated results from the stimulus-bias model, which implemented competing Pavlovian and instrumental valuation systems ($\alpha = .50$, $\beta = 0.50$, Pavlovian parameter = 0.20). Black bars = reliable to unreliable; gray bars = unreliable to reliable.

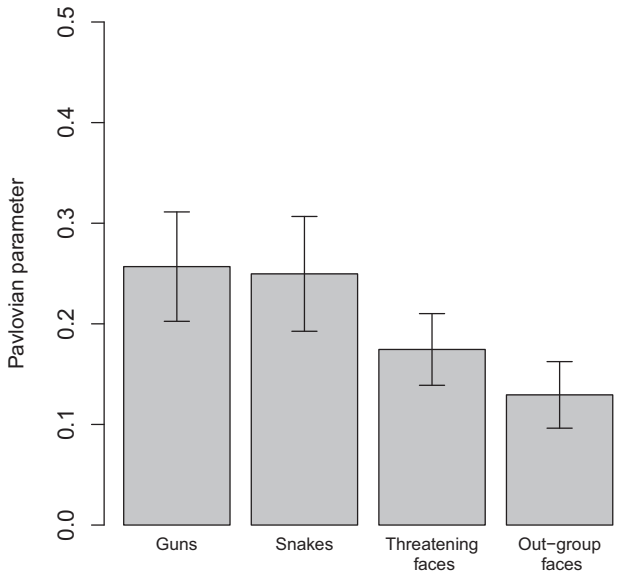


Figure 4. Magnitudes of the estimated Pavlovian parameter from the independent-systems model. The error bars represent standard errors of the mean.

The rated threat and danger in the first sample is shown in Figure 5. For the threat ratings (per paired t tests), only out-group F5 faces differed from the other fear-relevant stimuli ($ts > 8.00$, $ps < .001$, $ds > 0.96$). For the danger ratings, all comparisons, except that between guns and snakes, $t(56) = 1.84$, $p = .07$,[4] were highly Fn4 significant ($ps < .001$), with all effect sizes above 0.68. The effect

[3] We asked the participants in Experiment 4 to rate the gun stimulus for perceived threat and dangerousness. These ratings did not correlate with the estimated Pavlovian parameter ($p > .05$). However, because the ratings questionnaire was administered after the experiment and depicted the same stimulus that was used in the experiment, any potential correlation might be obscured by the aversive experiences of the experiment (i.e., a conditioning effect). Unfortunately, this appears to have been the case: The number of electric shocks received during the experiment correlated with the ratings of perceived dangerousness ($r = .420$, $p = .012$). This was not the case for the validated snake fear ($r = -.156$, $p = .39$) or racial bias questionnaires ($r = .12$, $p = .41$).

[4] Note that the specific gun stimulus used in Experiment 4 was selected on the basis of these criteria. We included three other gun exemplars in the first ratings study, two of which were rated as significantly more dangerous than the snake stimulus.

tapraid5/emo-emo/emo-emo/emo00315/emo3036d15z | xppws | S=1 | 4/7/15 | 4:24 | Art: 2014-0538

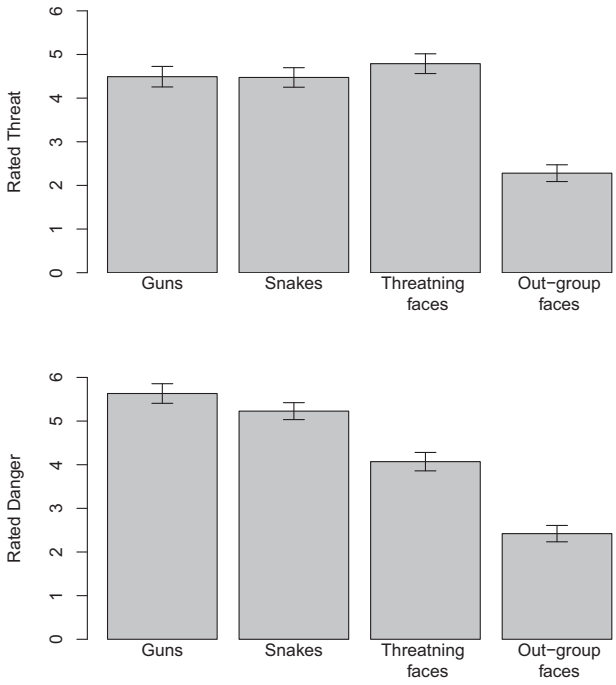


Figure 5. Threat and danger ratings for different fear-relevant stimuli. Error bars represent standard errors of the mean.

sizes of the difference between snakes and out-group faces (d = 1.49) and between guns and out-group faces (d = 1.52) were particularly large. Although the ratings of threat and danger were correlated, the correlation was not perfect ($rs < .64$, $r^2 < .41$), indicating that the two questions captured different constructs. The danger ratings (see Figure 5) appear to correspond most closely to the qualitative ordering of the estimated Pavlovian values from Experiment 1–4 (see Figure 4). Notably, the qualitative pattern of reported negative experiences with the fear-relevant stimuli types did not correspond to either the threat or the danger ratings.[5] Threatening faces were associated with the most negative experiences (*Mdn* = *some*), whereas reported negative experiences with guns, snakes, and out-group faces were very rare (*Mdn* = *none*). These results suggest that individual (direct) classical conditioning is unlikely to have underlain the influence of fear-relevant stimuli of behavior in our experiments.

Our second ratings study aimed to characterize the degree of negative exposure to the fear-relevant stimuli from other sources as opposed to individual experience. We asked the participants to estimate how much negatively valenced exposure they had experienced to the different fear-relevant stimuli through (a) individual experience, (b) family and friends, (c) news media, and (d) popular culture. These results are summarized in Figure S2 in the OSMs. In summary, the participants reported both the most individual exposure and exposure through family/friends to threatening faces, and secondarily to snakes. In contrast, for both news media and popular media, guns were associated with the most negative exposure, threatening faces and out-group faces with intermediate levels, and the least negative mass media exposure was reported for snakes. Together, these results suggest that different routes of learning might underlie the impact of the different fear-relevant stimuli.

Discussion

Across four independent experiments, we have, for the first time, demonstrated that fear-relevant stimuli have a strong influence on adaptive instrumental behavior. More specifically, this influence took the form of a Pavlovian stimulus-driven bias whose impact was critically dependent on the environment: When fear-relevant stimuli were reliable predictors of danger, adaptive behavior was enhanced, but when they were unreliable predictors of danger, adaptive behavior was corrupted. This pattern was consistent across all four types of fear-relevant stimuli: snakes, threatening faces, guns, and out-group faces. These results show that both archetypically phylogenetic (e.g., snakes) and ontogenetic (e.g., guns) fear-relevant stimuli similarly affect adaptive behavior.

The Pavlovian bias induced by fear-relevant stimuli appears to be learning independent (within the time scale of the experiment) and automatically triggered, as shown by the shift in the direction of the Pavlovian bias after the contingency reversal in our experiments (see Figure 2) and the impaired statistical fit of a RL model (changing-bias model), in which the bias could change over time. This view is consistent with the classical animal literature, which shows the extreme difficulties of learning instrumental contingencies that clash with the intrinsic Pavlovian value of an action (Breland & Breland, 1961). For example, in one experiment, squirrel monkeys were punished by electric shocks for pulling a leash that restrained them (Morse, Mead, & Kelleher, 1967). The optimal instrumental action in this situation would have been to remain still. Instead, the monkeys increased their leash-pulling behavior, presumably because the Pavlovian system automatically assigned value to avoidance in the domain of danger (Dayan et al., 2006). Such findings conceptually mirror our results, in which the mere presentation of fear-relevant stimuli corrupted adaptive behavior when the Pavlovian value and the optimal instrumental action were incompatible. Our results are also consistent with those of recent studies of how experimentally conditioned aversive stimuli inhibit instrumental approach behavior (Geurts et al., 2013; Huys et al., 2011), which have been modeled as competing valuation systems in RL models (Dayan et al., 2006; Guitart-Masip et al., 2012, 2014; Huys et al., 2011). It is, however, important to note that our general conclusion that fear-relevant stimuli bias instrumental behavior is not contingent on the details, or even validity, of the stimulus-bias model. Indeed, the described bias is evident as a robust impact on adaptive behavior across the four experiments and stimuli categories (see Figure 2).

Previous research has shown more persistent fear conditioning to both snakes and guns relative to control stimuli (Öhman & Mineka, 2001); to threatening, relative to neutral, in-group faces (Öhman & Mineka, 2001); and to racial out-group, relative to racial in-group, faces (Navarrete et al., 2009; Olsson et al., 2005). The present results provide an important extension of this literature by showing that the negative value intrinsic to these stimuli directly affects adaptive behavior (cf. Lindström et al., 2014). By showing the behavioral consequences of fear-relevant stimuli, our results lend support to the previous suggestion that Pavlovian biases have significant, potentially detrimental, consequences for

[5] The rating scale consisted of three qualitative scale steps (*none*, *some*, *many*). Therefore, the median is used as the estimate of central tendency.

tapraid5/emo-emo/emo-emo/emo00315/emo3036d15z | xppws | S=1 | 4/7/15 | 4:24 | Art: 2014-0538

humans (Dayan et al., 2006; Rangel et al., 2008). Our experimental approach was distinct from those of other studies of Pavlovian influences on instrumental behavior: By using the fear-relevant stimuli as choice stimuli rather than displayed in the background of an instrumental task, as in recent Pavlovian-to-instrumental transfer studies (Geurts et al., 2013; Huys et al., 2011), we have produced results that are directly relevant to real-world situations in which the same stimuli carry both Pavlovian and instrumental value. For example, in many social contexts, Pavlovian value and instrumental value co-occur, such as when one interacts with someone with facial features that trigger the Pavlovian system but the liking of whom is instrumentally valuable. Moreover, our experimental design was also distinct from those of studies showing Pavlovian biasing of action (approach and avoidance) tendencies (Chiu, Cools, & Aron, 2014; Guitart-Masip et al., 2014; Ly, Huys, Stins, Roelofs, & Cools, 2014), because using the fear-relevant stimulus as a choice stimulus precluded biased action or motor tendencies as the mechanism underlying the results. Instead, our experimental and modeling results suggest competition between the Pavlovian and instrumental systems at the decision stage via a Pavlovian bias of the action values associated with fear-relevant stimuli.

The fact that guns exerted the same influence as snakes on adaptive behavior clearly demonstrates that evolutionary predispositions are not necessary to explain the present results. Instead, the pattern of estimated Pavlovian values of the stimulus types matched their perceived dangerousness as rated by an independent sample. Perceived dangerousness or deadliness has been suggested as the factor underlying both phylogenetic and ontogenetic fear-relevant stimuli, thought to result in fear learning biases through an enhanced expectancy for associated negative events (Davey, 1995). What then determines perceived deadliness? A classic account of the etiology of human fears suggested that multiple pathways jointly contribute to the acquisition of fears (Rachman, 1977). These pathways are direct experience—observation (or vicarious) and verbal transmission of threat information. The effectiveness of the last two pathways has been repeatedly demonstrated in both children and adults (Askew & Field, 2008; Olsson & Phelps, 2007; Phelps et al., 2001). Our second ratings study, in which we asked participants to estimate the amount of negatively valenced exposure from different sources to the four types of fear-relevant stimuli, provides qualitative support for the idea that the Pavlovian value of all stimuli, except threatening faces, is at least partially acquired socially through observational and/or verbal pathways. Individual differences in negative exposure to fear-relevant stimuli would, from this learning-focused perspective, be expected to be expressed in corresponding differences in Pavlovian valuations. The correlation between individual differences in racial bias (presumably reflecting differences in negative exposure) and the estimated Pavlovian value of out-group faces provides preliminary support for this idea, but the lack of a similar correlation for snakes hints that other, as yet unknown, factors may mediate the relationship between individual differences and behavior.

In conclusion, we asked how fear-relevant stimuli influence adaptive behavior and how the computational basis of this influence was constituted. We demonstrated that fear-relevant stimuli can corrupt adaptive behavior in environments in which they are unreliable predictors of danger and enhance adaptive behavior when they are reliable predictors of danger. Further, phylogenetic and ontogenetic fear-relevant stimuli exerted a comparable influence, underscoring the critical role of learning in adaptive behavior. Together, these results shed new light on how emotional biases affect behavior by focusing on the interplay between behavior and the environment (Houston et al., 2007). These findings have implications for understanding how behavioral biases and phobias affect adaptive voluntary behavior.

References

Akrami, N., Ekehammar, B., & Araya, T. (2000). Classical and modern racial prejudice: A study of attitudes toward immigrants in Sweden. *European Journal of Social Psychology, 30,* 521–532.

Askew, C., & Field, A. P. (2008). The vicarious learning pathway to fear 40 years on. *Clinical Psychology Review, 28,* 1249–1265. http://dx.doi.org/10.1016/j.cpr.2008.05.003

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59,* 390–412. http://dx.doi.org/10.1016/j.jml.2007.12.005

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology, 100,* 603–617. http://dx.doi.org/10.1348/000712608X377117

Breland, K., & Breland, M. (1961). The misbehavior of organisms. *American Psychologist, 16,* 681–684. http://dx.doi.org/10.1037/h0040090

Chiu, Y.-C., Cools, R., & Aron, A. R. (2014). Opposing effects of appetitive and aversive cues on go/no-go behavior and motor excitability. *Journal of Cognitive Neuroscience, 26,* 1851–1860. http://dx.doi.org/10.1162/jocn_a_00585

Davey, G. C. L. (1995). Preparedness and phobias: Specific evolved associations or a generalized expectancy bias? *Behavioral and Brain Sciences, 18,* 289–297. http://dx.doi.org/10.1017/S0140525X00038498

Dayan, P., Niv, Y., Seymour, B., & Daw, N. D. (2006). The misbehavior of value and the discipline of the will. *Neural Networks, 19,* 1153–1160. http://dx.doi.org/10.1016/j.neunet.2006.03.002

Dayan, P., & Seymour, B. (2007). Values and actions in aversion. In P. W. Glimcher, C. F. Camerer, E. Fehr, & R. A. Poldrack (Eds.), *Neuroeconomics: Decision making and the brain* (pp. 175–192). Waltham, MA: Academic Press.

Fareri, D. S., Chang, L. J., & Delgado, M. R. (2012). Effects of direct social experience on trust decisions and neural reward circuitry. *Frontiers in Neuroscience, 6,* 148. http://dx.doi.org/10.3389/fnins.2012.00148

Ferrari, M. C. O., Gonzalo, A., Messier, F., & Chivers, D. P. (2007). Generalization of learned predator recognition: An experimental test and framework for future studies. *Proceedings of the Royal Society B: Biological Sciences, 274,* 1853–1859.

Foster, K. R., & Kokko, H. (2009). The evolution of superstitious and superstition-like behaviour. *Proceedings of the Royal Society B: Biological Sciences, 276,* 31–37. http://dx.doi.org/10.1098/rspb.2008.0981

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression*. Thousand Oaks, CA: Sage.

Geurts, D. E. M., Huys, Q. J. M., den Ouden, H. E. M., & Cools, R. (2013). Aversive Pavlovian control of instrumental behavior in humans. *Journal of Cognitive Neuroscience, 25,* 1428–1441.

Guitart-Masip, M., Duzel, E., Dolan, R., & Dayan, P. (2014). Action versus valence in decision making. *Trends in Cognitive Sciences, 18,* 194–202. http://dx.doi.org/10.1016/j.tics.2014.01.003

Guitart-Masip, M., Huys, Q. J. M., Fuentemilla, L., Dayan, P., Duzel, E., & Dolan, R. J. (2012). Go and no-go learning in reward and punishment: Interactions between affect and effect. *NeuroImage, 62,* 154–166. http://dx.doi.org/10.1016/j.neuroimage.2012.04.024

tapraid5/emo-emo/emo-emo/emo00315/emo3036d15z | xppws | S=1 | 4/7/15 | 4:24 | Art: 2014-0538

Haselton, M. G., & Nettle, D. (2006). The paranoid optimist: An integrative evolutionary model of cognitive biases. *Personality and Social Psychology Review, 10,* 47–66.

Houston, A. I., McNamara, J. M., & Steer, M. D. (2007). Do we expect natural selection to produce rational behaviour? *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences, 362,* 1531–1543.

Huys, Q. J. M., Cools, R., Gölzer, M., Friedel, E., Heinz, A., Dolan, R. J., & Dayan, P. (2011). Disentangling the roles of approach, activation and valence in instrumental and Pavlovian responding. *PLoS Computational Biology, 7*(4), e1002028. http://dx.doi.org/10.1371/journal.pcbi.1002028

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59,* 434–446. http://dx.doi.org/10.1016/j.jml.2007.11.007

Johnson, D. D. P., Blumstein, D. T., Fowler, J. H., & Haselton, M. G. (2013). The evolution of error: Error management, cognitive constraints, and adaptive decision-making biases. *Trends in Ecology & Evolution, 28,* 474–481. http://dx.doi.org/10.1016/j.tree.2013.05.014

Klorman, R., Weerts, T. C., Hastings, J. E., Melamed, B. G., & Lang, P. J. (1974). Psychometric description of some specific-fear questionnaires. *Behavior Therapy, 5,* 401–409.

LeDoux, J. (2012). Rethinking the emotional brain. *Neuron, 73,* 653–676. http://dx.doi.org/10.1016/j.neuron.2012.02.004

Lindström, B., Selbing, I., Molapour, T., & Olsson, A. (2014). Racial bias shapes social reinforcement learning. *Psychological Science, 25,* 711–719. http://dx.doi.org/10.1177/0956797613514093

Lundqvist, D., Flykt, A., & Öhman, A. (1998). *The Karolinska Directed Emotional Faces—KDEF*. Stockholm, Sweden: Karolinska Institutet, Department of Clinical Neuroscience, Psychology Section.

Ly, V., Huys, Q. J. M., Stins, J. F., Roelofs, K., & Cools, R. (2014). Individual differences in bodily freezing predict emotional biases in decision making. *Frontiers in Behavioral Neuroscience, 8,* 237. http://dx.doi.org/10.3389/fnbeh.2014.00237

Morse, W. H., Mead, R. N., & Kelleher, R. T. (1967, July 14). Modulation of elicited behavior by a fixed-interval schedule of electric shock presentation. *Science, 157,* 215–217. http://dx.doi.org/10.1126/science.157.3785.215

Navarrete, C. D., Olsson, A., Ho, A. K., Mendes, W. B., Thomsen, L., & Sidanius, J. (2009). Fear extinction to an out-group face: The role of target gender. *Psychological Science, 20,* 155–158. http://dx.doi.org/10.1111/j.1467-9280.2009.02273.x

Öhman, A. (2005, July 29). Conditioned fear of a face: A prelude to ethnic enmity? *Science, 309,* 711–713. http://dx.doi.org/10.1126/science.1116710

Öhman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review, 108,* 483–522. http://dx.doi.org/10.1037/0033-295X.108.3.483

Olsson, A., Ebert, J. P., Banaji, M. R., & Phelps, E. A. (2005, July 29). The role of social groups in the persistence of learned fear. *Science, 309,* 785–787. http://dx.doi.org/10.1126/science.1113551

Olsson, A., & Phelps, E. A. (2007). Social learning of fear. *Nature Neuroscience, 10,* 1095–1102. http://dx.doi.org/10.1038/nn1968

Phelps, E. A., O'Connor, K. J., Gatenby, J. C., Gore, J. C., Grillon, C., & Davis, M. (2001). Activation of the left amygdala to a cognitive representation of fear. *Nature Neuroscience, 4,* 437–441. http://dx.doi.org/10.1038/86110

Rachman, S. (1977). The conditioning theory of fear-acquisition a critical examination. *Behaviour Research and Therapy, 15,* 375–387.

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience, 9,* 545–556. http://dx.doi.org/10.1038/nrn2357

Seligman, M. E. P. (1971). Phobias and preparedness. *Behavior Therapy, 2,* 307–320. http://dx.doi.org/10.1016/S0005-7894(71)80064-3

Smith, S. M. (1975, February 28). Innate recognition of coral snake pattern by a possible avian predator. *Science, 187,* 759–760. http://dx.doi.org/10.1126/science.187.4178.759

Suzuki, S., Harasawa, N., Ueno, K., Gardner, J. L., Ichinohe, N., Haruno, M., . . . Nakahara, H. (2012). Learning to simulate others' decisions. *Neuron, 74,* 1125–1137. http://dx.doi.org/10.1016/j.neuron.2012.04.030

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., . . . Nelson, C. (2009). The NimStim set of facial expressions: Judgments from untrained research participants. *Psychiatry Research, 168,* 242–249. http://dx.doi.org/10.1016/j.psychres.2008.05.006

Veen, T., Richardson, D. S., Blaakmeer, K., & Komdeur, J. (2000). Experimental evidence for innate predator recognition in the Seychelles warbler. *Proceedings of the Royal Society B: Biological Sciences, 267,* 2253–2258. http://dx.doi.org/10.1098/rspb.2000.1276

Wiens, S., Peira, N., Golkar, A., & Öhman, A. (2008). Recognizing masked threat: Fear betrays, but disgust you can trust. *Emotion, 8,* 810–819. http://dx.doi.org/10.1037/a0013731

Received October 16, 2014
Revision received February 26, 2015
Accepted March 2, 2015 ■