Enhanced Pavlovian aversive conditioning to positive emotional stimuli

Stussi, Yoann; Pourtois, Gilles; Sander, David

How to cite

STUSSI, Yoann, POURTOIS, Gilles, SANDER, David. Enhanced Pavlovian aversive conditioning to positive emotional stimuli. In: Journal of Experimental Psychology: General, 2018, vol. 147, n° 6, p. 905–923. doi: 10.1037/xge0000424

This publication URL: https://archive-ouverte.unige.ch/unige:106271

Publication DOI: 10.1037/xge0000424

In press, *Journal of Experimental Psychology: General*

ENHANCED PAVLOVIAN AVERSIVE CONDITIONING

TO POSITIVE EMOTIONAL STIMULI

Yoann Stussi

University of Geneva

Gilles Pourtois

Ghent University

David Sander

University of Geneva

Word count: 13274

Corresponding Author:

Yoann Stussi, Campus Biotech, CISA – University of Geneva,

Chemin des Mines 9, CH-1202 Geneva

Phone: +41 (0)22 379 09 06

E-mail: yoann.stussi@unige.ch

Author Note

Yoann Stussi, Swiss Center for Affective Sciences, Campus Biotech, University of Geneva, and Laboratory for the study of Emotion Elicitation and Expression, Department of Psychology, University of Geneva; Gilles Pourtois, Cognitive & Affective Psychophysiology Laboratory, Department of Experimental Clinical & Health Psychology, Ghent University; David Sander, Swiss Center for Affective Sciences, Campus Biotech, University of Geneva, and Laboratory for the study of Emotion Elicitation and Expression, Department of Psychology, University of Geneva.

Portions of the results reported in this article were presented at the Consortium of European Research on Emotion conference, Leuven, the Netherlands, July 2016; at the Ninth European Meeting on Human Fear Conditioning, Hamburg, Germany, May 2017; and at the Fifteenth Swiss Psychological Society conference, Lausanne, Switzerland, September 2017. This research was supported by the National Center of Competence in Research (NCCR) Affective Sciences, financed by the Swiss National Science Foundation (51NF40-104897), and hosted by the University of Geneva, and by a Doc.CH grant from the Swiss National Science Foundation (P0GEP1_159057) to Yoann Stussi. We thank Aude Ferrero and Alessio Giarrizzo for their assistance with data collection, as well as Sylvain Delplanque, Eva Pool, and Vanessa Sennwald for insightful comments on this article. The data reported in this paper are available on the Open Science Framework (https://doi.org/10.17605/OSF.IO/2FYHQ).

Correspondence concerning this article should be addressed to Yoann Stussi or David Sander, Campus Biotech, CISA – University of Geneva, Chemin des Mines 9, CH-1202 Geneva, Switzerland. E-mail: yoann.stussi@unige.ch or david.sander@unige.ch.

Abstract

Pavlovian aversive conditioning is an evolutionarily well-conserved adaptation enabling organisms to learn to associate environmental stimuli with biologically aversive events. However, mechanisms underlying preferential (or enhanced) Pavlovian aversive conditioning remain unclear. Previous research has suggested that only specific stimuli that have threatened survival across evolution (e.g., snakes and angry faces) are preferentially conditioned to threat. Here, we challenge this view by showing that positive stimuli with biological relevance (baby faces and erotic stimuli) are likewise readily associated with an aversive event (electric stimulation) during Pavlovian aversive conditioning, thereby reflecting a learning bias to these stimuli. Across three experiments, our results reveal an enhanced persistence of the conditioned response to both threat-relevant and positive relevant stimuli compared with the conditioned response to neutral stimuli. These findings support the existence of a general mechanism underlying preferential Pavlovian aversive conditioning that is shared across negative and positive stimuli with high relevance to the organism, and provide new insights into the basic mechanisms underlying emotional learning in humans.

Keywords: Pavlovian conditioning; Emotion; Learning; Positive stimuli; Relevance detection

Enhanced Pavlovian aversive conditioning to positive emotional stimuli

In Pavlovian conditioning, a conditioned stimulus acquires a predictive and emotional value through a single or repeated contingent pairing with a biologically potent stimulus. This learning process represents a fundamental evolutionarily well-conserved adaptation enabling organisms to predict and detect stimuli in the environment, and shape appropriate responses to them. Pavlovian conditioning has substantially contributed to our understanding of the psychological and neurobiological underpinnings of learning, memory, and emotion (e.g., Büchel, Morris, Dolan, & Friston, 1998; LaBar & Cabeza, 2006; LaBar, Gatenby, Gore, LeDoux, & Phelps, 1998; LeDoux, 2000, 2012, 2014; Phelps, Delgado, Nearing, & LeDoux, 2004; Phelps & LeDoux, 2005; Schiller et al., 2010). Research on Pavlovian conditioning has essentially focused on unveiling the general principles of learning (Pavlov, 1927), delineating in particular the central role of prediction error (i.e., the discrepancy between the predicted and the actual outcome) and stimulus' associability (i.e., the degree to which the stimulus reliably predicts and easily enters into association with the outcome) in associative learning (see, e.g., Niv & Schoenbaum, 2008; Pearce & Hall, 1980; Rescorla & Wagner, 1972). However, this line of research has generally omitted to consider the relative importance of the stimuli at stake for the organism. Apart from this trend, preparedness theory (Seligman, 1970, 1971) posits that certain classes of "evolutionarily prepared" threat stimuli are preferentially associated with aversive events based on biological predispositions shaped by evolution. Consistent with this view, a series of empirical studies have shown that evolutionary threat-relevant stimuli – such as snakes, angry faces, or outgroup faces – are more readily associated with an aversive outcome than threat-irrelevant stimuli – such as flowers, happy faces, or ingroup faces (e.g., Öhman & Dimberg, 1978; Öhman, Fredrikson, Hugdahl, & Rimmö, 1976; Öhman & Mineka, 2001;

Olsson, Ebert, Banaji, & Phelps, 2005; but see Mallan, Lipp, & Cochrane, 2013, for a review of evidence showing that threat conditioned to social threat-relevant stimuli is more malleable than threat conditioned to animal threat-relevant stimuli). Extending preparedness theory, Öhman and Mineka (2001) proposed the existence of an evolved fear module centered on the amygdala in the human brain dedicated to processing threat-relevant stimuli from phylogenetic origin, thus subserving the preferential processing of, and the learning bias to, evolutionarily prepared threat stimuli.

In contrast, we suggest that preferential emotional learning is not specific to threat-related stimuli but extends to all stimuli that are relevant to the organism's concerns (Frijda, 1988). This alternative model holds that such preferential learning is driven by a general mechanism of relevance detection that is not specific to threat. Relevance detection is conceptualized as a rapid process, which enables the organism to detect and continuously appraise stimuli as a function of their affective relevance in relation to the organism's concerns (Pool, Brosch, Delplanque, & Sander, 2016; Sander, Grafman, & Zalla, 2003; Sander, Grandjean, & Scherer, 2005). A stimulus is therefore detected and appraised as relevant if "it increases the probability of satisfaction or dissatisfaction toward a major concern of the individual" (Sander, 2013, p. 22). Concerns refer to affective representations of psychological and physiological motives, needs, goals, and values that are of major importance to the organism (Frijda, 1988; Pool, Brosch, et al., 2016). According to this model, phylogenetically threat-relevant stimuli lead to preferential processing and learning because they are highly relevant to the organism's survival. More specifically, the relevance detection hypothesis predicts that stimuli detected as relevant to the organism benefit from enhanced processing (Brosch, Sander, Pourtois, & Scherer, 2008; Pool, Brosch, et al., 2016) and preferential learning regardless of their valence. If the organism does preferentially learn associations involving highly relevant stimuli irrespective of their valence, this implies – even if it might seem counterintuitive – that positive stimuli with high relevance to the organism should be likewise readily associated with an aversive outcome, as is the case for threat-relevant stimuli.

Here, we therefore assessed whether positive relevant stimuli are readily associated with a biologically significant stimulus in Pavlovian aversive conditioning, thus reflecting a learning bias. Such learning bias can be characterized by a faster acquisition of a conditioned response, the acquisition of a larger conditioned response, and/or enhanced resistance to extinction of that conditioned response (Öhman & Mineka, 2001). Although all of these different indicators are considered as inherently valid, preferential emotional learning has been most consistently evidenced in humans as an enhanced persistence of the learned threat response to threat-relevant stimuli, whereas the learned threat response to threat-irrelevant stimuli generally extinguishes rapidly (Öhman & Mineka, 2001). According to preparedness and fear module theories, evolutionarily prepared threat-relevant – but not positive relevant – stimuli are readily associated with an aversive event. These theories would therefore imply that a conditioned response to positive relevant stimuli should hence be similarly, or even more quickly, extinguished than a conditioned response to neutral stimuli (Öhman & Dimberg, 1978; Öhman & Mineka, 2001). Conversely and congruently with the predictions of the relevance detection model, we predicted that the conditioned response to both threat-relevant and positive relevant stimuli would be more persistent than the conditioned response to neutral stimuli with less relevance.

To test this competing hypothesis, we conducted three experiments examining whether, similar to threat-relevant stimuli, positive stimuli with biological relevance to the organism likewise induce a learning bias during Pavlovian aversive conditioning. In each experiment, we manipulated the conditioned stimuli's valence in a differential aversive conditioning paradigm by using three distinct conditioned stimulus categories: negative biologically relevant stimuli (angry faces in Experiments 1 and 2, and snakes in Experiment 3), positive biologically relevant stimuli (baby faces in Experiments 1 and 2, and erotic stimuli in Experiment 3), and neutral, less relevant stimuli (neutral faces in Experiments 1 and 2, and neutral colored squares in Experiment 3). This set of experiments thereby is key in order to test the hypothesis that preferential emotional learning is driven by a relevance detection mechanism, without being selective to negative threatening stimuli.

EXPERIMENTS 1 AND 2

In Experiments 1 and 2, we investigated whether angry faces and baby faces are preferentially conditioned to threat relative to neutral faces. Experiment 2 consisted of a direct replication of Experiment 1 with the aim of establishing the observed effects' reproducibility and robustness within an even more highly powered experiment. Baby faces were selected as positive relevant conditioned stimuli (CSs) because they represent a prototypical instance of stimuli being positive and highly biologically relevant for the survival of the species (Brosch et al., 2008; Kringelbach, Stark, Alexander, Bornstein, & Stein, 2016; Pool, Brosch, et al., 2016; see also Lorenz, 1943). In agreement with this view, baby faces have been shown to elicit positive evaluations (e.g., Brosch, Sander, & Scherer, 2007), to be readily prioritized for access to attentional resources (Brosch et al., 2007, 2008; Kringelbach et al., 2016; Pool, Brosch, et al., 2016), and to hold high motivational salience and a high reward value (Parsons, Young, Kumari, Stein, & Kringelbach, 2011), all of these characteristics serving as evolutionarily adaptive traits for promoting caregiving behaviors in adults and ultimately infant survival (Kringelbach et al., 2016; Lorenz, 1943). In both experiments, the differential aversive conditioning procedure comprised three contiguous phases, following standard methodology (see Lonsdorf et al., 2017). During the initial habituation phase, all CSs were presented without being reinforced. In the subsequent acquisition phase, one stimulus (reinforced stimulus [CS+]) from each CS category was systematically paired with a mild electric stimulation (unconditioned stimulus [US]) using a partial reinforcement schedule, whereas the other stimulus (unreinforced stimulus [CS-]) from each category was never associated with the electric stimulation. During the extinction phase that followed, no electric stimulation was delivered. Skin conductance responses (SCRs) were measured during all the phases. The conditioned response (CR) was operationalized as the differential SCR to the CS+ minus CS- from the same CS category (see, e.g., Olsson et al., 2005) and used as an index of learning. Our prediction was that the CR to both angry faces and baby faces would be more resistant to extinction than the CR to neutral faces.

Method

Participants

In Experiment 1, 52 participants were recruited at the University of Geneva. They provided informed consent prior to the start of the experiment, which was approved by the Faculty of Psychology and Educational Sciences Ethics committee at the University of Geneva, and they received either partial course credit or monetary compensation (20 Swiss francs) for their participation. Twelve participants were excluded from the analyses due to technical problems (*n* = 8), for displaying virtually no SCRs (*n* = 2), or for failing to acquire a CR to at least one of the three CSs predictive of the US delivery (*n* = 2). These exclusion criteria are commonly applied in the contemporary human conditioning literature (e.g., Olsson et al., 2005;

Olsson & Phelps, 2004; Phelps et al., 2004; Stussi, Brosch, & Sander, 2015) and were determined prior to data collection. The final sample comprised 40 participants (31 women, 9 men), aged between 18 and 52 years old (mean age = 23.85 ± 6.26 years). The sample size was determined based on a power analysis conducted with G*Power 3 (Faul, Erdfelder, Lang, & Buchner, 2007). The analysis revealed that a total sample of 34 participants would be required to obtain a power of 80% to detect a moderate effect (*d* = 0.5) as reported in a previous study (Stussi et al., 2015). For counterbalancing purposes, we aimed to recruit a sample of 40 participants exhibiting differential conditioning to at least one of the three CS categories and stopped collecting data when we ascertained that the required number of participants had been reached.

In Experiment 2, 88 undergraduate psychology students from the University of Geneva were tested. None of them took part in Experiment 1. They provided informed consent prior to the start of the experiment, which was approved by the Faculty of Psychology and Educational Sciences Ethics committee at the University of Geneva, and received partial course credit for their participation. Twenty-eight participants were excluded from the analyses due to technical problems (*n* = 7), for displaying virtually no SCRs (*n* = 8), or for failing to acquire a CR to at least one of the three CSs predictive of the US delivery (*n* = 13). The final sample consisted of 60 participants (46 women, 14 men), aged between 19 and 50 years old (mean age = 23.03 ± 6.25 years). The sample size was determined based on a power analysis, which indicated that at least 54 participants would be required to achieve a power of 95% to detect a moderate effect (*d* = 0.5). We therefore aimed to recruit a sample of 60 participants who were conditioned to at least one of the three CS categories and stopped data collection once this sample had been reached.

Stimuli and apparatus

The CSs consisted of six different (male) faces divided into three categories: two adult faces with an angry expression, two adult faces with a neutral expression, and two baby faces. The four adult faces were taken from the Radboud Faces Database (model numbers 23 and 46 for the angry faces, and model numbers 15 and 25 for the neutral faces; Langner et al., 2010). The baby faces were selected from a set of infant faces used in previous studies (Coppin et al., 2014; Van Duuren, Kendell-Scott, & Stark, 2003). The selected faces were cut out from their original background and placed on a solid, gray background. All stimulus images were grayscale-transformed. Quantitative analyses (see Delplanque, N'diaye, Scherer, & Grandjean, 2007) confirmed that the angry, neutral, and baby stimulus images did not differ statistically in terms of luminance, apparent contrast, or mean energy in spatial-frequency bands. Each face served both as a CS+ and a CS-, counterbalanced across participants. An independent rating study ($N = 63$; see supplemental materials) in which the stimuli used in Experiments 1 and 2 were evaluated on a visual analog scale (VAS) ranging from 0 (*very unpleasant*) to 100 (*very pleasant*) substantiated that the angry faces were evaluated as negative ($M = 30.17$, $SE = 2.07$), the neutral faces as neutral ($M = 50.71$, $SE = 1.53$), and the baby faces as positive ($M = 72.12$, $SE = 2.08$). In Experiment 1, the US consisted of a mild electric stimulation (200-ms duration, 50 pulses/s) delivered to the participants' right wrist through a Grass SD9 stimulator (Grass Medical Instruments, West Warwick, RI) charged by a stabilized current. In Experiment 2, the US was a mild electric stimulation (10-ms duration) delivered to the participants' right wrist through a unipolar pulse electric stimulator (STM200; BIOPAC Systems Inc., Goleta, CA).

In Experiment 1, the CR was assessed through SCR measured with two pre-gelled disposable Ag-AgCl electrodes (11-mm contact diameter). In Experiment 2, the CR was assessed through SCR measured with two Ag-AgCl electrodes (6-mm contact diameter) filled with 0.5%

NaCl electrolyte gel. In both experiments, the electrodes were attached to the distal phalanges of the second and third digits of the participants' left hand. The SCR data was continuously recorded with a sampling rate of 1000 Hz through a BIOPAC MP150 system (Santa Barbara, CA). SCR was analyzed offline with AcqKnowledge software (version 4.2 in Experiment 1, and version 4.4 in Experiment 2; BIOPAC Systems Inc. Goleta, CA).

Procedure

Before conditioning, a work-up procedure was conducted to individually set the stimulation intensity (*M* = 36.75 V, *SE* = 1.27 in Experiment 1, and *M* = 34.75 V, *SE* = 0.98 in Experiment 2) to a level reported as "uncomfortable, but not painful" by the participant (e.g., Lonsdorf et al., 2017; Olsson et al., 2005). The initial habituation phase of the differential aversive conditioning procedure comprised two unreinforced presentations of each of the six CSs. During the acquisition phase, each CS was presented seven times. This phase always started with a reinforced CS+ trial. Five of the seven presentations of each CS+ coterminated with the US delivery, whereas the presentations of each CS- were never paired with the US. We used a partial reinforcement schedule to potentiate the CR resistance to extinction, with the aim of optimizing the investigation of the differences in the persistence of learned emotional responses between the three CS categories used. The final extinction phase consisted of six unreinforced presentations of each CS. During all the conditioning phases, the CSs were presented for 6 s with an intertrial interval ranging from 12 to 15 s. The CSs' order of presentation was pseudorandomized into eight different orders to systematically counterbalance the associations between the face stimuli and CS type (CS+ vs. CS-) across the three CS categories (anger vs. baby vs. neutral).

After the extinction phase, participants completed subjective ratings of CS-US contingency and CS liking as manipulation checks in order to assess their awareness of the reinforcement contingencies and the CSs' pleasantness, respectively. In this procedure, the CSs were presented again, accompanied by a VAS. For the CS-US contingency ratings, participants were asked to rate to what extent the CS was predictive of the delivery of an electric stimulation, the VAS ranging from 0 (*never*) to 100 (*always*). For the CS liking ratings, participants were asked to rate to what extent the CS was unpleasant or pleasant, the VAS ranging from 0 (*very unpleasant*) to 100 (*very pleasant*). The order of the CS presentations and the questions was randomized across participants.

Response definition

SCR was measured for each trial as the peak-to-peak amplitude difference in skin conductance of the largest response starting in the 0.5 to 4.5 s temporal window following CS onset. The minimal response criterion was 0.02 µS. Responses below this criterion were scored as '0' and remained in the analyses. The SCR data was low-pass filtered (Blackman -92 dB, cutoff frequency = 1 Hz). SCRs were detected automatically with AcqKnowledge software as well as checked manually for artifacts and response detection. Trials containing artifacts influencing the coding of event-related SCRs or containing loss of SCR signal (1.78% in Experiment 1, and 0.003% in Experiment 2) were removed from the analyses. The raw SCR scores were square-root-transformed to normalize the distributions and scaled according to each participant's mean square-root-transformed unconditioned response (UR). The UR was scored as the peak-to-peak amplitude difference in skin conductance of the largest response starting in the 0.5 to 4.5 s temporal window following the US delivery, and the mean UR was calculated across all USs for each participant (see supplemental materials). The habituation means included the first two presentations of each CS (see Figure 1). To examine the CR acquisition speed, the acquisition means were separated into an early (i.e., the first three presentations of each CS following the first association of the CS+ with the US; Trials 4 to 6, see Figure 1) and a late (the subsequent three presentations of each CS; Trials 7 to 9, see Figure 1) phase (see, e.g., Lonsdorf et al., 2017; Stussi et al., 2015). The first acquisition trial for each CS was omitted from the analyses because the CSs+ were predictive of the US only after their first association with the electric stimulation. The extinction means comprised the last six presentations of each CS (i.e., Trials 10 to 15, see Figure 1). The analyses of the conditioning data were performed on the CR, which was calculated by subtracting the SCR to the CS- from the SCR to the CS+ from the same CS category (e.g., Olsson et al., 2005). This procedure permits to reduce the confounding role of preexisting differences in the CS categories' emotional salience (Olsson et al., 2005) and to specifically control for learning within participant.

Statistical analyses

As it is standardly done in the human conditioning literature (see, e.g., Lonsdorf et al., 2017), the SCR data was analyzed separately for the habituation, acquisition, and extinction phases. One-way repeated measures analyses of variance (ANOVAs) with CS category (anger vs. baby vs. neutral) as a within-participant factor were used to analyze the habituation and extinction data, whereas a two-way repeated measures ANOVA with CS category (anger vs. baby vs. neutral) and time (early vs. late) as within-participant factors was used for the acquisition data. One-sample *t*-tests were conducted to assess whether differential conditioning occurred to angry, baby, and neutral faces across the whole acquisition phase. To specifically test our a priori hypothesis, we performed a planned contrast analysis comparing the CR to both angry (contrast weight: +1) and baby (contrast weight: +1) faces vs. neutral faces (contrast weight: -2) in extinction. Following this main contrast, three further contrasts were conducted to examine more closely whether the CR would be more persistent to (a) angry (contrast weight: +1) vs. neutral (contrast weight: -1) faces and (b) baby (contrast weight: +1) vs. neutral (contrast weight: -1) faces, and to assess the possible differences between (c) angry (contrast weight: +1) and baby (contrast weight: -1) faces. Because these contrasts were non-orthogonal, a Holm-Bonferroni sequential procedure (Holm, 1979) was applied to correct for multiple comparisons. Specifically, the alpha level of the contrast with the lowest *p* value was set as $\alpha = .05/4 = .0125$, the alpha level of the contrast with the second lowest *p* value as $\alpha = .05/3 = .0167$, the alpha level of the contrast with the second highest *p* value as $\alpha = .05/2 = .025$, and the alpha level of the contrast with the highest *p* value as $\alpha = .05$. An alpha level of $\alpha = .05$ was adopted for all the other statistical analyses performed. For each contrast, we additionally computed the Bayes factor (BF_{10}) quantifying the likelihood of the data under the alternative hypothesis relative to the likelihood of the data under the null hypothesis (see, e.g., Dienes, 2011; Rouder, Speckman, Sun, Morey, & Iverson, 2009), using a Cauchy prior width of 0.5. For instance, a BF_{10} of 4 indicates that the data is four times more likely to be observed under the alternative hypothesis than under the null hypothesis. A BF_{10} larger than 3 (moderate evidence), larger than 10 (strong evidence), or larger than 30 (very strong evidence) is considered to provide evidence in favor of the alternative hypothesis relative to the null hypothesis, whereas a BF_{10} smaller than 0.333 (moderate evidence), smaller than 0.100 (strong evidence), or smaller than 0.033 (very strong evidence) is considered to provide evidence in favor of the null hypothesis over the alternative hypothesis (Jeffreys, 1961). We performed one-sided testing to test our a priori, theory-driven directional hypotheses (one-sample *t*-tests, main contrast and contrasts a and b), whereas two-sided testing was used when we did not have a directional prediction (contrast c).

The CS-US contingency and CS liking ratings were each analyzed with a two-way repeated measures ANOVA with CS type (CS+ vs. CS-) and CS category (anger vs. baby vs. neutral) as within-participant factors. Significant effects were followed up with a multiple comparison procedure using Tukey's HSD tests when applicable.

We report either partial η^2 or Hedges' g_{av} as estimates of effect size (see Lakens, 2013) and their 90% or 95% confidence interval (CI), respectively. Huynh-Feldt adjustments of degrees of freedom were applied when appropriate.

Results

Figure 1 displays the mean SCR magnitudes to angry, baby, and neutral faces throughout the habituation, acquisition, and extinction phases separately for the CS+ and the CS-. The conditioned response to angry, baby, and neutral faces during acquisition and extinction is depicted in Figure 2.

Experiment 1

Skin conductance response. In the habituation phase, no preexisting difference in differential SCRs to the CS categories was found, $F(2, 78) = 0.64$, $p = .533$, partial $\eta^2 = .016$, 90% CI [.000, .069]. Similarly, no statistical difference between the CS categories emerged during acquisition, $F(2, 78) = 0.44$, $p = .643$, partial $\eta^2 = .011$, 90% CI [.000, .057]. Moreover, the CR did not statistically differ between the early and late phases of acquisition, $F(1, 39) = 0.05$, $p = .816$, partial $\eta^2 = .001$, 90% CI [.000, .054]. No statistically significant interaction effect of CS category and time was observed, $F(2, 78) = 1.75$, $p = .180$, partial $\eta^2 = .043$, 90% CI [.000, .120], which indicates that there was no statistical difference in the speed of the CR acquisition across the CS categories. Further analyses revealed however a reliably greater SCR to

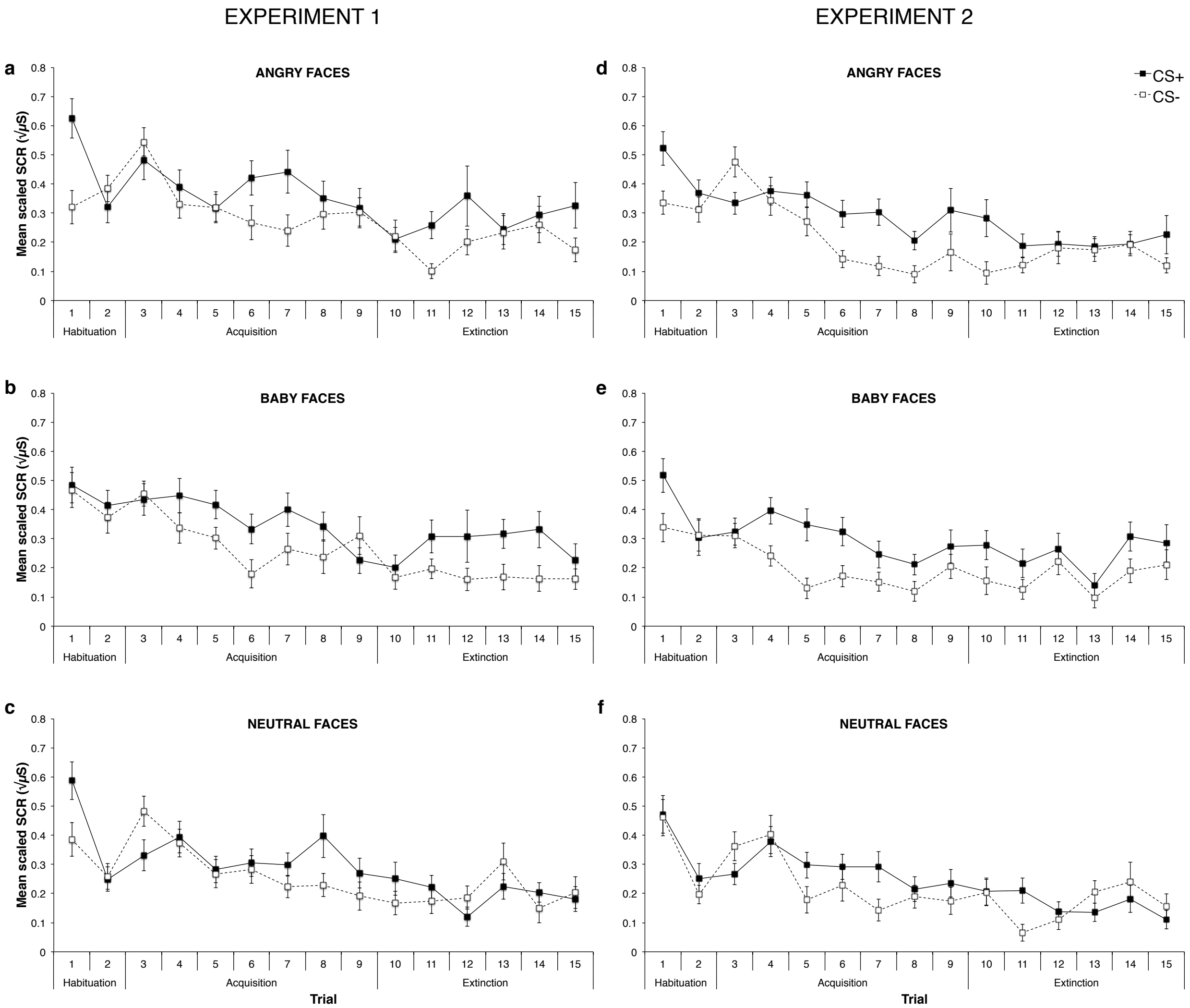

Figure 1. Mean scaled skin conductance response (SCR) to the conditioned stimuli as a function of the conditioned stimulus type (CS+ vs. CS-) across trials in (a-c) Experiment 1 and (d-f) Experiment 2. Mean scaled SCR to (a, d) angry faces, (b, e) baby faces, and (c, f) neutral faces. Errors bars indicate ± 1 *SEM* adjusted for within-participant designs (Morey, 2008).

the CS+ than CS- for angry, $t(39) = 2.31$, $p = .013$ (one-tailed), $g_{av} = 0.507$, 95% CI [0.061, 0.967], baby, $t(39) = 3.05$, $p = .002$ (one-tailed), $g_{av} = 0.669$, 95% CI [0.214, 1.141], and neutral faces, $t(39) = 2.61$, $p = .006$ (one-tailed), $g_{av} = 0.571$, 95% CI [0.122, 1.036], indicating successful differential conditioning to all three CS categories (see Figure 2a). Central to our hypothesis, analysis of the extinction phase showed that the CS categories differentially affected the persistence of the CR, $F(2, 78) = 4.51$, $p = .014$, partial $\eta^2 = .104$, 90% CI [.012, .204]. As predicted by the relevance detection hypothesis, the CR to both angry and baby faces was more resistant to extinction than the CR to neutral faces, $t(39) = 3.04$, $p = .002$ (one-tailed), $g_{av} = 0.598$, 95% CI [0.191, 1.021], $BF_{10} = 19.154$ (see Figure 2a). Direct comparisons revealed a more persistent CR to angry faces compared with neutral faces, $t(39) = 2.43$, $p = .010$ (one-tailed), $g_{av} = 0.472$, 95% CI [0.076, 0.881], $BF_{10} = 5.348$ (see Figure 2a). Importantly, the CR to baby faces was likewise more persistent than the CR to neutral faces, $t(39) = 2.73$, $p = .005$ (one-tailed), $g_{av} = 0.569$, 95% CI [0.141, 1.014], $BF_{10} = 9.679$, whereas there was no statistical difference in the resistance to extinction of the CR to angry faces compared with baby faces, $t(39) = -0.64$, $p = .524$ (two-tailed), $g_{av} = -0.132$, 95% CI [-0.545, 0.278], $BF_{10} = 0.279$ (see Figure 2a).

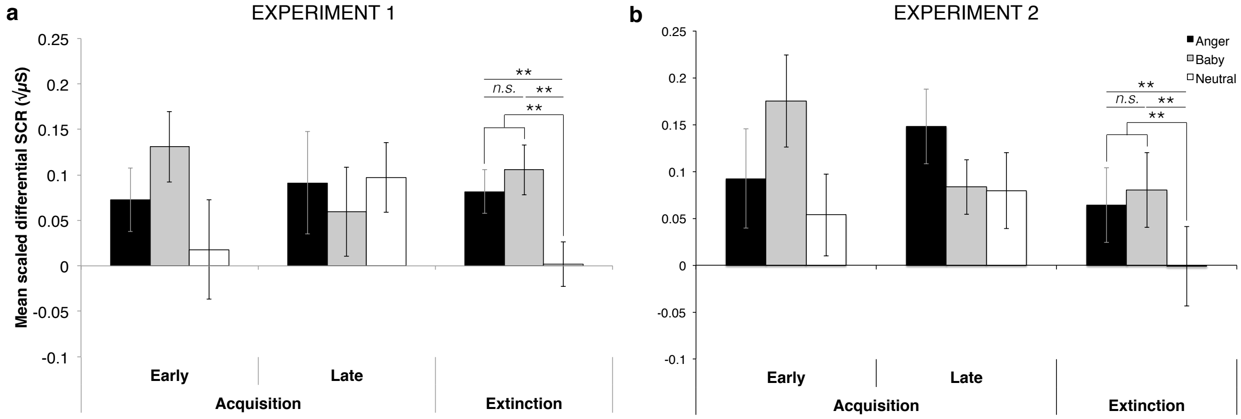

Figure 2. Mean conditioned response (scaled differential skin conductance response [SCR]) as a function of the conditioned stimulus category (anger vs. baby vs. neutral) during (early and late) acquisition and extinction in (a) Experiment 1 and (b) Experiment 2. Errors bars indicate ± 1 *SEM* adjusted for within-participant designs (Morey, 2008). Asterisks indicate statistically

significant differences between conditions (***p* < .01, one-tailed) and *ns* indicates a statistically nonsignificant difference.

Subjective ratings. The CS-US contingency ratings showed that the CSs+ were deemed more likely to be associated with the US than the CSs-, *F*(1, 39) = 75.25, *p* < .001, partial η^2 = .659, 90% CI [.494, .745], whereas there was no interaction between CS type and CS category, *F*(2, 78) = 0.73, *p* = .485, partial η^2 = .018, 90% CI [.000, .075]. Moreover, the CS categories differentially influenced the CS-US contingency ratings, *F*(1.69, 66.00) = 7.97, *p* = .001, partial η^2 = .170, 90% CI [.045, .291]. Follow-up analyses revealed that angry faces were rated as more likely to be predictive of the US than both baby faces (*p* = .011, g_{av} = 0.621, 95% CI [0.108, 1.151]) and neutral faces (*p* < .001, g_{av} = 0.878, 95% CI [0.399, 1.381]), whereas there was no statistical difference in the CS-US contingency ratings for baby faces relative to neutral faces (*p* = .681, g_{av} = 0.225, 95% CI [-0.196, 0.652]; see Figure 3a).

The CS liking ratings revealed that the CSs- were more liked than the CSs+, *F*(1, 39) = 5.75, *p* = .021, partial η^2 = .128, 90% CI [.011, .289], a significant main effect not qualified by an interaction with CS category, *F*(2, 78) = 0.25, *p* = .780, partial η^2 = .006, 90% CI [.000, .040]. The CS liking ratings were also modulated by the CS categories, *F*(1.78, 69.23) = 68.92, *p* < .001, partial η^2 = .639, 90% CI [.514, .710]. Follow-up analyses showed that baby faces were rated as more pleasant than angry faces (*p* < .001, g_{av} = 2.505, 95% CI [1.792, 3.302]) and neutral faces (*p* < .001, g_{av} = 1.386, 95% CI [0.918, 1.898]), and that neutral faces were rated as more pleasant than angry faces (*p* < .001, g_{av} = 1.310, 95% CI [0.796, 1.863]; see Figure 3b).

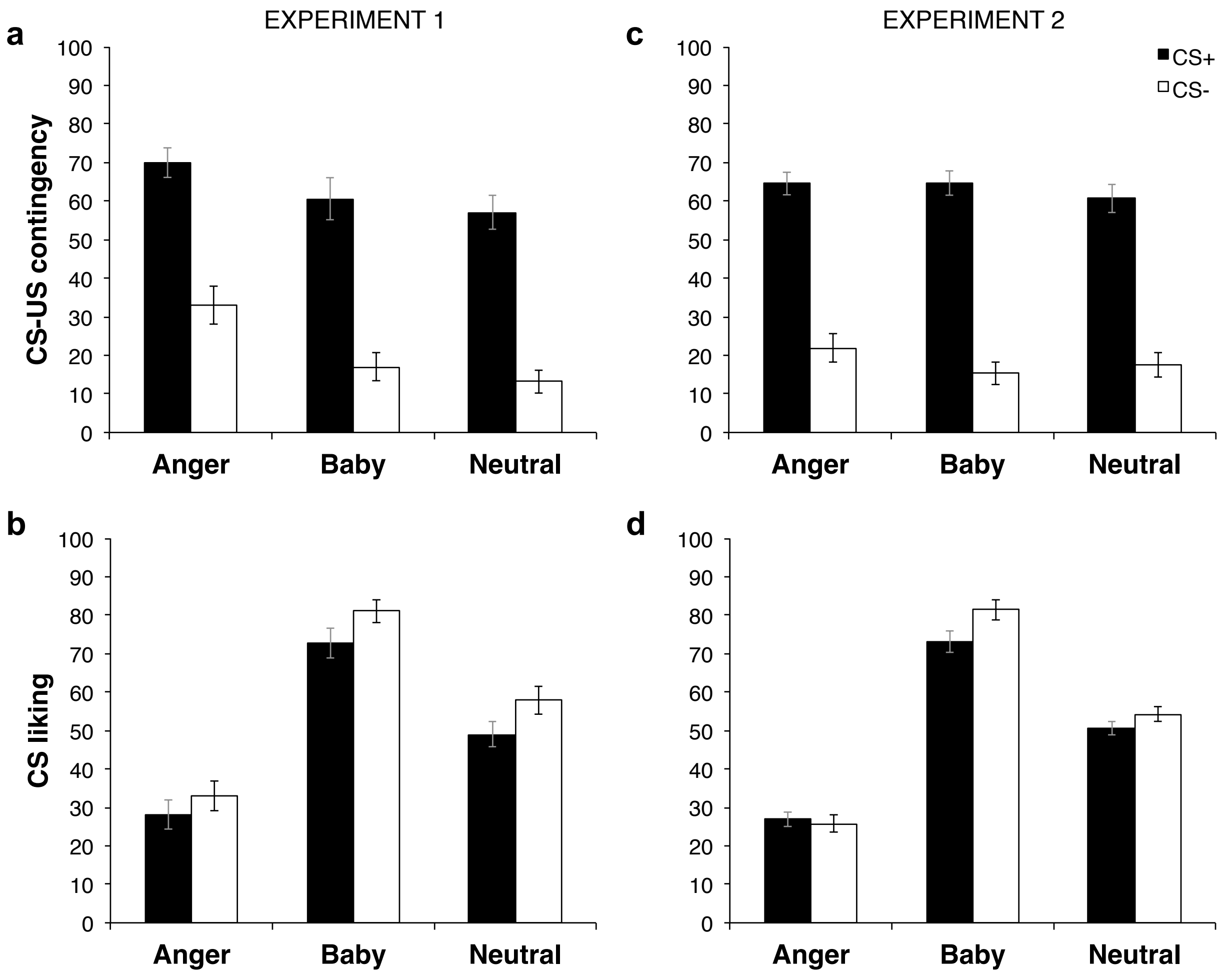

Figure 3. Mean subjective ratings as a function of the conditioned stimulus type (CS+ vs. CS-) and the conditioned stimulus category (anger vs. baby vs. neutral) in (a-b) Experiment 1 and (c-d) Experiment 2. Mean (a, c) CS-US contingency ratings and (b, d) CS liking ratings. Errors bars indicate ± 1 *SEM* adjusted for within-participant designs (Morey, 2008).

Experiment 2

Skin conductance response. During habituation, there was no statistical difference in differential SCRs to the different CS categories, $F(1.80, 105.96) = 0.76$, $p = .459$, partial $\eta^2 = .013$, 90% CI [.000, .057]. Likewise, the CR did not statistically differ across the three CS categories during the acquisition phase, $F(1.84, 108.67) = 1.72$, $p = .186$, partial $\eta^2 = .028$, 90% CI [.000, .087]. No statistically significant main effect of time was found, $F(1, 59) = 0.02$, $p = .881$, partial $\eta^2 = .0004$, 90% CI [.000, .016]. The interaction between CS category and time did not yield statistical significance either, $F(1.78, 104.89) = 1.53$, $p = .222$, partial $\eta^2 = .025$, 90% CI [.000, .083], which suggests that the CR acquisition speed did not differ across the CS categories. As in Experiment 1, one-sample *t* tests showed a greater SCR to the CS+ than CS- for angry, $t(59) = 4.80$, $p < .001$ (one-tailed), $g_{av} = 0.865$, 95% CI [0.482, 1.264], baby, $t(59) = 4.45$, $p < .001$ (one-tailed), $g_{av} = 0.801$, 95% CI [0.422, 1.195], and neutral faces, $t(59) = 1.96$, $p = .027$ (one-tailed), $g_{av} = 0.353$, 95% CI [-0.007, 0.720],[1] reflecting successful differential conditioning to all three CS categories (see Figure 2b). Analysis of the extinction phase revealed that the CS categories differentially modulated the CR resistance to extinction, $F(2, 118) = 4.93$, $p = .009$, partial $\eta^2 = .077$, 90% CI [.012, .153]. Replicating results from Experiment 1, the CR to both angry and baby faces was more persistent than the CR to neutral faces, $t(59) = 3.21$, $p = .001$ (one-tailed), $g_{av} = 0.444$, 95% CI [0.162, 0.735], $BF_{10} = 31.123$ (see Figure 2b). Direct comparisons showed that the CR to angry faces was more resistant to extinction relative to neutral faces, $t(59) = 2.45$, $p = .009$ (one-tailed), $g_{av} = 0.352$, 95% CI [0.063, 0.647], $BF_{10} = 5.363$ (see Figure 2b). Critically, the CR to baby faces was also more resistant to extinction than the CR to neutral faces, $t(59) = 2.99$, $p = .002$ (one-tailed), $g_{av} = 0.451$, 95% CI [0.144, 0.765], $BF_{10} = 17.861$, whereas the CR persistence to angry faces did not statistically differ from the CR persistence to baby faces, $t(59) = -0.57$, $p = .571$ (two-tailed), $g_{av} = -0.094$, 95% CI [-0.423, 0.233], $BF_{10} = 0.225$ (see Figure 2b).[2]

Subjective ratings. The CS-US contingency ratings indicated that the CSs+ were rated as being more predictive of the US than the CSs-, $F(1, 59) = 108.15$, $p < .001$, partial $\eta^2 = .647$, 90% CI [.518, .724] (see Figure 3c), whereas the interaction between CS type and CS category did not reach statistical significance, $F(2, 118) = 1.12$, $p = .331$, partial $\eta^2 = .019$, 90% CI [.000, .065]. In contrast to Experiment 1, no main effect of CS category was found, $F(2, 118) = 1.47$, $p = .235$, partial $\eta^2 = .024$, 90% CI [.000, .076].

The CS liking ratings revealed a main effect of CS type, $F(1, 59) = 4.55$, $p = .037$, partial $\eta^2 = .072$, 90% CI [.002, .191], and a main effect of CS category, $F(1.66, 98.16) = 196.77$, $p < .001$, partial $\eta^2 = .769$, 90% CI [.701, .810]. These main effects were however qualified by the higher-order interaction between CS type and CS category, $F(2, 118) = 3.37$, $p = .038$, partial $\eta^2 = .054$, 90% CI [.002, .122]. Follow-up analyses showed that baby faces were rated as more pleasant than angry faces (all *p*s < .001, $2.41 < g_{av}$s < 2.96) and neutral faces (all *p*s < .001, $1.02 < g_{av}$s < 1.80), while neutral faces were rated as more pleasant than angry faces (all *p*s < .001, $1.59 < g_{av}$s < 1.80). Furthermore, whereas the CS- was evaluated as more pleasant than the CS+ for baby faces ($p = .021$, $g_{av} = 0.397$, 95% CI [0.068, 0.734]), there was no statistical difference in rated pleasantness between the CS- and the CS+ for angry faces ($p = .997$, $g_{av} = -0.072$, 95% CI [-0.323, 0.179]) and neutral faces ($p = .711$, $g_{av} = 0.270$, 95% CI [-0.080, 0.626]; see Figure 3d).

Discussion

In line with the relevance detection model's prediction, Experiments 1 and 2 revealed that both angry faces and baby faces produced a learning bias during Pavlovian aversive conditioning, as shown by the enhanced conditioned response persistence to angry faces and baby faces compared with neutral faces. Whereas the results for angry faces replicate previous findings (e.g., Öhman & Dimberg, 1978; Öhman & Mineka, 2001), the greater resistance to extinction of the conditioned response to baby faces expands the existing human conditioning literature, and suggests that positive stimuli with biological relevance can likewise be preferentially conditioned to threat, thereby demonstrating that preferential Pavlovian aversive conditioning is not specific to threat-related stimuli.

In contrast, we found no evidence for faster or stronger acquisition of the conditioned response to angry or baby faces relative to neutral faces. Such absence of differences across conditioned stimulus categories during acquisition is however not surprising when considering the human conditioning literature, which has generally shown a lack of experimental support for faster or stronger aversive conditioning to specific stimulus classes, such as threat-relevant stimuli (see McNally, 1987; Öhman & Mineka, 2001, for reviews). Although enhanced resistance to extinction has been frequently demonstrated to threat-relevant stimuli (Öhman & Mineka, 2001), evidence for faster or larger aversive conditioning to threat-relevant stimuli remains by comparison very scarce (Ho & Lipp, 2014; Öhman, Eriksson, & Olofsson, 1975). A potential explanation for this absence of significant effect relates to the use of a relatively high reinforcement rate whereby the CSs+ reliably predicted the US, which may have entailed rapid aversive conditioning to all the conditioned stimulus categories within a few pairings between the CSs+ and the US, and consequently led to ceiling effects in the conditioned response acquisition readiness, thereby potentially obscuring the emergence of differences in learning patterns among the stimulus categories (see Ho & Lipp, 2014; Lissek, Pine, & Grillon, 2006).

Further, it should also be noted that the pattern of skin conductance responses in Experiment 1 was somewhat unusual at the descriptive level in comparison with what is generally observed in human aversive conditioning studies. Whereas the difference between the CS+ and the CS- is usually evident at the end of acquisition and at the onset of extinction, there seemed to be no such difference at the last acquisition trial and first extinction trial for angry faces (see Figure 1a) and baby faces (see Figure 1b). It could be speculated that this pattern may be due to the use of a within-participant design using six different CSs, instead of a between-participant design (e.g., Öhman & Dimberg, 1978; Öhman et al., 1976) or a within-participant design including only two to four conditioned stimuli (e.g., Ho & Lipp, 2014; Olsson et al., 2005), which might have entailed a stronger habituation of skin conductance responses to the CS+ than commonly observed. The subsequent reemergence of differences between the CS+ and the CS- could then have been induced by the change of contingency between the CS+ and the US, thus possibly leading to dishabituation effects. However, it remains unclear why this relative lack of evident CS+/CS- differentiation at the last acquisition trial and first extinction trial was observed for angry faces and baby faces but not for neutral faces, and why it was observed in Experiment 1, but not in Experiment 2, which suggests that it may otherwise simply reflect noise in the data.

It is also noteworthy that the observed enhanced resistance to extinction effects might be interpreted as reflecting selective sensitization, a nonassociative process, in addition to – or rather than – a conditioning process (Lovibond, Siddle, & Bond, 1993). Selective sensitization has been proposed as a putative mechanism responsible for enhanced responding to threat-relevant CSs+ during extinction, emerging as a result of the activation of preexisting response tendencies to these stimuli under certain conditions, such as threat or a state of arousal (e.g., Lovibond et al., 1993). In the present case, it could then be argued that the angry and the baby face CSs+ may have led to a greater resistance to extinction of the conditioned response than the neutral face CS+ because of their inherent potential to elicit enhanced responses in a state of arousal (i.e., induced by threat of electric stimulation). Even though we cannot completely rule out this possibility, it is unlikely that selective sensitization was the sole factor accounting for our results. Selective sensitization, as a relatively short-lived phenomenon (e.g., Lipp, Cronin, Alhadad, & Luck, 2015), has been suggested to be insufficient to explain the long-lasting effects classically observed in human aversive conditioning studies using threat-relevant stimuli (Öhman & Mineka, 2001). Furthermore, analyses of the SCRs during the habituation phase in Experiments 1 and 2 provided no support for a selective sensitization to angry and baby faces compared with neutral faces,[3] thereby suggesting that the enhanced resistance to extinction to angry and baby faces primarily resulted from an associative learning process.

In Experiments 1 and 2, subjective ratings showed that the CS+ was evaluated as being more likely to be predictive of the US delivery than the CS- across the three stimulus categories, indicating that, overall, participants were aware of the contingencies. In Experiment 1, angry faces were deemed more predictive of the US than baby and neutral faces, which might suggest that negative threat-relevant stimuli are more likely to be associated with an aversive outcome at the explicit level irrespective of the actual contingencies (Davey, 1992; Tomarken, Mineka, & Cook, 1989). However, this interpretation should be considered with caution as subjective ratings were collected exclusively after extinction but not after acquisition. Moreover, this effect did not replicate in Experiment 2, highlighting that the boundary conditions of such potential expectancy or covariation bias remain to be determined. As anticipated, baby faces were evaluated as more pleasant than neutral and angry faces, and neutral faces were rated as more pleasant than angry faces after the extinction phase in both experiments, thus reflecting an efficient manipulation of the CSs’ valence. In Experiment 1, aversive conditioning had a similar effect on the CS+’s and the CS-’s rated pleasantness across the three stimulus categories; however, the CS- was evaluated as statistically significantly more pleasant than the CS+ only for baby faces in Experiment 2. Although not central to the present study's aims, these results likely stem from the fact that the electric stimulation was shorter in Experiment 2 than in Experiment 1 (10-ms vs. 200-ms duration), thus being less aversive and perceived as less intense,[4] which might have induced less robust evaluative conditioning effects (see Hofmann, De Houwer, Perugini, Baeyens, & Crombez, 2010).

In sum, the occurrence of a Pavlovian learning bias to both angry faces and baby faces supports the view that preferential emotional learning is underlain by a relevance detection mechanism rather than a threat- or valence-specific mechanism, such as a fear module (Öhman & Mineka, 2001). Nonetheless, we only used a single instance of positive relevant stimuli in both experiments, thus entailing the possibility that the observed effects are selective to baby faces. The relevance detection model however predicts that positive biologically relevant stimuli induce a learning bias during Pavlovian aversive conditioning, this learning bias thereby not being confined to baby faces. Findings showing that other categories of positive relevant stimuli are preferentially conditioned to threat as well would hence provide additional empirical evidence in favor of this model. Therefore, we tested in Experiment 3 whether an enhanced Pavlovian aversive conditioning to positive relevant stimuli also occurs in response to another category of positive emotional stimuli that are relevant to the organism, namely erotic stimuli (see, e.g., Bradley, Codispoti, Cuthbert, & Lang, 2001; Panksepp, 1998; Sennwald et al., 2016).

EXPERIMENT 3

In Experiment 3, we aimed to replicate and extend the findings from Experiments 1 and 2 with different categories of stimuli. More specifically, we investigated whether both snakes and erotic stimuli are preferentially conditioned to threat in comparison with neutral stimuli. To this end, we used a differential aversive conditioning procedure, in which snake images, erotic images, and colored squares were presented as CSs. Erotic stimuli were selected as positive biologically relevant CSs because they are typically positive and rewarding, and hold high relevance for the species' reproduction and survival, thereby being biologically and motivationally relevant to the organism (Berridge & Kringelbach, 2015; Bradley et al., 2001; Georgiadis & Kringelbach, 2012; Panksepp, 1998; Pool, Brosch, et al., 2016; Sander et al., 2003; Schultz, 2015; Sennwald et al., 2016). Snakes were selected as negative biologically relevant CSs because they constitute the prototypical instance of negative threat-relevant stimuli from phylogenetic origin that have threatened the survival of the species (see, e.g., Öhman & Mineka, 2001). The differential aversive conditioning procedure was identical to the one used in Experiments 1 and 2. After the habituation phase, during which all CSs were presented without being reinforced, the CS+ from each CS category was systematically paired with a mild electric stimulation (US) using a partial reinforcement schedule during acquisition, whereas the CS- from each category was never associated with the electric stimulation. In the subsequent extinction phase, the electric stimulation was no longer delivered. As in Experiments 1 and 2, the CR was operationalized as the differential SCR to the CS+ minus CS- from the same CS category (see, e.g., Olsson et al., 2005) and used as an index of learning. Our prediction was that the CR to both snake images and erotic images would be more resistant to extinction than the CR to neutral colored squares.

Method

Participants

Fifty-five male volunteers were recruited at the University of Geneva. They provided informed consent prior to the start of the experiment, which was approved by the Regional Research Ethics Committee in Geneva, and received monetary compensation (20 Swiss francs) for their participation. As visual sexual stimuli are primarily tailored for men, who are accordingly thought to be generally more interested in such stimuli than women (e.g., Hamann, Herman, Nolan, & Wallen, 2004; but see, e.g., Rupp & Wallen, 2008, for a discussion of the role of the stimulus materials used), only men were included in the experiment. Fifteen participants were excluded from the analyses due to technical problems (*n* = 2), for displaying virtually no SCRs (*n* = 4), for failing to acquire a CR to at least one of the three CSs predictive of the US delivery (*n* = 6), or for withdrawing from the experiment early (*n* = 3). The final sample consisted of 40 men aged between 19 and 42 years old (mean age = 24.80 ± 5.43 years). The sample size was established on the basis of a power analysis (see Experiment 1) with the aim of recruiting a sample of 40 participants exhibiting differential conditioning to at least one of the three CS categories. We stopped collecting data when the required number of participants had been reached.

Stimuli and apparatus

The CSs were selected individually for each participant among a set of 12 snake images taken from the International Affective Picture System[5] (IAPS; Lang, Bradley, & Cuthbert, 2008), 24 erotic images (12 images of nude or partially nude men and 12 images of nude or partially nude women; Sennwald et al., 2018), and 12 colored squares. Based on each participant's ratings, the two most disliked snake images, the two most liked erotic images, and the two most neutral colored squares were used as CSs. In the event that several images had identical liking ratings within a CS category, the two most arousing images were selected for the snake and erotic CS categories, respectively, whereas the two least arousing colored squares were selected for the neutral CS category. If the liking and arousal ratings were identical for several images within a CS category, the images that had been the most recently presented were chosen. The attribution of the CS+ and CS- roles to the two selected stimuli for each CS category was counterbalanced across participants. The rationale for the CSs' selection procedure was to take into account individual differences in response to erotic stimuli, the responses to such stimuli being notoriously highly variable, by adequately considering individual preferences (see Kagerer et al., 2014; Sennwald et al., 2018). This way we could ensure that the erotic stimuli were rewarding, thereby increasing the chances of these stimuli to be motivationally relevant for the participants' sexual concerns (see Sennwald et al., 2018). The selection procedure was likewise applied to the snake and neutral CSs to ensure the equal treatment of each CS category, as well as to ensure that the snake CSs were deemed negative and the neutral CSs neutral. The US was a mild electric stimulation (200-ms duration, 50 pulses/s) delivered to the participants' dominant wrist through a Grass SD9 stimulator (Grass Medical Instruments, West Warwick, RI) charged by a stabilized current.

The CR was assessed through SCR measured with two Ag-AgCl electrodes (6-mm contact diameter) filled with 0.5% NaCl electrolyte gel. The electrodes were attached to the distal phalanges of the second and third digits of the participants' non-dominant hand. The SCR data was continuously recorded with a sampling rate of 1000 Hz through a BIOPAC MP150 system (Santa Barbara, CA). SCR was analyzed offline with AcqKnowledge software (version 4.2; BIOPAC Systems Inc., Goleta, CA).

Questionnaires

The Sexual Desire Inventory 2 (SDI-2; Spector, Carey, & Steinberg, 1996) and a questionnaire on sexual orientation were used in this experiment. The SDI-2 consists of a 14-item inventory indexing dyadic (summed score from 0 to 62) and solitary sexual desire (summed score from 0 to 23), as well as general sexual desire (summed score from 0 to 109). It was used to examine whether there might be an association between participants' sexual desire and their CR to erotic stimuli during the acquisition and extinction phases of the aversive conditioning procedure (see supplemental materials). Participants reported a mean dyadic sexual desire of 42.05 (*SE* = 1.02, range = 27-60), a mean solitary sexual desire of 10.70 (*SE* = 0.88, range = 0-23), and a mean general sexual desire of 66.08 (*SE* = 1.69, range = 47-93). The sexual orientation questionnaire was used to establish participants' sexual orientation using the Kinsey scale (Kinsey, Pomeroy, & Martin, 1948) on four different aspects of sexual orientation (i.e., sexual attraction, sexual behavior, sexual fantasies, and sexual identity).

Procedure

Prior to the experiment, participants were asked to fill out the SDI-2 and the sexual orientation questionnaire. Subsequently, they were asked to rate the 48 stimulus images according to their liking and felt arousal. The liking ratings measured how much participants liked seeing the displayed image on a VAS ranging from 0 (*not at all*) to 100 (*extremely*), whereas the arousal ratings measured how much participants felt physiologically aroused by the displayed image on a VAS ranging from 0 (*very weakly*) to 100 (*very strongly*). The stimulus images' presentation order was randomized across participants.

Once the CSs' selection procedure was completed, participants first underwent a work-up procedure in order to individually set the electric stimulation intensity (*M* = 29.75 V, *SE* = 1.16), and then the differential aversive conditioning procedure. Finally, participants completed subjective ratings of CS-US contingency and CS liking as manipulation checks to assess their awareness of the reinforcement contingencies and the CSs' pleasantness, respectively. All these procedures were identical to the ones used in Experiments 1 and 2.

Response definition

Response definition was strictly the same as in Experiments 1 and 2. Trials containing artifacts influencing the coding of event-related SCRs (0.005%) were removed from the analyses.

Statistical analyses

We performed repeated measures ANOVAs with CS type (CS+ vs. CS-) and CS category (snake vs. erotic vs. neutral) as within-participant factors on the liking and arousal ratings collected during the CSs' selection procedure to ensure (a) that there were no preexisting differences in the liking and arousal ratings between the selected CS+ and CS- within each CS category, and (b) that the selected erotic images were more liked than the selected snake images and the selected neutral colored squares, and that the selected neutral colored squares were more liked than the selected snake images. A multiple comparison procedure using Tukey's HSD tests was applied to follow up significant effects when applicable. Statistical analyses of the SCR data and the subjective ratings (i.e., CS-US contingency and CS liking ratings) were identical to the ones used in Experiments 1 and 2.

As in Experiments 1 and 2, we report either partial η^2 or Hedges' g_{av} as estimates of effect size (see Lakens, 2013) and their 90% or 95% CI, respectively. Huynh-Feldt adjustments of degrees of freedom were applied when appropriate.

Results

Figure 4 displays the mean SCR magnitudes to snake, erotic, and neutral stimuli across the habituation, acquisition, and extinction phases separately for the CS+ and the CS-. The conditioned response to snake, erotic, and neutral stimuli during acquisition and extinction is shown in Figure 5.

Conditioned stimuli's evaluation. Table 1 shows the mean liking and arousal ratings of the CSs selected for each CS category. No main effect of CS type was found for the liking ratings of the selected CSs, $F(1, 39) = 0.73$, $p = .397$, partial $\eta^2 = .018$, 90% CI [.000, .132]. Likewise, the interaction between CS type and CS category was not statistically significant, $F(1.79, 69.77) = 0.31$, $p = .710$, partial $\eta^2 = .008$, 90% CI [.000, .053]. These results indicate that the selected CS+ and CS- did not statistically differ in terms of rated liking within each CS category. As expected, a significant main effect of CS category for the liking ratings was observed, $F(2, 78) = 284.71$, $p < .001$, partial $\eta^2 = .880$, 90% CI [.835, .902]. Follow-up analyses confirmed that the selected erotic images were more liked than the selected snake images ($p < .001$, $g_{av} = 5.769$, 95% CI [4.494, 7.260]) and the selected neutral colored squares ($p < .001$, $g_{av} = 3.560$, 95% CI [2.699, 4.548]), whereas the selected colored squares were more liked than the selected snake images ($p < .001$, $g_{av} = 1.932$, 95% CI [1.329, 2.598]).

Similarly to the liking ratings, the main effect of CS type for the arousal ratings of the selected CSs was not statistically significant, $F(1, 39) = 1.03$, $p = .316$, partial $\eta^2 = .026$, 90% CI [.000, .148], and no interaction effect between CS type and CS category was found, $F(2, 78) = 0.25$, $p = .779$, partial $\eta^2 = .006$, 90% CI [.000, .040], reflecting that the selected CS+ and CS- did not statistically differ in terms of rated arousal within each CS category. As expected, the CS categories differentially influenced the arousal ratings of the selected CSs, $F(2, 78) = 75.45$, $p < .001$, partial $\eta^2 = .659$, 90% CI [.548, .723]. Follow-up tests showed that the selected snake

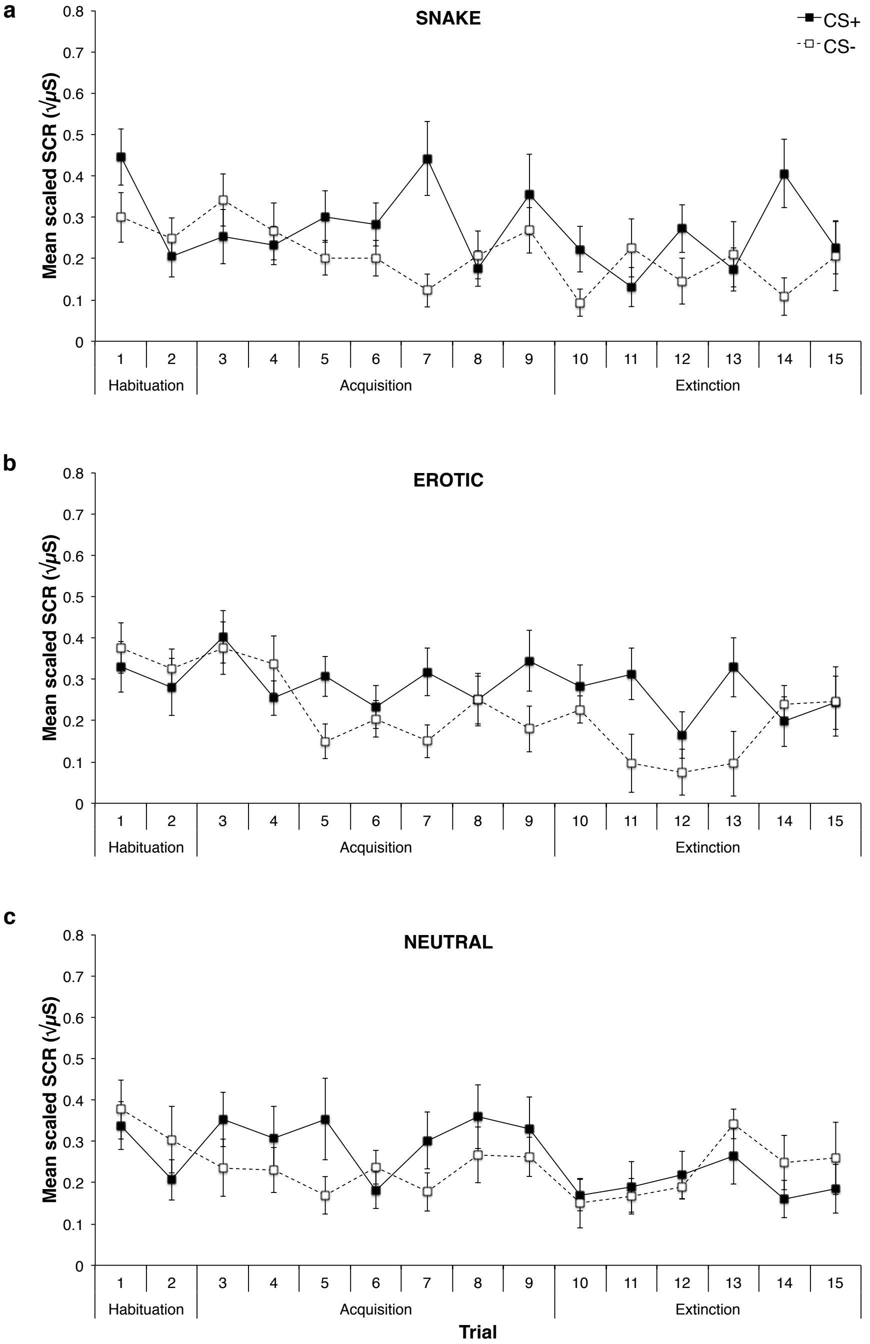

Figure 4. Mean scaled skin conductance response (SCR) to the conditioned stimuli as a function of the conditioned stimulus type (CS+ vs. CS-) across trials in Experiment 3. Mean scaled SCR to (a) snake stimuli, (b) erotic stimuli, and (c) neutral stimuli. Errors bars indicate ± 1 *SEM* adjusted for within-participant designs (Morey, 2008).

images were rated as more arousing than the selected neutral colored squares ($p < .001$, $g_{av} = 0.843$, 95% CI [0.410, 1.301]), and that the selected erotic images were rated as more arousing than the selected colored squares ($p < .001$, $g_{av} = 3.249$, 95% CI [2.441, 4.172]). In addition, the selected erotic images were evaluated as more arousing than the selected snake images ($p < .001$, $g_{av} = 1.523$, 95% CI [1.017, 2.076]).[6]

Table 1

Mean ratings (and standard errors) of the selected conditioned stimuli (CSs) in Experiment 3.

	Snake		Erotic		Neutral	
CS type	Liking	Arousal	Liking	Arousal	Liking	Arousal
CS+	13.66 (2.48)	47.36 (5.33)	93.21 (1.75)	86.85 (2.22)	43.72 (2.70)	22.76 (4.11)
CS-	12.53 (2.58)	49.35 (5.30)	91.99 (1.87)	86.93 (2.13)	43.84 (2.56)	24.97 (3.89)

Skin conductance response. In the habituation phase, no preexisting difference in differential SCRs to the CS categories was observed, $F(2, 78) = 1.06$, $p = .353$, partial $\eta^2 = .026$, 90% CI [.000, .091]. In the acquisition phase, the CR did not statistically differ across the CS categories either, $F(2, 78) = 0.03$, $p = .967$, partial $\eta^2 = .001$, 90% CI [.000, .017], and there was no statistically significant main effect of time, $F(1, 39) = 1.41$, $p = .243$, partial $\eta^2 = .035$, 90% CI [.000, .164]. Similarly, no statistically significant interaction effect of CS category and time was found, $F(1.73, 67.50) = 0.20$, $p = .789$, partial $\eta^2 = .005$, 90% CI [.000, .043], reflecting that there was no statistical difference in the CR acquisition speed among the CS categories. Further analyses revealed that the SCR to the CS+ was greater than to the CS- for snake images, $t(39) =$

2.50, p = .008 (one-tailed), g_{av} = 0.547, 95% CI [0.099, 1.010], erotic images, t(39) = 2.29, p = .014 (one-tailed), g_{av} = 0.502, 95% CI [0.056, 0.962], and neutral colored squares, t(39) = 2.46, p = .009 (one-tailed), g_{av} = 0.540, 95% CI [0.092, 1.002], indicating successful differential conditioning to all three CS categories (see Figure 5). Analysis of the extinction phase showed that the CR persistence was differentially affected by the CS categories, F(1.73, 67.62) = 4.68, p = .016, partial η^2 = .107, 90% CI [.012, .218]. As predicted by the relevance detection model, the CR to both snake and erotic images was more persistent than the CR to neutral colored squares, t(39) = 2.62, p = .006 (one-tailed), g_{av} = 0.496, 95% CI [0.109, 0.898], BF_{10} = 7.777 (see Figure 5). Pairwise comparisons revealed that the CR to snake images was more resistant to extinction than colored squares, t(39) = 2.52, p = .008 (one-tailed), g_{av} = 0.432, 95% CI [0.082, 0.794], BF_{10} = 6.397. The CR to erotic images was likewise more resistant to extinction compared with the CR to colored squares, t(39) = 2.38, p = .011 (one-tailed), g_{av} = 0.504, 95% CI [0.072, 0.950], BF_{10} = 4.815, whereas no statistical difference in CR resistance to extinction emerged between snake images and erotic images, t(39) = -0.51, p = .610 (two-tailed), g_{av} = -0.095, 95% CI [-0.466, 0.274], BF_{10} = 0.261 (see Figure 5).

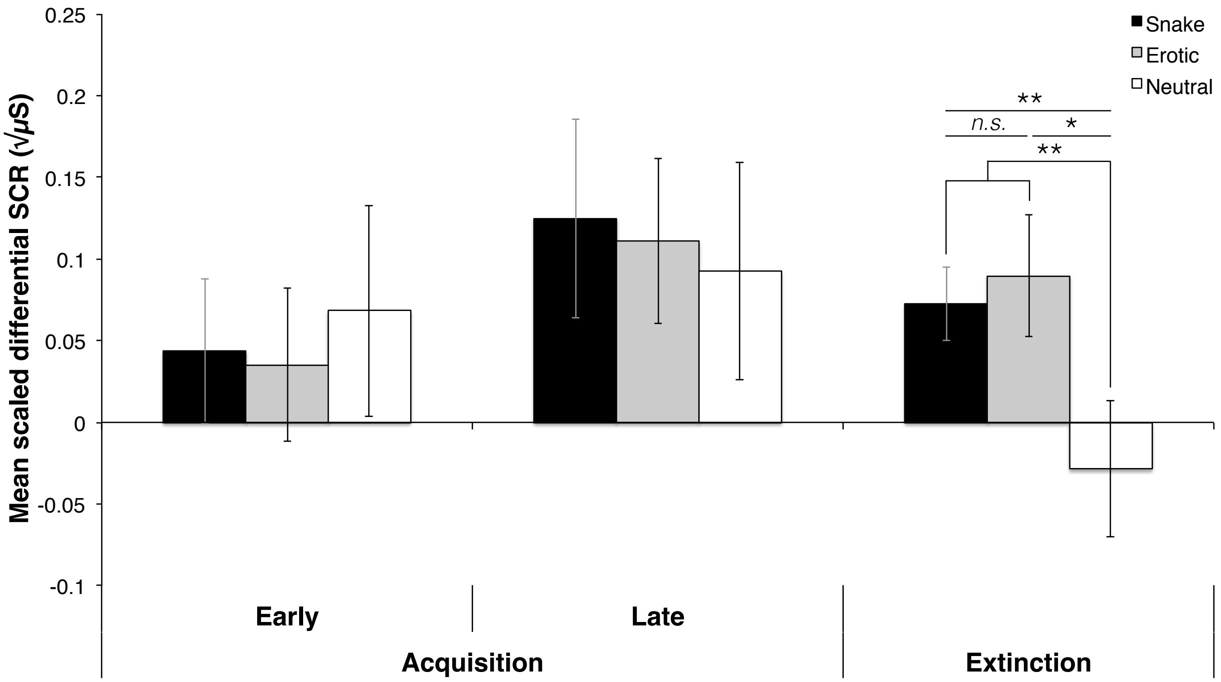

Figure 5. Mean conditioned response (scaled differential skin conductance response [SCR]) as a function of the conditioned stimulus category (snake vs. erotic vs. neutral) during (early and late) acquisition and extinction in Experiment 3. Errors bars indicate ± 1 *SEM* adjusted for within-participant designs (Morey, 2008). Asterisks indicate statistically significant differences between conditions (***p* < .01, **p* < .05, one-tailed) and *ns* indicates a statistically nonsignificant difference.

Subjective ratings. The CS-US contingency ratings showed that the CSs+ were more likely to be associated with the US than the CSs-, $F(1, 39) = 26.62$, $p < .001$, partial $\eta^2 = .406$, 90% CI [.203, .547], while the interaction between CS type and CS category did not reach statistical significance, $F(2, 78) = 2.66$, $p = .076$, partial $\eta^2 = .064$, 90% CI [.000, .152]. Moreover, the CS-US contingency ratings were differentially modulated by the CS categories, $F(2, 78) = 3.55$, $p = .034$, partial $\eta^2 = .083$, 90% CI [.004, .178]. Follow-up tests indicated that erotic images were rated as being more predictive of the US compared with colored squares (p =

.038, g_{av} = 0.479, 95% CI [0.055, 0.917]), but not relative to snake images (p = .890, g_{av} = 0.093, 95% CI [-0.309, 0.497]), whereas snake images were not evaluated as more predictive of the US than colored squares (p = .109, g_{av} = 0.388, 95% CI [0.037, 0.750]; see Figure 6a).

The CS liking ratings revealed that the CSs- were not deemed more pleasant than the CSs+ after the extinction phase, F(1, 39) = 0.56, p = .459, partial η^2 = .014, 90% CI [.000, .122]. Expectedly, a main effect of CS category was found, F(2, 78) = 135.20, p < .001, partial η^2 = .776, 90% CI [.697, .818]. This main effect was not qualified by an interaction with CS type, F(2, 78) = 0.22, p = .801, partial η^2 = .006, 90% CI [.000, .037]. Follow-up analyses showed that erotic images were evaluated as more pleasant than snake images (p < .001, g_{av} = 3.801, 95% CI [2.879, 4.860]) and colored squares (p < .001, g_{av} = 2.654, 95% CI [1.963, 3.438]), while colored squares were rated as more pleasant than snake images (p = .001, g_{av} = 0.797, 95% CI [0.337, 1.279]; see Figure 6b).

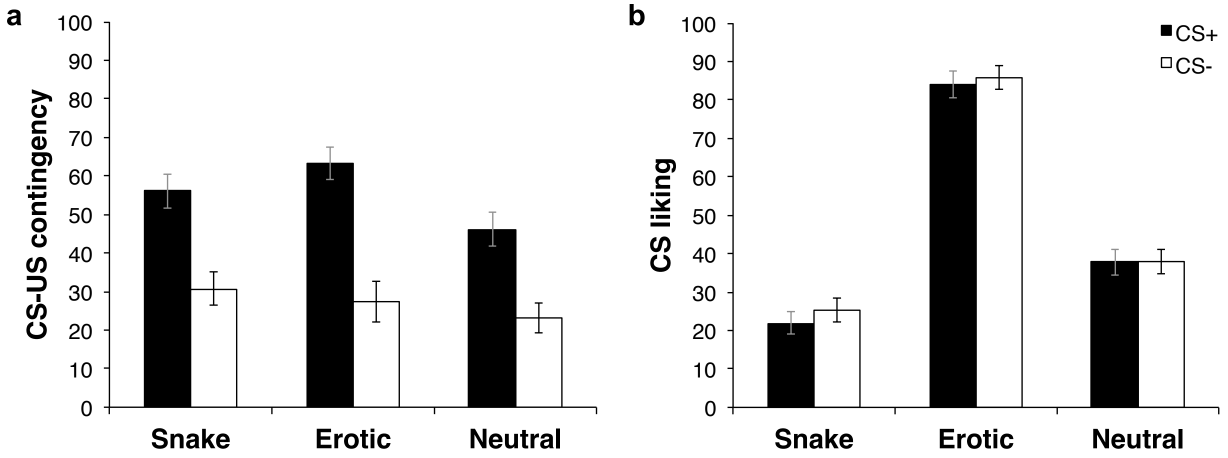

Figure 6. Mean subjective ratings as a function of the conditioned stimulus type (CS+ vs. CS-) and the conditioned stimulus category (snake vs. erotic vs. neutral) in Experiment 3. Mean (a) CS-US contingency ratings and (b) CS liking ratings. Errors bars indicate ± 1 *SEM* adjusted for within-participant designs (Morey, 2008).

Discussion

Experiment 3 replicated and extended the key findings of Experiments 1 and 2 by demonstrating that, like threat-relevant stimuli, positive stimuli with biological relevance to the organism are preferentially conditioned to threat, and, in particular, that these findings generalize beyond baby faces. Results indeed showed that the conditioned response to snake images was more resistant to extinction than the conditioned response to neutral colored squares, which concurs with previous research in the human conditioning literature (e.g., Öhman et al., 1976; Öhman & Mineka, 2001). Of critical importance, the conditioned response to erotic images was likewise more resistant to extinction relative to neutral colored squares, thereby reflecting that both snake and erotic stimuli induced a learning bias during Pavlovian aversive conditioning.

Of note, previous studies by Hamm and colleagues (Hamm, Greenwald, Bradley, & Lang, 1993; Hamm & Stark, 1993; Hamm & Vaitl, 1996) have also used erotic stimuli as CSs in a differential aversive conditioning procedure. Although these studies showed a greater responding in SCR to the CS+ than the CS- across the various stimulus categories used (e.g., threatening animals, mutilations, household objects, and nature scenes) during extinction, none of them seemed to suggest an enhanced resistance to extinction to erotic stimuli, thus contrasting with the current findings. Nonetheless, it is important to note that these studies did not take into account individual preferences for erotic stimuli, and thereby did not directly consider erotic stimuli's affective relevance for the individual's sexual concerns, which may potentially account for the discrepancy between their results and ours.

In line with prior reports in the human conditioning literature (see McNally, 1987; Öhman & Mineka, 2001, for reviews), we observed no reliable differences among the conditioned stimulus categories during the acquisition phase, thus providing no evidence for faster or larger acquisition of a conditioned response to snake images and erotic stimuli compared with neutral stimuli. As for Experiments 1 and 2, this absence of effect might be explained by the specifics of the experimental paradigm used here, in which the various CSs+ predicted relatively unambiguously the US, thereby possibly masking the emergence of differences in the conditioned response acquisition readiness across the conditioned stimulus categories (Ho & Lipp, 2014; Lissek et al., 2006).

Overall, the CSs' ratings during the CSs' selection procedure confirmed that the selected snake stimuli were deemed negative, the selected neutral stimuli neutral, and the selected erotic stimuli positive. The selected erotic and snake stimuli were additionally rated as more arousing than the selected neutral stimuli, whereas the erotic stimuli were also rated as more arousing than the snake stimuli. This latter effect might have occurred because some participants may have misinterpreted the notion of physiological arousal as sexual arousal, thus entailing a possible undervaluation of the actual snake stimuli's arousal value. Importantly, there was however no statistical difference between the selected CS+ and the selected CS- within each stimulus category in the liking and arousal ratings, thereby reflecting an appropriate selection of the conditioned stimuli for each stimulus category.

Subjective ratings collected after extinction revealed that the CSs+ were evaluated as more predictive of the US than the CSs- across the three stimulus categories, indicating that, overall, participants were aware of the contingencies. Moreover, erotic stimuli were deemed more likely to be associated with the US than neutral stimuli regardless of the actual contingencies. This might suggest that expectancy (Davey, 1992) and/or covariation (Tomarken et al., 1989) biases are not selective to associations involving negative threat-relevant stimuli, but can also encompass certain associations between positive biologically relevant stimuli and aversive outcomes. However, this interpretation should be considered with caution because we collected subjective ratings only after extinction, but not after acquisition. In addition, the fact that we did not find such an effect either in Experiment 1 or 2 highlights that further research is needed to explore its determinants, along with its reproducibility and robustness. The CS liking ratings confirmed that erotic stimuli were still evaluated as more pleasant than neutral and snake stimuli after extinction, whereas neutral stimuli were still rated as more pleasant than snake stimuli. In contrast to Experiments 1 and 2 as well as previous reports in the human conditioning literature (e.g., Hamm et al., 1993; Hamm & Vaitl, 1996), no resistant-to-extinction evaluative effects were observed in this experiment. A potential explanation for this discrepancy could be that the addition of CSs' prior ratings during the CSs' selection procedure may have biased participants' postextinction ratings of the same CSs, leading to reduced evaluative conditioning effects (see Lipp & Purkis, 2006).

In brief, Experiment 3 aligns with Experiments 1 and 2 in suggesting that preferential aversive conditioning is not selective to threat-related stimuli, but extends to positive biologically relevant stimuli as well. Experiment 3 thus provides further evidence supporting the hypothesis that stimuli that are relevant to the organism's concerns benefit from preferential emotional learning independently of their valence.

General discussion

In the present study, we aimed at directly testing the predictions of two competing models of emotion with respect to emotional learning; more specifically, we aimed to test the appraisal-based hypothesis that preferential emotional learning is driven by a relevance detection mechanism that is not selective to threat, an hypothesis that is opposed to the fear module hypothesis according to which preferential emotional learning is driven by a fear-specific mechanism that is selective to threat. In order to do so, we investigated whether, similar to threat-relevant stimuli, positive stimuli that are biologically relevant to the organism are likewise preferentially conditioned to threat. In three experiments, we used a differential aversive conditioning paradigm, in which negative biologically relevant stimuli (angry faces, snakes), positive biologically relevant stimuli (baby faces, erotic stimuli), and neutral, less relevant stimuli (neutral faces, colored squares) were used as conditioned stimuli. Taken together, results demonstrate a preferential Pavlovian aversive conditioning to both threat-relevant and positive relevant stimuli.

The enhanced persistence of the learned threat response to threat-relevant stimuli compared with neutral stimuli replicates the basic finding of preferential emotional learning to threat-relevant stimuli consistently reported in the human conditioning literature (e.g., Öhman & Dimberg, 1978; Öhman et al., 1976; Öhman & Mineka, 2001; Olsson et al., 2005; see also Mallan et al., 2013). More importantly, our findings showing an enhanced persistence of the conditioned response to positive relevant stimuli relative to neutral stimuli reflect that positive stimuli with biological relevance are likewise readily associated with a biologically significant event during Pavlovian aversive conditioning, even if this event is naturally aversive. In contradiction to the fear module theory, and somewhat counterintuitively, our hypotheses-driven findings therefore demonstrate that preferential aversive conditioning is not limited to negative stimuli carrying threatening information, but can be extended to positive stimuli that are biologically relevant to the organism. In this respect, our results concur with prior empirical findings in the field of emotional attention, which have shown that attention is not exclusively biased toward negative threatening stimuli, but also orients preferentially and quickly toward positive relevant stimuli (Brosch et al., 2008; Pool, Brosch, et al., 2016). In addition, our data also align with neurobiological evidence suggesting the existence of shared mechanisms across negative and positive valence. Indeed, the encoding and processing of negative and positive stimulus' values has been shown to rely on overlapping brain structures (e.g., Canli, Sivers, Whitfield, Gotlib, & Gabrieli, 2002; Janak & Tye, 2015; Jin, Zelano, Gottfried, & Mohanty, 2015; Namburi et al., 2015; Paton, Belova, Morrison, & Salzman, 2006; Seymour, Daw, Dayan, Singer, & Dolan, 2007; Shabel & Janak, 2009) and neurotransmitter systems (e.g., Matsumoto & Hikosaka, 2009). However, the occurrence of a learning bias to threat-relevant and positive relevant stimuli strongly contrasts with previous research suggesting that preferential aversive conditioning is restricted to specific classes of stimuli that have provided threats to the survival of our ancestors across evolution (Öhman & Dimberg, 1978; Öhman et al., 1976; Öhman & Mineka, 2001; Olsson et al., 2005; Seligman, 1970, 1971). Our findings challenge the view that threat-relevant stimuli are readily associated with an aversive event because they have been correlated with threat through evolution, and alternatively suggest that the key factor underlying preferential emotional learning to threat-relevant stimuli in humans is their high affective relevance to the organism. Our study thereby provides strong support for the existence of a general relevance detection mechanism underlying emotional learning in humans that is common across negative and positive stimuli with biological relevance to the organism.

Nonetheless, it might be proposed that the enhanced persistence of the conditioned response to both threat-relevant and positive relevant stimuli was driven by their a priori negative and positive valence, respectively. Such an account appears nevertheless unlikely because learned threat to happy faces, which represent a typical instance of highly positive stimuli with a relatively low level of general relevance to the organism (Brosch et al., 2008; Pool, Brosch, et al., 2016) and the processing of which is likely to be sensitive to individual differences (Canli et al., 2002), has been shown to rapidly extinguish (e.g., Öhman & Dimberg, 1978; Rowles, Lipp, & Mallan, 2012).

As negative and positive biologically relevant stimuli are typically highly arousing, it could be possible that our findings were mediated by the stimuli's arousal value, the respective contributions of relevance detection and arousal to enhanced aversive conditioning being difficult to disentangle from one another (Montagrin & Sander, 2016; Pool, Brosch, et al., 2016; Sander, 2013). In fact, appraisal theories (e.g., Sander et al., 2003, 2005) posit that stimuli that are appraised as relevant to the organism's concerns also very often elicit a motivational state, which is reflected in a consequent physiological state of arousal that may be felt consciously (Pool, Brosch, et al., 2016). However, the relevance detection and arousal accounts fundamentally differ in terms of the hypothesized psychological mechanisms underlying preferential emotional learning. Whereas the arousal account suggests that the stimulus' arousal value directly drives learning bias, the relevance detection hypothesis explicitly states that the stimulus' affective relevance to the organism's concerns determines learning bias. Accordingly, the mechanism responsible for enhanced emotional learning lies in the emotion elicitation process for the relevance detection account; by contrast, it lies in one component of the emotional response for the arousal account. Indirect evidence in favor of the relevance detection hypothesis comes from a recent meta-analysis on attentional bias for positive stimuli (Pool, Brosch, et al., 2016), which has demonstrated that, whereas both arousal and affective relevance modulated the attentional bias magnitude, only affective relevance remained a significant predictor of the magnitude of the attentional bias when the contributions of arousal and affective relevance were tested by statistically controlling their respective variances, thus implying that relevance detection is more likely to constitute the key mechanism underlying biases in emotional attention than arousal. Additional evidence challenging the arousal account can also be found in studies by Hamm and colleagues (Hamm et al., 1993; Hamm & Stark, 1993; Hamm & Vaitl, 1996), which have shown that highly arousing positive and negative stimuli, without considering their affective relevance to the organism's concerns, did not lead to enhanced resistance to extinction compared with stimuli with a lower arousal level. These results hence indicate that arousal alone might not be sufficient for triggering enhanced Pavlovian aversive conditioning, thereby suggesting that relevance detection provides a more appropriate and plausible mechanism to account for our findings.

Alternatively, it could be argued that preferential emotional learning to threat-relevant stimuli relies on a fear module on the one hand, whereas preferential emotional learning to positive relevant stimuli is triggered by another module dedicated to processing positive, appetitive, or reward-related stimuli with high relevance on the other hand. However, increasing converging evidence shows that the amygdala, which plays a fundamental role in emotional learning (e.g., Büchel et al., 1998; Janak & Tye, 2015; LaBar et al., 1998; LeDoux, 2000, 2012; Phelps & LeDoux, 2005) and was historically conceived as a fear module (Öhman & Mineka, 2001), is not specifically involved in the processing of threat-relevant stimuli, but in the processing of stimuli that are relevant to the organism (Cunningham & Brosch, 2012; Pessoa & Adolphs, 2010; Sander et al., 2003; Sergerie, Chochol, & Armony, 2008), including positive or rewarding stimuli (Gottfried, O'Doherty, & Dolan, 2003; Sergerie et al., 2008). Furthermore, the amygdala has been shown to be a core brain structure of the motivational neural circuits underlying reinforcement learning, directly contributing not only to aversive but also to appetitive reinforcement learning (Averbeck & Costa, 2017). In particular, the amygdala is implicated in the computation of both prediction error (Boll, Gamer, Gluth, Finsterbusch, & Büchel, 2013) and stimulus' associability (Boll et al., 2013; Li, Schiller, Schoenbaum, Phelps, & Daw, 2011), which are fundamental determinants of associative learning in computational models of Pavlovian conditioning (e.g., Li et al., 2011; Niv & Schoenbaum, 2008; Pearce & Hall, 1980; Rescorla & Wagner, 1972). In light of this evidence, we argue that relevance detection constitutes a parsimonious and plausible account of the learning bias to both threat-relevant and positive relevant stimuli during Pavlovian aversive conditioning in humans.

A wider consideration of computational models of Pavlovian conditioning (e.g., Li et al., 2011; Pearce & Hall, 1980; Rescorla & Wagner, 1972) however raises the question as to whether the existence of a learning bias to negative and positive stimuli with biological relevance is adequately captured, and can be characterized, by such Pavlovian learning models. Given the critical role of prediction error and stimulus' associability in associative learning, it could be hypothesized that stimulus' biological relevance may bias Pavlovian conditioning by altering such learning signals. A potential computational learning mechanism whereby the influence of stimulus' biological relevance may operate is stimulus salience, which constitutes a key parameter determining the learning rate and ultimately affecting the impact of prediction error and associability in a number of computational models of conditioning (e.g., Pearce & Hall, 1980; Rescorla & Wagner, 1972).

Stimulus salience traditionally refers to a bottom-up perceptual process based on the stimulus' physical properties (see, e.g., Öhman & Mineka, 2001; Parkhurst, Law, & Niebur, 2002; Pearce & Hall, 1980). Although more salient or intense stimuli – in the sense of physical or perceptual salience – have been shown to be more easily conditioned than less salient or intense stimuli (e.g., Pearce & Hall, 1980; Rescorla, 1988; Rescorla & Wagner, 1972), it has been demonstrated that neutral stimuli with a high perceptual salience do not produce enhanced resistance to extinction compared with neutral stimuli with a low perceptual salience (Öhman et al., 1976), thereby reflecting that physical salience alone provides an insufficient and unlikely explanation for the effects observed in our three experiments (see also McNally, 1987; Öhman & Mineka, 2001). However, stimulus salience has not solely been discussed in the literature as a mere characteristic of the stimulus, but has also been discussed in terms of motivational contingencies relating to the organism's needs and goals (see Cunningham & Brosch, 2012; Öhman & Mineka, 2001; Rescorla, 1988). In this respect, various stimuli can be considered as motivationally salient, such as the threat-relevant and positive relevant stimuli used in our study (see, e.g., Öhman & Mineka, 2001; Parsons et al., 2011; Schultz, 2015). It has been argued that the process of incentive salience is conceptually very closely related to the construct of relevance detection as used in appraisal theories of emotion (see Pool, Sennwald, Delplanque, Brosch, & Sander, 2016; Sennwald, Pool, & Sander, 2017). For instance, it has been suggested that the human amygdala is the key brain system involved in relevance detection (Sander et al., 2003), an idea that is conceptually very similar to the proposal that the amygdala is the key region involved in motivational salience (Cunningham & Brosch, 2012). Of course, the constructs of relevance detection and motivational salience have different conceptual historical roots, and are used in different research traditions but share a fundamental aspect underlying why a post-hoc explanation of our results in terms of motivational salience would closely mirror our a priori prediction in terms of relevance detection: Both constructs suggest that the key factor responsible for our results stems from the interaction between the stimulus and the organism's current concerns.

Critically, our findings of enhanced resistance to extinction of the learned emotional response to both threat-relevant and positive relevant stimuli are however in stark contrast with the predictions of the influential Rescorla-Wagner (Rescorla & Wagner, 1972) and Pearce-Hall (Pearce & Hall, 1980) models of Pavlovian conditioning, as well as previous empirical data from animal research (e.g., Kamin & Gaioni, 1974; Kremer, 1978; Taylor & Boakes, 2002). Although these models predict and account for the accelerated acquisition of the conditioned response to more salient stimuli during conditioning (e.g., Pearce & Hall, 1980; Rescorla, 1988; Rescorla & Wagner, 1972), they also predict that, all else being equal, the conditioned response to more salient stimuli will extinguish faster than the conditioned response to less salient stimuli (see Siddle & Bond, 1988; see also Kamin & Gaioni, 1974; Kremer, 1978; Taylor & Boakes, 2002, for studies in rats providing either direct or indirect support for this prediction). A salience parameter as implemented in the Rescorla-Wagner and Pearce-Hall models therefore does not seem to provide a plausible computational learning mechanism that is able to adequately capture and characterize the influence of the type of stimulus' biological relevance that we investigated in our series of experiments. In line with this view, additional computational analyses of our data using simple reinforcement learning models (Li et al., 2011; Pearce & Hall, 1980; Rescorla & Wagner, 1972; see supplemental materials) suggest that the influence of both negative and positive biologically relevant stimuli, relative to neutral stimuli with less relevance, might be specifically characterized by a lower learning rate for negative prediction error (i.e., when the expected outcome is omitted or when the outcome is less than predicted) that biases inhibitory learning – which includes, without being limited to, extinction learning (Dunsmoor, Niv, Daw, & Phelps, 2015) – through a reduced impact of negative prediction error on associative strength, thus potentially accounting for the enhanced persistence of the conditioned response.

Nonetheless, the computational mechanisms by which the influence of stimulus' affective relevance on Pavlovian conditioning operates remain yet to be better elucidated and characterized.

In conclusion, this series of three experiments suggests that relevance detection drives Pavlovian aversive conditioning in humans. Relevance detection constitutes a rapid (e.g., Grandjean & Scherer, 2008) and flexible (e.g., Moors, 2010) mechanism that enables the organism to adaptively and dynamically trigger the preferential processing and learning of stimuli that are detected as highly relevant. Importantly, the relevance detection account also allows for the accommodation and reinterpretation of existing evidence on preferential aversive conditioning to evolutionary threat stimuli, as these stimuli are a highly relevant signal for the organism. However, a relevance detection mechanism should trigger preferential emotional learning not only to biologically relevant stimuli but also to stimuli that are relevant to the organism's concerns independently of their evolutionary status *per se*. Primary evidence of this point still remains inconclusive. Some studies have shown a similar persistence of learned threat to threatening stimuli from both phylogenetic (i.e., snakes) and ontogenetic (i.e., pointed guns) origin (Flykt, Esteves, & Öhman, 2007; Hugdahl & Johnsen, 1989), while other studies have reported a greater persistence of learned threat to phylogenetically threat-relevant stimuli compared with ontogenetically threat-relevant stimuli (Cook, Hodes, & Lang, 1986; Hugdahl & Kärker, 1981). Further research will thus have to pinpoint whether preferential emotional learning is limited to evolutionary relevant stimuli or extends to stimuli with high relevance to the organism beyond biological and evolutionary considerations. As neural circuits underlying threat-related responses and behaviors have been shown to respond differently to actual threats posed by predators as opposed to standard aversive conditioning paradigms commonly used in laboratory settings (Mobbs & Kim, 2015), another interesting and important avenue for future research will be to investigate whether the role of relevance detection generalizes across more ethologically valid paradigms (e.g., using virtual reality) mirroring the ecological conditions under which threats and rewards typically occur in the organism's natural environment. By postulating a common mechanism of emotional learning not only across negative and positive stimuli but also across aversive and appetitive contingencies, the relevance detection approach offers a new perspective that may contribute to a better understanding of the functioning of human emotional learning, as well as its alteration in specific disorders. Although the generality of a relevance detection mechanism remains to be determined in appetitive conditioning, our study provides new insights into the basic mechanisms underlying emotional learning in humans.

Context of the research

The present set of experiments originates from a research program that aims to investigate the links between the appraisal processes involved in emotion elicitation and the basic mechanisms underlying learning in humans. In this research program, we seek to challenge the dominant view that only threat-related stimuli induce preferential emotional learning by offering an alternative theoretical framework based on appraisal theories of emotion (e.g., Sander et al., 2003, 2005), which holds that emotional learning is driven by a process of relevance detection that is not specific to threat. Our goal is therefore to systematically test the theoretical prediction that stimuli that are detected as highly relevant to the organism's concerns benefit from enhanced Pavlovian conditioning, independently of their intrinsic valence. In this perspective, the findings reported here provide initial evidence for the existence of a relevance detection mechanism underlying emotional learning in humans, and suggest that appraisal theories may offer a promising framework to foster better insights into the understanding of human emotional learning. Ultimately, this framework might also be valuable to account for the high flexibility and large inter-individual differences typically observed in emotional learning across varying contexts and situations, as well as some impairments in this process preceding or following the onset and maintenance of specific emotional disorders. Accordingly, future research will focus on expanding the current findings with the aim of further establishing and characterizing the role of relevance detection in emotional learning.

References

Averbeck, B. B., & Costa, V. D. (2017). Motivational neural circuits underlying reinforcement learning. *Nature Neuroscience, 20*, 505-512. doi:10.1038/nn.4506

Berridge, K. C., & Kringelbach, M. L. (2015). Pleasure systems in the brain. *Neuron, 86*, 646-664. doi:10.1016/j.neuron.2015.02.018

Boll, S., Gamer, M., Gluth, S., Finsterbusch, J., & Büchel, C. (2013). Separate amygdala subregions signal surprise and predictiveness during associative fear learning in humans. *European Journal of Neuroscience, 37*, 758-767. doi:10.1111/ejn.12094

Bradley, M. M., Codispoti, M., Cuthbert, B. N., & Lang, P. J. (2001). Emotion and motivation I: Defensive and appetitive reactions in picture processing. *Emotion, 1*, 276-298. doi:10.1037/1528-3542.1.3.276

Brosch, T., Sander, D., Pourtois, G., & Scherer, K. R. (2008). Beyond fear: Rapid spatial orienting toward positive emotional stimuli. *Psychological Science, 19*, 362-370. doi:10.1111/j.1467-9280.2008.02094.x

Brosch, T., Sander, D., & Scherer, K. R. (2007). That baby caught my eye… Attention capture by infant faces. *Emotion, 7*, 685-689. doi:10.1037/1528-3542.7.3.685

Büchel, C., Morris, J., Dolan, R. J., & Friston, K. J. (1998). Brain systems mediating aversive conditioning: An event-related fMRI study. *Neuron, 20*, 947-957. doi:10.1016/S0896-6273(00)80476-6

Canli, T., Sivers, H., Whitfield, S. L., Gotlib, I. H., & Gabrieli, J. D. E. (2002). Amygdala response to happy faces as a function of extraversion. *Science, 296*, 2191. doi:10.1126/science.1068749

Cook, E. W., III, Hodes, R. L., & Lang, P. J. (1986). Preparedness and phobia: Effects of stimulus content on human visceral conditioning. *Journal of Abnormal Psychology, 95*, 195-207. doi:10.1037/0021-843X.95.3.195

Coppin, G., Delplanque, S., Bernard, C., Cekic, S., Porcherot, C., Cayeux, I., & Sander, D. (2014). Choice both affects and reflects preferences. *The Quarterly Journal of Experimental Psychology, 67*, 1415-1427. doi:10.1080/17470218.2013.863953

Cunningham, W. A., & Brosch, T. (2012). Motivational salience: Amygdala tuning from traits, needs, values, and goals. *Current Directions in Psychological Science, 21*, 54-59. doi:10.1177/0963721411430832

Davey, G. C. L. (1992). An expectancy model of laboratory preparedness effects. *Journal of Experimental Psychology: General, 121*, 24-40. doi:10.1037/0096-3445.121.1.24

Delplanque, S., N'diaye, K., Scherer, K., & Grandjean, D. (2007). Spatial frequencies or emotional effects? A systematic measure of spatial frequencies for IAPS pictures by a discrete wavelet analysis. *Journal of Neuroscience Methods, 165*, 144-150. doi:10.1016/j.jneumeth.2007.05.030

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6*, 274-290. doi:10.1177/1745691611406920

Dunsmoor, J. E., Niv, Y., Daw, N., & Phelps, E. A. (2015). Rethinking extinction. *Neuron, 88*, 47-63. doi:10.1016/j.neuron.2015.09.028

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191. doi:10.3758/BF03193146

Flykt, A., Esteves, F., & Öhman, A. (2007). Skin conductance responses to masked conditioned stimuli: Phylogenetic/ontogenetic factors versus direction of threat? *Biological Psychology, 74*, 328-336. doi:10.1016/j.biopsycho.2006.08.004

Frijda, N. H. (1988). The laws of emotion. *American Psychologist, 43*, 349-358. doi:10.1037/0003-066X.43.5.349

Georgiadis, J. R., & Kringelbach, M. L. (2012). The human sexual response cycle: Brain imaging evidence linking sex to other pleasures. *Progress in Neurobiology, 98*, 49-81. doi:10.1016/j.pneurobio.2012.05.004

Gershman, S. J. (2015). Do learning rates adapt to the distribution of rewards? *Psychonomic Bulletin & Review, 22*, 1320-1327. doi:10.3758/s13423-014-0890-3

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology, 71*, 1-6. doi:10.1016/j.jmp.2016.01.006

Gottfried, J. A., O'Doherty, J., & Dolan, R. J. (2003). Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science, 301*, 1104-1107. doi:10.1126/science.1087919

Grandjean, D., & Scherer, K. R. (2008). Unpacking the cognitive architecture of emotion processes. *Emotion, 8*, 341-351. doi:10.1037/1528-3542.8.3.341

Hamann, S., Herman, R. A., Nolan, C. N., & Wallen, K. (2004). Men and women differ in amygdala response to visual sexual stimuli. *Nature Neuroscience, 7*, 411-416. doi:10.1038/nn1208

Hamm, A. O., Greenwald, M. K., Bradley, M. M., & Lang, P. J. (1993). Emotional learning, hedonic change, and the startle probe. *Journal of Abnormal Psychology, 102*, 453-465. doi:10.1037/0021-843X.102.3.453

Hamm, A. O., & Stark, R. (1993). Sensitization and aversive conditioning: Effects on the startle reflex and electrodermal responding. *Integrative Physiological & Behavioral Science, 28*, 171-176. doi:10.1007/BF02691223

Hamm, A. O., & Vaitl, D. (1996). Affective learning: Awareness and aversion. *Psychophysiology, 33*, 698-710. doi:10.1111/j.1469-8986.1996.tb02366.x

Ho, Y., & Lipp, O. V. (2014). Faster acquisition of conditioned fear to fear-relevant than to nonfear-relevant conditional stimuli. *Psychophysiology, 51*, 810-813. doi:10.1111/psyp.12223

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: A meta-analysis. *Psychological Bulletin, 136*, 390-421. doi:10.1037/a0018916

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65-70.

Hugdahl, K., & Johnsen, B. H. (1989). Preparedness and electrodermal fear-conditioning: Ontogenetic vs phylogenetic explanations. *Behaviour Research and Therapy, 27*, 269-278. doi:10.1016/0005-7967(89)90046-6

Hugdahl, K., & Kärker, A.-C. (1981). Biological vs experiential factors in phobic conditioning. *Behaviour Research and Therapy, 19*, 109-115. doi:10.1016/0005-7967(81)90034-6

Janak, P. H., & Tye, K. M. (2015). From circuits to behaviour in the amygdala. *Nature, 517*, 284-292. doi:10.1038/nature14188

Jeffreys, H. (1961). *The theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.

Jin, J., Zelano, C., Gottfried, J. A., & Mohanty, A. (2015). Human amygdala represents the complete spectrum of subjective valence. *The Journal of Neuroscience, 35*, 15145-15156. doi:10.1523/JNEUROSCI.2450-15.2015

Kagerer, S., Wehrum, S., Klucken, T., Walter, B., Vaitl, D., & Stark, R. (2014). Sex attracts: Investigating individual differences in attentional bias to sexual stimuli. *PLoS ONE, 9*, e107795. doi:10.1371/journal.pone.0107795

Kamin, L. J., & Gaioni, S. J. (1974). Compound conditioned emotional response conditioning with differentially salient elements in rats. *Journal of Comparative and Physiological Psychology, 87*, 591-597. doi:10.1037/h0036989

Kinsey, A. C., Pomeroy, W. B., & Martin, C. E. (1948). *Sexual behavior in the human male*. Philadelphia, PA: W. B. Saunders.

Kremer, E. F. (1978). The Rescorla-Wagner model: Losses in associative strength in compound conditioned stimuli. *Journal of Experimental Psychology: Animal Behavior Processes, 4*, 22-36. doi:10.1037/0097-7403.4.1.22

Kringelbach, M. L., Stark, E. A., Alexander, C., Bornstein, M. H., & Stein, A. (2016). On cuteness: Unlocking the parental brain and beyond. *Trends in Cognitive Sciences, 20*, 545-558. doi:10.1016/j.tics.2016.05.003

LaBar, K. S., & Cabeza, R. (2006). Cognitive neuroscience of emotional memory. *Nature Reviews Neuroscience, 7*, 54-64. doi:10.1038/nrn1825

LaBar, K. S., Gatenby, J. C., Gore, J. C., LeDoux, J. E., & Phelps, E. A. (1998). Human amygdala activation during conditioned fear acquisition and extinction: A mixed-trial fMRI study. *Neuron, 20*, 937-945. doi:10.1016/S0896-6273(00)80475-4

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology, 4*, 863. doi:10.3389/fpsyg.2013.00863

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). *International affective picture system (IAPS): Affective ratings of pictures and instruction manual*. Tech. Rep. No A-8. Gainesville, FL: University of Florida.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. *Cognition and Emotion, 24,* 1377-1388. doi:10.1080/02699930903485076

LeDoux, J. E. (2000). Emotion circuits in the brain. *Annual Review of Neuroscience, 23*, 155-184. doi:10.1146/annurev.neuro.23.1.155

LeDoux, J. E. (2012). Rethinking the emotional brain. *Neuron, 73*, 653-676. doi:10.1016/j.neuron.2012.02.004

LeDoux, J. E. (2014). Coming to terms with fear. *Proceedings of the National Academy of Sciences of the United States of America, 111*, 2871-2878. doi:10.1073/pnas.1400335111

Li, J., Schiller, D., Schoenbaum, G., Phelps, E. A., & Daw, N. (2011). Differential roles of human striatum and amygdala in associative learning. *Nature Neuroscience, 14*, 1250-1252. doi:10.1038/nn.2904

Lipp, O. V., Cronin, S. L., Alhadad, S. S. J., & Luck, C. C. (2015). Enhanced sensitization to animal, interpersonal, and intergroup fear-relevant stimuli (but no evidence for selective one-trial fear learning). *Psychophysiology, 52*, 1520-1528. doi:10.1111/psyp.12513

Lipp, O. V., & Purkis, H. M. (2006). The effects of assessment type on verbal ratings of conditional stimulus valence and contingency judgments: Implications for the extinction of evaluative learning. *Journal of Experimental Psychology: Animal Behavior Processes, 32*, 431-440. doi:10.1037/0097-7403.32.4.431

Lissek, S., Pine, D. S., & Grillon, C. (2006). The strong situation: A potential impediment to studying psychobiology and pharmacology of anxiety disorders. *Biological Psychology, 72*, 265-270. doi:10.1016/j.biopsycho.2005.11.004

Lonsdorf, T. B., Menz, M. M., Andreatta, M., Fullana, M. A., Golkar, A., Haaker, J., …Merz, C. J. (2017). Don't fear 'fear conditioning': Methodological considerations for the design and analysis of studies on human fear acquisition, extinction, and return of fear. *Neuroscience and Biobehavioral Reviews, 77*, 247-285. doi:10.1016/j.neubiorev.2017.02.026

Lorenz, K. (1943). Die angeborenen Formen möglicher Erfahrung [The innate forms of potential experience]. *Zeitschrift für Tierpsychologie, 5*, 235-409. doi:10.1111/j.1439-0310.1943.tb00655.x

Lovibond, P. F., Siddle, D. A. T., & Bond, N. W. (1993). Resistance to extinction of fear-relevant stimuli: Preparedness or selective sensitization? *Journal of Experimental Psychology: General, 122*, 449-461. doi:10.1037/0096-3445.122.4.449

Mallan, K. M., Lipp, O. V., & Cochrane, B. (2013). Slithering snakes, angry men and out-group members: What and whom are we evolved to fear? *Cognition and Emotion, 27*, 1168-1180. doi:10.1080/02699931.2013.778195

Matsumoto, M., & Hikosaka, O. (2009). Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature, 459*, 837-841. doi:10.1038/nature08028

McNally, R. (1987). Preparedness and phobias: A review. *Psychological Bulletin, 101*, 283-303. doi:10.1037/0033-2909.101.2.283

Mobbs, D., & Kim, J. J. (2015). Neuroethological studies of fear, anxiety, and risky decision-making in rodents and humans. *Current Opinion in Behavioral Sciences, 5*, 8-15. doi:10.1016/j.cobeha.2015.06.005

Montagrin, A., & Sander, D. (2016). Emotional memory: From affective relevance to arousal. *Behavioral and Brain Sciences, 39*, e216. doi:10.1017/S0140525X15001879

Moors, A. (2010). Automatic constructive appraisal as a candidate cause of emotion. *Emotion Review, 2*, 139-156. doi:10.1177/1754073909351755

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology, 4*, 61-64. doi:10.20982/tqmp.04.2.p061

Namburi, P., Beyeler, A., Yorozu, S., Calhoon, G. G., Halbert, S. A., Wichmann, R., …Tye, K. M. (2015). A circuit mechanism for differentiating positive and negative associations. *Nature, 520*, 675-678. doi:10.1038/nature14366

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *The Journal of Neuroscience, 32*, 551-562. doi:10.1523/JNEUROSCI.5498-10.2012

Niv, Y., & Schoenbaum, G. (2008). Dialogues on prediction error. *Trends in Cognitive Sciences, 12*, 265-272. doi:10.1016/j.tics.2008.03.006

Öhman, A., & Dimberg, U. (1978). Facial expressions as conditioned stimuli for electrodermal responses: A case of "preparedness"? *Journal of Personality and Social Psychology, 36*, 1251-1258. doi:10.1037/0022-3514.36.11.1251

Öhman, A., Eriksson, A., & Olofsson, C. (1975). One-trial learning and superior resistance to extinction of autonomic responses conditioned to potentially phobic stimuli. *Journal of Comparative and Physiological Psychology, 88*, 619-627. doi:10.1037/h0078388

Öhman, A., Fredrikson, M., Hugdahl, K., & Rimmö, P.-A. (1976). The premise of equipotentiality in human classical conditioning: Conditioned electrodermal responses to potentially phobic stimuli. *Journal of Experimental Psychology: General, 105*, 313-337. doi:10.1037/0096-3445.105.4.313

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

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

Olsson, A., & Phelps, E. A. (2004). Learned fear of "unseen" faces after Pavlovian, observational, and instructed fear. *Psychological Science, 15*, 822-828. doi:10.1111/j.0956-7976.2004.00762.x

Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotions*. New York, NY: Oxford University Press.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42*, 107-123. doi:10.1016/S0042-6989(01)00250-4

Parsons, C. E., Young, K. S., Kumari, N., Stein, A., & Kringelbach, M. L. (2011). The motivational salience of infant faces is similar for men and women. *PLoS ONE, 6*, e20632. doi:10.1371/journal.pone.0020632

Paton, J. J., Belova, M. A., Morrison, S. E., & Salzman, C. D. (2006). The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature, 439*, 865-870. doi:10.1038/nature04490

Pavlov, I. P. (1927). *Conditioned reflexes*. London, UK: Oxford University Press.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not unconditioned stimuli. *Psychological Review, 87*, 532-552. doi:10.1037/0033-295X.87.6.532

Pessoa, L., & Adolphs, R. (2010). Emotion processing and the amygdala: From a ‘low road’ to ‘many roads’ for evaluating biological significance. *Nature Reviews Neuroscience, 11*, 773-783. doi:10.1038/nrn2920

Phelps, E. A., Delgado, M. R., Nearing, K. I., & LeDoux, J. E. (2004). Extinction learning in humans: Role of the amygdala and vmPFC. *Neuron, 43*, 897-905. doi:10.1016/j.neuron.2004.08.042

Phelps, E. A., & LeDoux, J. E. (2005). Contributions of the amygdala to emotion processing: From animal models to human behavior. *Neuron, 48*, 175-187. doi:10.1016/j.neuron.2005.09.025

Pool, E., Brosch, T., Delplanque, S., & Sander, D. (2016). Attentional bias for positive emotional stimuli: A meta-analytic investigation. *Psychological Bulletin, 142*, 79-106. doi:10.1037/bul0000026

Pool, E., Sennwald, V., Delplanque, S., Brosch, T., & Sander, D. (2016). Measuring wanting and liking from animals to humans: A systematic review. *Neuroscience and Biobehavioral Reviews, 63*, 124-142. doi:10.1016/j.neurobiorev.2016.01.006

Prévost, C., McNamee, D., Jessup, R. K., Bossaerts, P., & O'Doherty, J. P. (2013). Evidence for model-based computations in the human amygdala during Pavlovian conditioning. *PLoS Computational Biology, 9*, e1002918. doi:10.1371/journal.pcbi.1002918

Rescorla, R. A. (1988). Behavioral studies of Pavlovian conditioning. *Annual Reviews of Neuroscience, 11*, 329-352. doi:10.1146/annurev.ne.11.030188.001553

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prosky (Eds.), *Classical conditioning II: Current research and theory* (pp. 64-99). New York, NY: Appleton-Century-Crofts.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting or rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225-237. doi:10.3758/PBR.16.2.225

Rowles, M. E., Lipp, O. V., & Mallan, K. M. (2012). On the resistance to extinction of fear conditioned to angry faces. *Psychophysiology, 49*, 375-380. doi:10.1111/j.1469-8986.2011.01308.x

Rupp, H. A., & Wallen, K. (2008). Sex differences in response to visual sexual stimuli: A review. *Archives of Sexual Behavior, 37*, 206-218. doi:10.1007/s10508-007-9217-9

Sander, D. (2013). Models of emotion: The affective neuroscience approach. In J. L. Armony & P. Vuilleumier (Eds.), *The Cambridge Handbook of human affective neuroscience* (pp. 5-53). Cambridge, UK: Cambridge University Press.

Sander, D., Grafman, J., & Zalla, T. (2003). The human amygdala: An evolved system for relevance detection. *Reviews in the Neurosciences, 14*, 303-316. doi:10.1515/REVNEURO.2003.14.4.303

Sander, D., Grandjean, D., & Scherer, K. R. (2005). A systems approach to appraisal mechanisms in emotion. *Neural Networks, 18*, 317-352. doi:10.1016/j.neunet.2005.03.001

Schiller, D., Monfils, M.-H., Raio, C. M., Johnson, D. C., LeDoux, J. E., & Phelps, E. A. (2010). Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature, 463*, 49-53. doi:10.1038/nature08637

Schultz, W. (2015). Neuronal reward and decision signals: From theories to data. *Physiological Review, 95*, 853-951. doi:10.1152/physrev.00023.2014

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461-464. doi:10.1214/aos/1176344136

Seligman, M. E. P. (1970). On the generality of the laws of learning. *Psychological Review, 77*, 406-418. doi:10.1037/h0029790

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

Sennwald, V., Pool, E., Brosch, T., Delplanque, S., Bianchi-Demicheli, F., & Sander, D. (2016). Emotional attention for erotic stimuli: Cognitive and brain mechanisms. *The Journal of Comparative Neurology, 524*, 1668-1675. doi:10.1002/cne.23859

Sennwald, V., Pool, E., Delplanque, S., Brosch, T., Bianchi-Demicheli, F., & Sander, D. (2018). *Inter-individual differences underlie cue-triggered 'wanting' for sexual reward.* Manuscript in preparation.

Sennwald, V., Pool, E., & Sander, D. (2017). Considering the influence of the Pavlovian system on behavior: Appraisal and value representation. *Psychological Inquiry, 28*, 52-55. doi:10.1080/1047840X.2017.1259951

Sergerie, K., Chochol, C., & Armony, J. L. (2008). The role of the amygdala in emotional processing: A quantitative meta-analysis of functional neuroimaging studies. *Neuroscience and Biobehavioral Reviews, 32*, 811-830. doi:10.1016/j.neubiorev.2007.12.002

Seymour, B., Daw, N., Dayan, P., Singer, T., & Dolan, R. (2007). Differential encoding of losses and gains in the human striatum. *The Journal of Neuroscience, 27*, 4826-4831. doi:10.1523/JNEUROSCI.0400-07.2007

Siddle, D. A. T., & Bond, N. W. (1988). Avoidance learning, Pavlovian conditioning, and the development of phobias. *Biological Psychology, 27*, 167-183. doi:10.1016/0301-0511(88)90048-8

Shabel, S. J., & Janak, P. H. (2009). Substantial similarity in amygdala neuronal activity during conditioned appetitive and aversive emotional arousal. *Proceedings of the National Academy of Sciences of the United States of America, 106*, 15031-15036. doi:10.1073/pnas.0905580106

Spector, I. P., Carey, M. P., & Steinberg, L. (1996). The sexual desire inventory: Development, factor structure, and evidence of reliability. *Journal of Sex & Marital Therapy, 22*, 175-190. doi:10.1080/00926239608414655

Stussi, Y., Brosch, T., & Sander, D. (2015). Learning to fear depends on emotion and gaze interaction: The role of self-relevance in fear learning. *Biological Psychology, 109*, 232-238. doi:10.1016/j.biopsycho.2015.06.008

Taylor, K. M., & Boakes, R. A. (2002). Extinction of conditioned taste aversions: Effects of concentration and overshadowing. *The Quarterly Journal of Experimental Psychology B:*

Comparative and Physiological Psychology, 55, 213-239. doi:10.1080/02724990143000270

Tomarken, A. J., Mineka, S., & Cook, M. (1989). Fear-relevant selective associations and covariation bias. *Journal of Abnormal Psychology, 98*, 381-394. doi:10.1037/0021-843X.98.4.381

Van Duuren, M., Kendell-Scott, L., & Stark, N. (2003). Early aesthetic choices: Infant preferences for attractive premature infant faces. *International Journal of Behavioral Development, 27*, 212-219. doi:10.1080/01650250244000218

Zhang, S., Mano, H., Ganesh, G., Robbins, T., & Seymour, B. (2016). Dissociable learning processes underlie human pain conditioning. *Current Biology, 26*, 52-58. doi:10.1016/j.cub.2015.10.066

Footnotes

[1] The descriptively less robust aversive conditioning to neutral faces across the acquisition phase in Experiment 2 was mainly driven by the presence of an outlier (-4.77 *SD* from the mean conditioned response to neutral faces), who strongly conditioned to the neutral face CS-. The one-sample *t* test excluding this outlier indeed revealed a stronger differential conditioning to neutral faces, $t(58) = 3.26$, $p < .001$ (one-tailed), $g_{av} = 0.593$, 95% CI [0.221, 0.975]. However, since we had no a priori reason to exclude this outlier, we kept it in the analyses.

[2] Given the nature of the stimuli used, we also analyzed the SCR data of Experiments 1 and 2 including a gender factor (men vs. women) to explore potential gender differences during conditioning. In Experiment 1, this analysis revealed that men exhibited a greater conditioned response than women across CS categories during the habituation phase, as shown by a main effect of gender, $F(1, 38) = 5.03$, $p = .031$, partial $\eta^2 = .117$, 90% CI [.006, .278]. No other main effect or interaction effect of gender reached statistical significance (all *F*s < 2.65, all *p*s > .07). In Experiment 2, no statistically significant main effect or interaction effect of gender was found (all *F*s < 0.86, all *p*s > .42). These results thus suggest that no gender difference emerged among the CS categories during conditioning.

[3] In order to examine whether angry and baby faces elicited enhanced sensitization in comparison with neutral faces, we performed a repeated measures ANOVA with CS type (CS+ vs. CS-) and CS category (angry vs. baby vs. neutral) as within-participant factors on SCR during the habituation phase both in Experiment 1 and 2. Although our experiments were not explicitly designed to assess selective sensitization effects, such analysis allows for a test thereof when an electric stimulation workup procedure preceding habituation is included, this workup

procedure being supposedly sufficient to induce sensitization (see Lipp et al., 2015). The outcome of these analyses revealed no main effect of CS category either in Experiment 1, $F(2, 78) = 1.41$, $p = .250$, partial $\eta^2 = .035$, 90% CI [.000, .107], or in Experiment 2, $F(2, 118) = 0.77$, $p = .468$, partial $\eta^2 = .013$, 90% CI [.000, .053], thus failing to provide evidence for the occurrence of selective sensitization to angry and baby faces.

[4] A Welch's *t* test for unequal sample sizes supported this interpretation by showing that the mean square-root-transformed unconditioned response in Experiment 2 ($M = 0.72$, $SE = 0.04$) was overall smaller than in Experiment 1 ($M = 1.48$, $SE = 0.08$), $t(62.04) = 8.78$, $p < .001$, $g_s = 1.923$, 95% CI [1.451, 2.418], suggesting that the unconditioned stimulus was indeed less intense in Experiment 2 than in Experiment 1.

[5] IAPS numbers of the snake images used in Experiment 3: 1022, 1026, 1033, 1040, 1050, 1051, 1052, 1070, 1090, 1113, 1114, 1120.

[6] A repeated measures ANOVA with CS type (CS+ vs. CS-) and CS category (Snake vs. Erotic vs. Neutral) as within-participant factors on SCR during the habituation phase however showed no main effect of CS category, $F(1.54, 59.96) = 0.31$, $p = .676$, partial $\eta^2 = .008$, 90% CI [.000, .064], indicating there was no statistical difference between the different CS categories in terms of physiological arousal as measured by SCR. Similarly, no main effect of CS type ($F(1, 39) = 0.41$, $p = .528$, partial $\eta^2 = .010$, 90% CI [.000, .111]) or interaction effect between CS type and CS category ($F(2, 78) = 1.06$, $p = .353$, partial $\eta^2 = .026$, 90% CI [.000, .091]) were found. Of note, the absence of a statistically significant main effect of CS category also did not provide evidence for the occurrence of selective sensitization to snakes and erotic stimuli relative to neutral colored squares.

Supplemental Materials

Enhanced Pavlovian aversive conditioning to positive emotional stimuli

by Y. Stussi, G. Pourtois, and D. Sander

Supplemental Method and Results

Independent rating study

Sixty-three volunteers (49 women and 14 men) aged between 18 to 48 years old (*M* = 27.54 ± 5.73 years) participated in an independent rating study to ensure that the angry faces used in Experiments 1 and 2 were evaluated as negative, the baby faces as positive, and the neutral faces as relatively neutral.

The independent rating study consisted of an online study using qualtrics® (https://www.qualtrics.com), in which the six different stimuli used in Experiments 1 and 2 were presented to participants, accompanied by a visual analog scale (VAS). Participants were asked to rate to what extent the face displayed onscreen was unpleasant or pleasant, the VAS ranging from 0 (*very unpleasant*) to 100 (*very pleasant*). The order of the face presentations was randomized across participants. The stimulus liking ratings were analyzed with a one-way repeated measures analysis of variance (ANOVA) with stimulus category (Anger vs. Baby vs. Neutral) as a within-participant factor. The main effect of stimulus category was followed up with a multiple comparison procedure using Tukey's HSD tests if applicable.

Table S1 reports the mean liking ratings for each stimulus separately. The one-way repeated measures ANOVA revealed that the liking ratings were modulated by the stimulus category, $F(1.66, 102.91) = 127.54$, $p < .001$, partial $\eta^2 = .673$, 90% CI [.583, .729]. Follow-up analyses showed that participants rated the baby faces ($M = 72.12$, $SE = 2.08$) as more pleasant than both the angry faces (*M* = 30.17, *SE* = 2.07; $p < .001$, $g_{av} = 2.519$, 95% CI [1.958, 3.133]) and the neutral faces (*M* = 50.71, *SE* = 1.53; $p < .001$, $g_{av} = 1.462$, 95% CI [1.062, 1.891]), while the neutral faces were evaluated as more pleasant than the angry faces ($p < .001$, $g_{av} = 1.406$, 95% CI [1.031, 1.810]). Overall, the independent rating study thus confirmed that the selected angry faces were evaluated as negative, the selected baby faces as positive, and the selected neutral faces as relatively neutral.

Table S1

Mean liking ratings (and standard errors) of the stimuli used in Experiments 1 and 2 in the independent rating study.

	Angry faces		Baby faces		Neutral faces	
	Face 1	Face 2	Face 1	Face 2	Face 1	Face 2
	34.79 (2.33)	25.54 (2.38)	74.30 (2.54)	69.94 (2.09)	49.37 (1.84)	52.06 (1.81)
Source	RaFD model 23 (Langner et al., 2010)	RaFD model 46 (Langner et al., 2010)	Coppin et al. (2014); Van Duuren et al. (2003)	Coppin et al. (2014); Van Duuren et al. (2003)	RaFD model 15 (Langner et al., 2010)	RaFD model 25 (Langner et al., 2010)

Note. RaFD = Radboud Faces Database.

Unconditioned response analysis

Across the three experiments, we analyzed the unconditioned response (UR) to the unconditioned stimulus (US; i.e., electric stimulation) using two-way repeated measures ANOVAs with CS category (Anger vs. Baby vs. Neutral in Experiments 1 and 2, Snake vs. Erotic vs. Neutral in Experiment 3) and US trial (US trial 1 vs. US trial 2 vs. US trial 3 vs. US trial 4 vs. US trial 5) as within-participant factors in order to explore whether the CS categories differentially modulated the UR, and to investigate the UR changes across trials. Due to missing values on some trials, 33 participants could be included in the UR analysis in Experiment 1, 52 in Experiment 2, and 38 participants in Experiment 3.

In Experiment 1, the UR was not differentially influenced by the CS categories, $F(2, 64) = 1.01$, $p = .369$, partial $\eta^2 = .031$, 90% CI [.000, .106], and did not significantly change across trials, $F(2.84, 91.02) = 1.80$, $p = .155$, partial $\eta^2 = .053$, 90% CI [.000, .120] (see Figure S1a). Similarly, no statistically significant interaction between CS category and US trial was observed, $F(5.64, 180.34) = 1.31$, $p = .259$, partial $\eta^2 = .039$, 90% CI [.000, .066].

In Experiment 2, we found no statistically significant main effect of the CS categories on the UR, $F(2, 102) = 0.27$, $p = .764$, partial $\eta^2 = .005$, 90% CI [.000, .033]. In contrast with Experiment 1, we observed a statistically significant main effect of US trial, $F(2.86, 145.95) = 20.18$, $p < .001$, partial $\eta^2 = .284$, 90% CI [.175, .365], reflecting that the UR decreased over trials (see Figure S1b). This main effect was not qualified by an interaction with the CS categories, $F(6.82, 348.07) = 0.70$, $p = .671$, partial $\eta^2 = .013$, 90% CI [.000, .018].

In Experiment 3, the UR was not modulated by the CS categories, $F(2, 74) = 1.69$, $p = .191$, partial $\eta^2 = .044$, 90% CI [.000, .123]. However, a statistically significant main effect of US trial emerged, $F(2.91, 107.51) = 11.55$, $p < .001$, partial $\eta^2 = .238$, 90% CI [.115, .330], indicating that the UR decreased across trials (see Figure S1c). This main effect was not qualified by a higher order interaction with CS category, $F(7.35, 272.03) = 1.07$, $p = .385$, partial $\eta^2 = .028$, 90% CI [.000, .040].

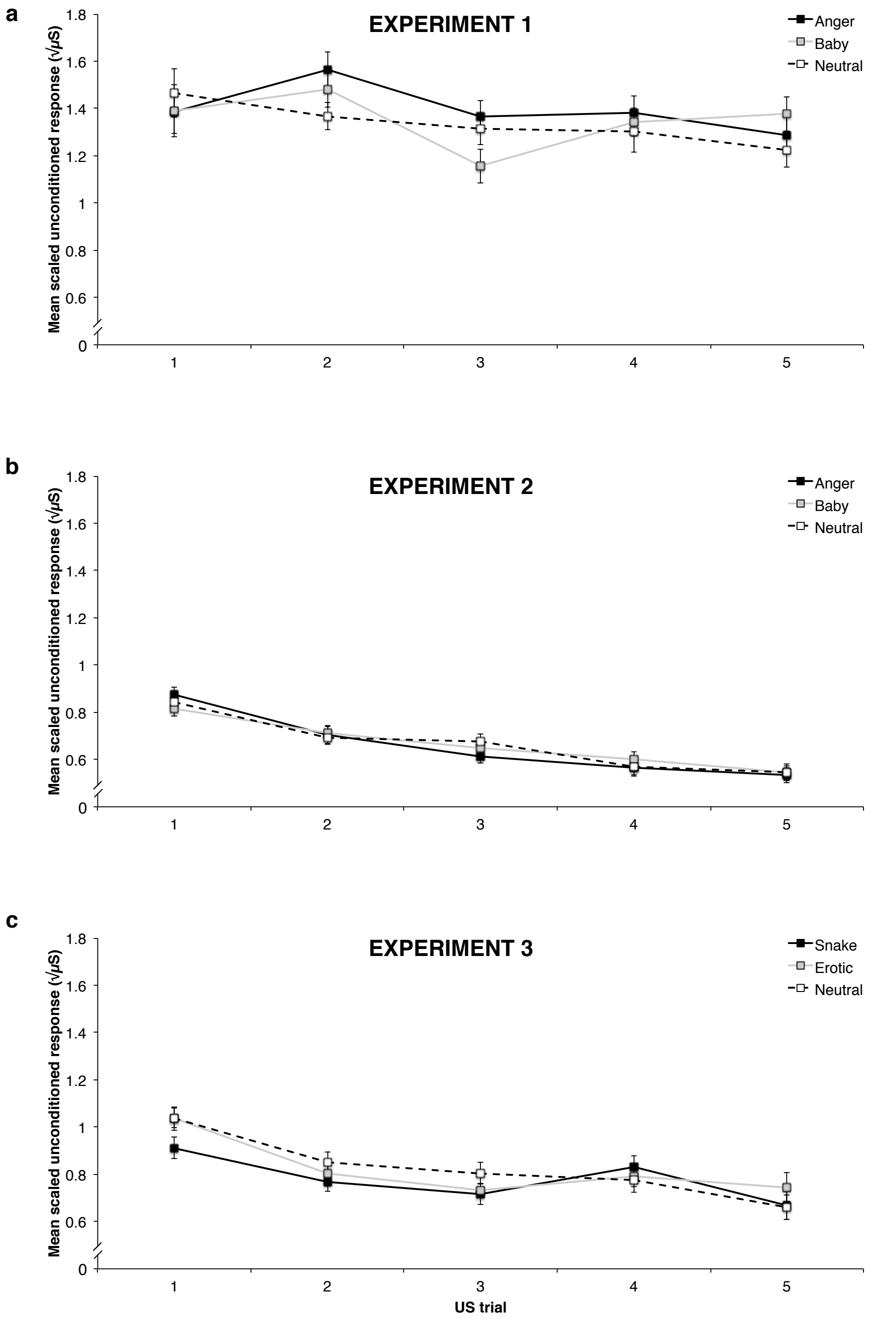

Figure S1. Mean scaled unconditioned response to the unconditioned stimulus (US; electric stimulation) as a function of the conditioned stimulus category and unconditioned stimulus trial in (a) Experiment 1, (b) Experiment 2, and (c) Experiment 3. Error bars indicate ± 1 *SEM* adjusted for within-participant designs (Morey, 2008).

Pavlovian learning models

We constructed simple reinforcement learning models (Li, Schiller, Schoenbaum, Phelps, & Daw, 2011; Pearce & Hall, 1980; Rescorla & Wagner, 1972) to characterize the influence of negative and positive stimuli with biological relevance (i.e., angry faces/snake images and baby faces/erotic stimuli, respectively), relative to neutral stimuli with less relevance (i.e., neutral faces/colored squares), on Pavlovian aversive conditioning. We fitted these models to the SCR data for each CS category separately for parameter estimation and model comparison, and we then compared the parameter estimates of the best-fitting model for each CS category.

Rescorla-Wagner model. The Rescorla-Wagner model (Rescorla & Wagner, 1972) is a classical and standard account of associative learning, in which learning is directly driven by the discrepancy between the actual and the predicted outcome, that is by prediction error. In this model, the value (or associative strength) V at trial t + 1 of a given conditioned stimulus j is updated based on the sum of the current expected value V_j at trial t, and the prediction error between the expected value V_j and the outcome R at trial t, weighted by a constant learning rate α:

$$V_j(t+1)=V_j(t)+\alpha\cdot(R(t)-V_j(t))$$

where the learning rate α is a free parameter within the range [0, 1]. If the unconditioned stimulus was delivered on the current trial t, $R(t) = 1$, else $R(t) = 0$.

Hybrid model. The hybrid model introduced by Li et al. (2011) combines both the Rescorla-Wagner model and the Pearce-Hall model (Pearce & Hall, 1980), where the Rescorla-Wagner algorithm is implemented for error-driven value update, and the Pearce-Hall associability mechanism is substituted for the constant learning rate, thus acting as a dynamic learning rate. According to the Pearce-Hall algorithm, the conditioned stimulus' associability decreases when the conditioned stimulus correctly and reliably predicts the actual outcome, whereas it increases when the conditioned stimulus does not reliably predict the actual outcome. In the hybrid model, the value *V* of a given conditioned stimulus *j* is updated as follows:

$$V_j(t+1)=V_j(t)+\kappa\cdot\alpha_j(t)\cdot(R(t)-V_j(t))$$

$$\alpha_j(t+1)=\eta\cdot\left|R(t)-V_j(t)\right|+(1-\eta)\cdot\alpha_j(t)$$

where the initial associability α_0, the learning rate κ, and the weighting factor η are free parameters within the range [0, 1]. If the unconditioned stimulus was delivered on the current trial t, $R(t) = 1$, else $R(t) = 0$.

Rescorla-Wagner model with dual learning rates. As we predicted that both negative and positive biologically relevant stimuli would induce a learning bias, as reflected by an enhanced resistance to extinction and consequently a diminished inhibitory learning, we also implemented a dual-learning-rate model using the Rescorla-Wagner algorithm (see, e.g., Gershman, 2015; Niv, Edlund, Dayan, & O'Doherty, 2012), where the learning rate differed as a function of whether the prediction error was positive (i.e., excitatory learning) or negative (i.e., inhibitory learning). To this end, we modified the Rescorla-Wagner model to allow for different learning rates for positive prediction error and for negative prediction error. In the dual-learning-rate Rescorla-Wagner model, the value *V* of a given conditioned stimulus *j* is updated as follows:

$$V_j(t+1)=\begin{cases} V_j(t)+\alpha^{+}\cdot(R(t)-V_j(t))\ if\ R(t)-V_j(t)>0 \\ V_j(t)+\alpha^{-}\cdot(R(t)-V_j(t))\ if\ R(t)-V_j(t)<0 \end{cases}$$

where the learning rate for positive prediction error α^{+} and the learning rate for negative prediction error α^{-} are free parameters within the range [0, 1]. If the unconditioned stimulus was delivered on the current trial t, $R(t) = 1$, else $R(t) = 0$.

Hybrid model with dual learning rates. We additionally considered a modified hybrid model implementing dual learning rates by allowing for different learning rates for positive prediction error and for negative prediction error. In this model, the value *V* of a given conditioned stimulus *j* is updated as follows:

$$V_j(t+1)=\begin{cases} V_j(t)+\kappa^{+}\cdot\alpha_j(t)\cdot(R(t)-V_j(t))\ if\ R(t)-V_j(t)>0 \\ V_j(t)+\kappa^{-}\cdot\alpha_j(t)\cdot(R(t)-V_j(t))\ if\ R(t)-V_j(t)<0 \end{cases}$$

$$\alpha_j(t+1)=\eta\cdot\left|R(t)-V_j(t)\right|+(1-\eta)\cdot\alpha_j(t)$$

where the initial associability α_0, the learning rate for positive prediction error κ^{+}, the learning rate for negative prediction error κ^{-}, and the weighting factor η are free parameters within the range [0, 1]. If the unconditioned stimulus was delivered on the current trial *t*, $R(t) = 1$, else $R(t) = 0$.

Model and parameter fitting. The free parameters of the models were optimized using maximum a posteriori estimation, which found the set of parameters maximizing the probability of individual participant’s trial-by-trial normalized (i.e., scaled and square-root-transformed) skin conductance response (SCR) measured following the conditioned stimulus (CS) given the model, constrained by regularizing priors (see Gershman, 2016; Niv et al., 2012). All the free parameters were constrained with a Beta (1.2, 1.2) prior distribution favoring a normal distribution of the parameter estimates. For the Rescorla-Wagner model (RW[*V*]) and the dual-learning-rate Rescorla-Wagner model (dual RW[*V*]), the trial-by-trial time series of CS values *V*(*t*) was used to optimize the free parameters; for the Hybrid model and the dual-learning-rate Hybrid model, the free parameters were optimized separately for each possible combination using the trial-by-trial time series of CS values *V*(*t*) (Hybrid[*V*] and dual Hybrid[*V*]), the trial-by-trial time series of

CS associability $\alpha(t)$ (Hybrid[α] and dual Hybrid[α]), or the combination of both (Hybrid[V+α] and dual Hybrid[V+α]; see Li et al., 2011; Zhang et al., 2016). Initial values (V_0) for each CS were set to 0.5, as participants expected to receive electric stimulations due to the work-up procedure and the instructions. The models were fit using a separate set of free parameters for each participant (i) across all trials, and (ii) separately for each CS category (Boll, Gamer, Gluth, Finsterbusch, & Büchel, 2013), thereby allowing for comparing the best-fitting parameter estimates among the different CS categories.

Model comparison. Model comparison was conducted using Bayesian information criterion (BIC; Schwarz, 1978; see also, e.g., Zhang et al., 2016), which quantitatively measures the models' goodness of fit, while taking into account and penalizing for the number of free parameters included in each model. The BIC value was calculated for each model averaged across participants using models with individual participant's parameter estimates. To ensure that the models outperformed a model with random predictions, we also compared the models against a baseline model, in which the value $V_j(t)$ and the prediction error were updated at each trial by adding random noise from a uniform random distribution within the range [-0.1, 0.1] (Prévost, McNamee, Jessup, Bossaerts, & O'Doherty, 2013). The BIC values for each model across the three experiments are reported in Table S2.

Table S2

Goodness of fit to skin conductance responses for individual models using the mean Bayesian Information Criterion (BIC) in Experiment 1 (N = 40), Experiment 2 (N = 59), and Experiment 3 (N = 40).

		Model								
Exp.	CS category	RW(*V*)	Dual RW(*V*)	Hybrid (*V*)	Hybrid (*α*)	Hybrid (*V*+*α*)	Dual Hybrid (*V*)	Dual Hybrid (*α*)	Dual Hybrid (*V*+*α*)	Baseline
1	All	41.84	38.22	49.37	40.74	41.45	45.40	43.90	44.42	51.78
	Anger	16.28	16.13	22.78	18.18	18.66	22.06	20.60	21.26	21.52
	Baby	13.65	13.50	19.49	16.47	17.30	19.32	17.93	18.38	19.51
	Neutral	7.93	7.51	14.36	9.09	9.23	13.02	11.89	12.13	13.11
2	All	41.49	35.84	48.77	39.46	39.82	42.88	42.58	42.61	51.34
	Anger	11.08	9.89	17.19	12.43	12.72	15.72	14.73	15.04	16.84
	Baby	12.02	10.86	18.42	14.21	14.72	16.91	15.93	16.21	17.40
	Neutral	12.24	11.46	18.28	13.72	13.75	16.64	15.68	15.93	18.14
3	All	53.75	53.74	61.46	57.27	57.38	61.55	60.35	60.63	61.70
	Snake	17.67	18.86	24.24	21.87	21.98	24.99	23.20	24.37	21.75
	Erotic	16.28	16.41	22.26	19.08	19.43	22.30	21.18	21.67	21.62
	Neutral	15.26	17.32	21.74	19.41	20.08	23.46	21.73	22.85	20.08

Note. Exp. = Experiment, RW = Rescorla-Wagner model, *V* = model values, *α* = associabilities, Dual = dual learning rates.

Relationship between modeled learning signals and participants' normalized skin conductance responses. To investigate whether and to what extent modeled learning signals from the optimized model predicted participants' trial-by-trial normalized SCRs, we computed a linear regression, in which we regressed value and prediction error time series generated using individual parameter estimates from the best-fitting model and averaged across participants against the trial-by-trial normalized SCRs averaged across participants. Across the three experiments, the best-fitting model based on all trials consisted of the Rescorla-Wagner model implementing dual learning rates (see Table S2). In Experiment 1, the results of the multiple linear regression analysis showed that value and prediction error signals generated from the Rescorla-Wagner model with dual learning rates explained a statistically significant amount of variance of trial-by-trial normalized SCRs ($R^2 = .361$, $R^2_{adj} = .346$, $F(2, 87) = 24.56$, $p < .001$). Value signals were found to statistically significantly predict trial-by-trial normalized SCRs, $\beta = .42$, $t(87) = 7.01$, $p < .001$ (see Figure S2a), which was not the case for prediction error signals, $\beta = .01$, $t(87) = 0.32$, $p = .751$.

In Experiment 2, value and prediction error signals generated from the Rescorla-Wagner model with dual learning rates likewise explained a statistically significant portion of variance of the trial-by-trial normalized SCRs ($R^2 = .426$, $R^2_{adj} = .413$, $F(2, 87) = 32.31$, $p < .001$). As in Experiment 1, trial-by-trial normalized SCRs were predicted by value signals, $\beta = .41$, $t(87) = 8.03$, $p < .001$ (see Figure S2b), but not by prediction error signals, $\beta = -.001$, $t(87) = -0.04$, $p = .969$.

In Experiment 3, the multiple linear regression indicated that value and prediction errors signals generated from the dual-learning-rate Rescorla-Wagner model explained a statistically significant, though considerably lower, amount of variance of trial-by-trial normalized SCRs

($R^2 = .251$, $R^2_{adj} = .234$, $F(2, 87) = 14.62$, $p < .001$). Value signals statistically significantly predicted trial-by-trial normalized SCRs, $\beta = .23$, $t(87) = 5.21$, $p < .001$ (see Figure S2c), while prediction error signals were only a marginally significant predictor, $\beta = .05$, $t(87) = 1.82$, $p = .072$.

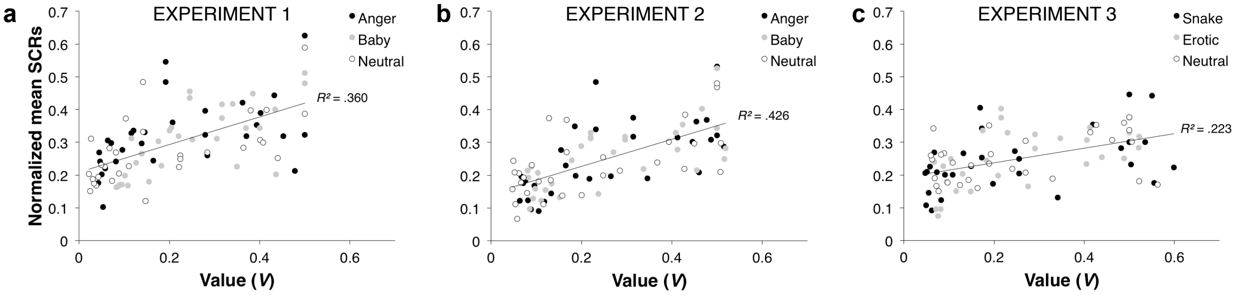

Figure S2. Relationship between modeled value (*V*) and trial-by-trial normalized skin conductance responses (SCRs) averaged across participants using the individual best-fitting parameters for the Rescorla-Wagner model implementing dual learning rates in (a) Experiment 1, (b) Experiment 2, and (c) Experiment 3. The curve represents the best-fitting line using least squares estimation.

Parameter estimates analyses. As model comparison using the BIC indicated that the Rescorla-Wagner model implementing dual learning rates provided the best fit to the data based on all trials compared with the other models in all the three experiments (see Table S2), we therefore analyzed the estimated parameters from this model for each CS category. In Experiment 1, a one-way ANOVA with CS category (Anger vs. Baby vs. Neutral) as a within-participant factor on the learning rate parameter estimates for positive prediction error showed no statistically significant difference between the CS categories, $F(2, 78) = 1.44$, $p = .243$, partial $\eta^2 = .036$, 90% CI [.000, .108] (see Figure S3a). In contrast, the CS categories differentially influenced the learning rate parameter estimates for negative prediction error, $F(2, 78) = 3.87$, $p = .025$, partial $\eta^2 = .090$, 90% CI [.006, .187]. A planned contrast analysis revealed that the learning rate for negative prediction error was lower for both angry (contrast weight: -1) and baby (contrast weight: -1) faces than for neutral faces (contrast weight: +2), $t(39) = 2.71$, $p = .005$ (one-tailed), $g_{av} = 0.596$, 95% CI [0.145, 1.064], $BF_{10} = 9.356$ (see Figure S3a), suggesting that angry and baby faces biased inhibitory learning through a diminished impact of negative prediction error. Further pairwise comparisons showed that the estimated learning rate for negative prediction error was lower for baby faces (contrast weight: -1) than for neutral faces (contrast weight: +1), $t(39) = 2.59$, $p = .007$ (one-tailed), $g_{av} = 0.611$, 95% CI [0.127, 1.112], $BF_{10} = 7.224$ (see Figure S3a), while it was marginally lower for angry faces (contrast weight: -1) relative to neutral faces (contrast weight: +1) with respect to the corrected alpha level for this contrast ($\alpha = .025$) using the Holm-Bonferroni sequential procedure (Holm, 1979), $t(39) = 1.91$, $p = .032$ (one-tailed), $g_{av} = 0.382$, 95% CI [-0.021, 0.794], $BF_{10} = 2.112$ (see Figure S3a). The estimated learning rate for negative prediction error did not statistically differ for angry faces (contrast weight: -1) compared with baby faces (contrast weight: +1), $t(39) = -1.06$, $p = .293$ (two-tailed), $g_{av} = -0.257$, 95% CI [-0.746, 0.225], $BF_{10} = 0.381$ (see Figure S3a).

In Experiment 2, one participant was removed from the Pavlovian learning models analyses since their individual parameters for baby faces could not be estimated due to a lack of SCR to all the baby face CSs during the whole experiment. A one-way ANOVA with CS category (Anger vs. Baby vs. Neutral) as a within-participant factor on the learning rate parameter for positive prediction error revealed no statistically significant difference between the CS categories, $F(2, 116) = 0.28$, $p = .757$, partial $\eta^2 = .005$, 90% CI [.000, .030] (see Figure S3b). However, the learning rate for negative prediction error parameter estimates were differentially modulated by the CS categories, $F(2, 116) = 4.23$, $p = .017$, partial $\eta^2 = .068$, 90% CI [.007, .142]. Both angry (contrast weight: -1) and baby (contrast weight: -1) faces exhibited a lower learning rate for negative prediction error than neutral faces (contrast weight: +2), $t(58) = 2.80$, $p = .003$ (one-tailed), $g_{av} = 0.433$, 95% CI [0.120, 0.754], $BF_{10} = 11.487$ (see Figure S3b), reflecting that angry and baby faces biased inhibitory learning. Further comparisons showed that the estimated learning rate for negative prediction error was lower for angry faces (contrast weight: -1) than for neutral faces (contrast weight: +1), $t(58) = 3.03$, $p = .002$ (one-tailed), $g_{av} = 0.465$, 95% CI [0.153, 0.786], $BF_{10} = 19.866$ (see Figure S3b), whereas it was marginally lower for baby faces (contrast weight: -1) compared with neutral faces (contrast weight: +1) with respect to the corrected alpha level for this contrast ($\alpha = .025$) (Holm, 1979), $t(58) = 1.92$, $p = .030$ (one-tailed), $g_{av} = 0.318$, 95% CI [-0.014, 0.656], $BF_{10} = 1.922$ (see Figure S3b). The estimated learning rate for negative prediction error for angry faces (contrast weight: -1) did not statistically differ from that for baby faces (contrast weight: +1), $t(58) = 0.77$, $p = .446$ (two-tailed), $g_{av} = 0.126$, 95% CI [-0.200, 0.455], $BF_{10} = 0.256$ (see Figure S3b).

In Experiment 3, analysis of the estimated learning rate for positive prediction error showed that the main effect of CS category did not reach statistical significance, $F(2, 78) = 2.40$, $p = .098$, partial $\eta^2 = .058$, 90% CI [.000, .143] (see Figure S3c). In contrast to Experiments 1 and 2, the estimated learning rate for negative prediction error was likewise not differentially modulated by the CS categories, $F(2, 78) = 0.50$, $p = .606$, partial $\eta^2 = .013$, 90% CI [.000, .061] (see Figure S3c).

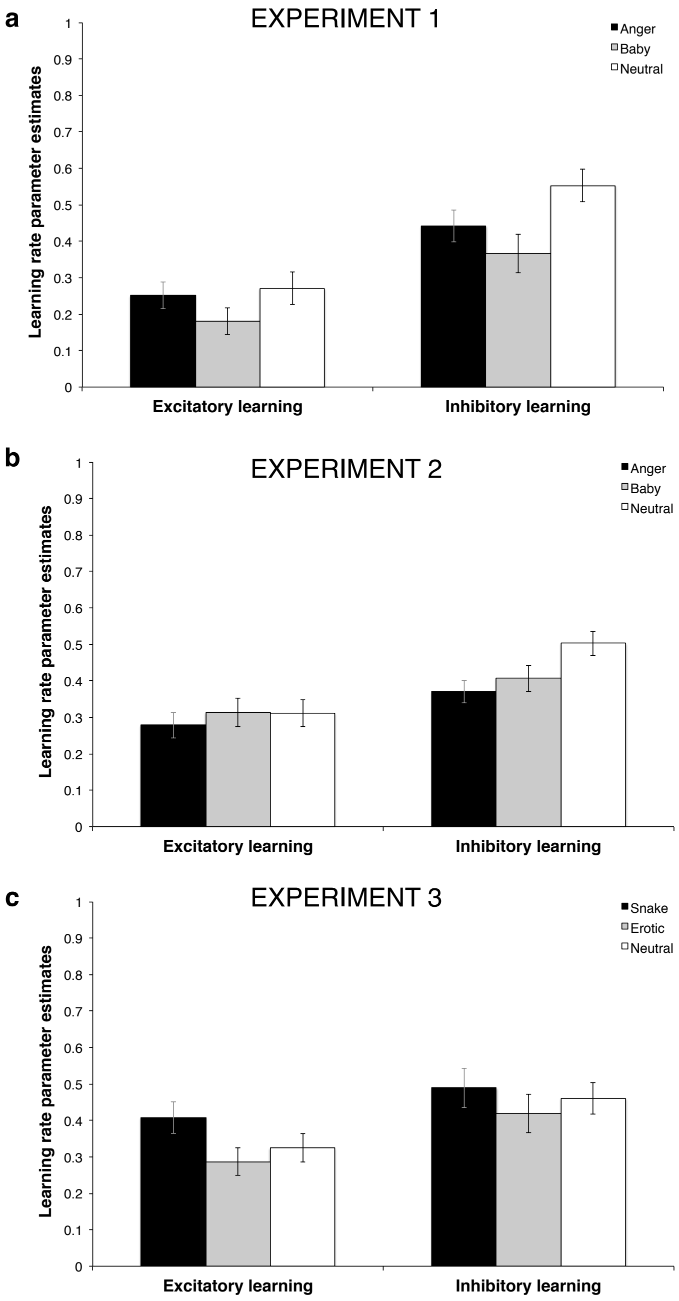

Figure S3. Learning rate parameter estimates of the Rescorla-Wagner model implementing dual learning rates using individual best-fitting parameters for positive prediction error (excitatory learning) and negative prediction error (inhibitory learning) as a function of the conditioned stimulus category in (a) Experiment 1, (b) Experiment 2, and (c) Experiment 3. Error bars indicate ± 1 *SEM* adjusted for within-participant designs (Morey, 2008).

Altogether, the computational analyses using simple reinforcement learning models in Experiments 1 and 2 suggest that the influence of stimulus' biological relevance on Pavlovian aversive conditioning could be best characterized by a lower inhibitory learning rate diminishing the impact of negative prediction error on associative strength, thereby reflecting a learning bias that may account for the enhanced persistence of the conditioned response to both negative and positive stimuli with biological relevance compared with the conditioned response to neutral stimuli with less relevance. These findings thus seem to provide further evidence for the existence of a shared mechanism underlying preferential Pavlovian aversive conditioning in humans that is common across negative and positive relevant stimuli, as predicted by the relevance detection hypothesis. However, these effects were not observed in Experiment 3, where no statistical difference in learning rate for either positive or negative prediction error across the CS categories was found, possibly because of somewhat noisier SCR data, as suggested by the reduced fit to the data observed in this experiment (see Figure S2). For these reasons and as the present findings represent only a first attempt to characterize at the computational level the influence of stimulus' biological relevance on Pavlovian conditioning in humans, it is important to highlight that further research is needed to better outline the computational characterization of the influence of stimulus' affective relevance on Pavlovian learning.

Exploratory correlational analysis in Experiment 3

In Experiment 3, we carried out an exploratory correlational analysis using Pearson's correlation coefficients to test whether the participants' CR to erotic stimuli during acquisition and extinction were associated with their dyadic, solitary, and general sexual desire measured with the Sexual Desire Inventory 2 (Spector, Carey, & Steinberg, 1996). One participant was excluded from the correlational analysis between participants' solitary and general sexual desire and their CR to erotic images during acquisition and extinction due to missing data preventing the computation of his solitary and general sexual desire score.

The correlational analysis did not show that participants' dyadic sexual desire was associated with their CR to erotic images during the acquisition (*r*(38) = -.121, *p* = .457, 95% CI [-0.416, 0.197]) or extinction (*r*(38) = -.125, *p* = .441, 95% CI [-0.420, 0.194]) phases. Similarly, no significant correlation was found between participants' solitary sexual desire and their CR to erotic images during acquisition (*r*(37) = .172, *p* = .296, 95% CI [-0.151, 0.462]) or extinction (*r*(37) = -.042, *p* = .798, 95% CI [-0.352, 0.277]). Furthermore, participants' general sexual desire did not correlate with their CR to erotic images in the acquisition phase (*r*(37) = .019, *p* = .911, 95% CI [-0.298, 0.332]) or in the extinction phase (*r*(37) = -.116, *p* = .482, 95% CI [-0.416, 0.207]).

References

Boll, S., Gamer, M., Gluth, S., Finsterbusch, J., & Büchel, C. (2013). Separate amygdala subregions signal surprise and predictiveness during associative fear learning in humans. *European Journal of Neuroscience, 37*, 758-767. doi:10.1111/ejn.12094

Coppin, G., Delplanque, S., Bernard, C., Cekic, S., Porcherot, C., Cayeux, I., & Sander, D. (2014). Choice both affects and reflects preferences. *The Quarterly Journal of Experimental Psychology, 67*, 1415-1427. doi:10.1080/17470218.2013.863953

Gershman, S. J. (2015). Do learning rates adapt to the distribution of rewards? *Psychonomic Bulletin & Review, 22*, 1320-1327. doi:10.3758/s13423-014-0890-3

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology, 71*, 1-6. doi:10.1016/j.jmp.2016.01.006

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65-70.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. *Cognition & Emotion, 24,* 1377-1388. doi:10.1080/02699930903485076

Li, J., Schiller, D., Schoenbaum, G., Phelps, E. A., & Daw, N. (2011). Differential roles of human striatum and amygdala in associative learning. *Nature Neuroscience, 14*, 1250-1252. doi:10.1038/nn.2904

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology, 4*, 61-64. doi:10.20982/tqmp.04.2.p061

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience, 32*, 551-562. doi:10.1523/JNEUROSCI.5498-10.2012

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not unconditioned stimuli. *Psychological Review, 87*, 532-552. doi:10.1037/0033-295X.87.6.532

Prévost, C., McNamee, D., Jessup, R. K., Bossaerts, P., & O'Doherty, J. P. (2013). Evidence for model-based computations in the human amygdala during Pavlovian conditioning. *PLoS Computational Biology, 9*, e1002918. doi:10.1371/journal.pcbi.1002918

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prosky (Eds.), *Classical conditioning II: Current research and theory* (pp. 64-99). New York, NY: Appleton-Century-Crofts.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461-464. doi:10.1214/aos/1176344136

Spector, I. P., Carey, M. P., & Steinberg, L. (1996). The sexual desire inventory: Development, factor structure, and evidence of reliability. *Journal of Sex & Marital Therapy, 22*, 175-190. doi:10.1080/00926239608414655

Van Duuren, M., Kendell-Scott, L., & Stark, N. (2003). Early aesthetic choices: Infant preferences for attractive premature infant faces. *International Journal of Behavioral Development, 27*, 212-219. doi:10.1080/01650250244000218

Zhang, S., Mano, H., Ganesh, G., Robbins, T., & Seymour, B. (2016). Dissociable learning processes underlie human pain conditioning. *Current Biology, 26*, 52-58. doi:10.1016/j.cub.2015.10.066