Learning biases to angry and happy faces during Pavlovian aversive conditioning

Stussi, Yoann; Pourtois, Gilles; Olsson, Andreas; Sander, David

How to cite

STUSSI, Yoann et al. Learning biases to angry and happy faces during Pavlovian aversive conditioning. In: Emotion, 2021, vol. 21, n° 4, p. 742–756. doi: 10.1037/emo0000733

This publication URL: https://archive-ouverte.unige.ch/unige:134035
Publication DOI: 10.1037/emo0000733

In press, *Emotion*

LEARNING BIASES TO ANGRY AND HAPPY FACES DURING PAVLOVIAN AVERSIVE CONDITIONING

Yoann Stussi

University of Geneva

Gilles Pourtois

Ghent University

Andreas Olsson

Karolinska Institutet

David Sander

University of Geneva

Corresponding author:

Yoann Stussi, Department of Psychology, Harvard University,

Northwest Lab Building, 52 Oxford Street, Cambridge, MA 02138.

E-mail: ystussi@fas.harvard.edu

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 (FPSE), University of Geneva; Gilles Pourtois, Cognitive & Affective Psychophysiology Laboratory, Department of Experimental Clinical & Health Psychology, Ghent University; Andreas Olsson, Department of Clinical Neuroscience, Division of Psychology, Karolinska Institutet; 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 (FPSE), University of Geneva.

Yoann Stussi is now at Department of Psychology, Harvard University, Cambridge.

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 (P0GEP1_159057) and an Early Postdoc.Mobility fellowship (P2GEP1_187911) from the Swiss National Science Foundation to Y.S. We thank Chloé Da Silva Coelho for her help with data collection, as well as Sylvain Delplanque and Eva R. Pool for their insightful comments on this work. The data reported in the present study and the code used for data analysis are available on the Open Science Framework (https://doi.org/10.17605/OSF.IO/DK2NP).

Correspondence concerning this article should be addressed to Yoann Stussi, Department of Psychology, Harvard University, Northwest Lab Building, 52 Oxford Street, Cambridge, MA 02138. Email: ystussi@fas.harvard.edu.

Abstract

Learning biases in Pavlovian aversive conditioning have been found in response to specific categories of threat-relevant stimuli, such as snakes or angry faces. This has been suggested to reflect a selective predisposition to preferentially learn to associate stimuli that provided threats to survival across evolution with aversive outcomes. Here, we contrast with this perspective by highlighting that both threatening (angry faces) and rewarding (happy faces) social stimuli can produce learning biases during Pavlovian aversive conditioning. Using a differential aversive conditioning paradigm, the present study (N = 107) showed that the conditioned response to angry and happy faces was more readily acquired and more resistant to extinction than the conditioned response to neutral faces. Strikingly, whereas the effects for angry faces were of moderate size, the conditioned response persistence to happy faces was of relatively small size and influenced by inter-individual differences in their affective evaluation, as indexed by a Go/No-Go Association Task. Computational reinforcement learning analyses further suggested that angry faces were associated with a lower inhibitory learning rate than happy faces, thereby inducing a greater decrease in the impact of negative prediction error signals that contributed to weakening extinction learning. Altogether, these findings provide further evidence that the occurrence of learning biases in Pavlovian aversive conditioning is not specific to threat-related stimuli and depends on the stimulus' affective relevance to the organism.

Keywords: Pavlovian conditioning; Learning; Emotion; Happy faces; Angry faces

Introduction

Learning to predict and anticipate impending threats in the environment holds a critical survival value to organisms (e.g., LeDoux & Daw, 2018). A basic form of learning whereby this skill is achieved is Pavlovian aversive conditioning (e.g., Delgado, Olsson, & Phelps, 2006; LaBar, Gatenby, Gore, LeDoux, & Phelps, 1998; Phelps & LeDoux, 2005). In this procedure, organisms learn to associate a stimulus from the environment (the conditioned stimulus) with a biologically aversive outcome (the unconditioned stimulus) through single or repeated contingent pairing (Pavlov, 1927; Rescorla, 1988), thereby endowing the conditioned stimulus with a predictive and emotional value eliciting an anticipatory response (the conditioned response). Research on Pavlovian conditioning has generally focused on identifying principles that apply across different types of stimuli irrespective of their nature (Pavlov, 1927; Pearce & Hall, 1980; Rescorla & Wagner, 1972). Certain associations have, however, been revealed to be more easily formed and maintained than others (Garcia & Koelling, 1966; Öhman & Mineka, 2001; Seligman, 1970, 1971). Surprisingly, mechanisms underlying such learning biases remain yet not well elucidated.

Major theoretical models put forward, such as the preparedness (Seligman, 1970, 1971) and fear module (Öhman & Mineka, 2001) theories, adopt an evolutionary perspective according to which organisms are biologically predisposed by evolution to preferentially associate stimuli that provided threats to the species' survival with aversive events. In agreement with this view, learning biases have been found in response to stimuli from specific animal and social threat-relevant categories, such as snakes, angry faces, or outgroup faces, in that these stimuli are more readily and persistently associated with an aversive outcome than nonthreatening stimuli, such as birds, happy faces, or ingroup faces (e.g., Ho & Lipp, 2014; Öhman & Dimberg, 1978; Öhman, Eriksson, & Olofsson, 1975; Öhman,

Fredrikson, Hugdahl, & Rimmö, 1976; Öhman & Mineka, 2001; Olsson, Ebert, Banaji, & Phelps, 2005; but see Åhs et al., 2018; Davey, 1995; Mallan, Lipp, & Cochrane, 2013).

An alternative framework to these accounts derives from appraisal theories of emotion (e.g., Sander, Grafman, & Zalla, 2003; Sander, Grandjean, & Scherer, 2005, 2018), and proposes that the occurrence of learning biases in Pavlovian aversive learning is driven by a mechanism of relevance detection that is not selective to threat (Stussi, Brosch, & Sander, 2015; Stussi, Pourtois, & Sander, 2018; Stussi, Ferrero, Pourtois, & Sander, 2019). This model holds that stimuli detected as relevant to the individual's concerns–such as their goals, needs, or values (Frijda, 1986; Pool, Brosch, Delplanque, & Sander, 2016)–benefit from enhanced Pavlovian conditioning beyond stimulus valence and evolutionary status per se, and that such preferential learning is critically dependent on individual differences in stimulus affective evaluation. Congruent with this hypothesis, initial evidence (Stussi et al., 2018) has shown that, similar to threat-relevant stimuli (angry faces or snakes), positive stimuli with high biological relevance to the organism (baby faces or erotic stimuli) can likewise induce learning biases during Pavlovian aversive conditioning.

Here, we sought to gain further insights into the mechanisms that modulate emotional learning in humans by comparing these two competing models through the investigation of Pavlovian aversive conditioning to threatening (angry faces), rewarding (happy faces), and neutral (neutral faces) social stimuli. On the one hand, extant evidence has documented the existence of learning biases to angry but not to happy faces in Pavlovian aversive conditioning (see, e.g., Bramwell, Mallan, & Lipp, 2014; Dimberg & Öhman, 1996; Esteves, Parra, Dimberg, & Öhman, 1994; Mazurski, Bond, Siddle, & Lovibond, 1996; Öhman & Dimberg, 1978; Öhman & Mineka, 2001; Rowles, Mallan, & Lipp, 2012), thereby mostly supporting the predictions of the preparedness and fear module theories[1]. On the other hand,

[1] Of note, Bramwell et al. (2014) reported resistance to extinction to outgroup race happy faces, thereby indicating that happy faces may lead to preferential aversive learning under certain circumstances. This effect was not due to negative

the relevance detection model predicts that both angry and happy faces should be preferentially learned during Pavlovian conditioning relative to neutral faces because of their higher affective relevance, but that learning biases to happy faces should be smaller than to angry faces and more sensitive to inter-individual differences in their affective evaluation. Indeed, happy faces have been suggested to generally have a lower level of relevance to the organism than stimuli with heightened biological relevance, such as angry or baby faces (Brosch, Pourtois, & Sander, 2010; Brosch, Sander, Pourtois, & Scherer, 2008; Pool et al., 2016). Whereas the latter stimuli are likely to be consistently detected as highly relevant across individuals due to their importance for the organism and species' survival, happy faces can carry several meanings (Ambadar, Cohn, & Reed, 2009; Martin, Rychlowska, Wood, & Niedenthal, 2017) and their processing may vary as a function of the situation and individual differences, such as extraversion for instance (Canli, Sivers, Whitfield, Gotlib, & Gabrieli, 2002). Nonetheless, prior research has mainly used small sample sizes (typical *n* by group ranged between 15 and 25), hence undermining the possibility to detect potentially small learning biases and explore whether learning biases to happy faces can be mapped onto inter-individual differences.

In the present study, we therefore implemented a differential Pavlovian aversive conditioning paradigm in a relatively large sample size ($N = 107$) to test the predictions of the relevance detection model. Two angry, happy, and neutral faces were used as conditioned stimuli (CSs). One stimulus (CS+) from each CS category was systematically associated with a mild electric stimulation, whereas the other stimulus (CS-) was never paired with the stimulation. We operationalized the conditioned response (CR) as the differential skin conductance response (SCR) to the CS+ minus CS- from the same CS category, which served

evaluation of outgroup happy faces, which were evaluated as more pleasant than ingroup happy faces at the explicit level, whereas no difference in positive or negative evaluation was found between them at the implicit level. Nevertheless, no resistance to extinction was observed to ingroup happy faces, which suggests that the enhanced persistence of threat conditioned to outgroup happy faces was likely driven by the faces' race category.

as an index of learning (e.g., Olsson et al., 2005; Stussi et al., 2015, 2018, 2019). We also used computational modeling (Li, Schiller, Schoenbaum, Phelps, & Daw, 2011; Lindström, Golkar, & Olsson, 2015; Rescorla & Wagner, 1972; Stussi et al., 2018) to characterize the learning biases associated with angry and happy faces as opposed to neutral faces by extracting and comparing learning parameters for these CS categories. Additionally, we examined inter-individual differences in affective evaluation of happy faces in two ways. First, we considered participants' extraversion (see Canli et al., 2002) based on the rationale that individuals high in extraversion should tend to appraise happy faces as more relevant to their concerns than individuals lower in this trait (Sander et al., 2003, 2005). Second, we assessed implicit associations between the face categories and importance (e.g., Critcher & Ferguson, 2016) through a Go/No-go Association Task (GNAT; Nosek & Banaji, 2001). This task aimed at measuring the strength with which participants associated the face categories with the attribute of importance, thereby serving as a proxy of individuals' affective relevance evaluation of the faces. Specifically, we reasoned that the more individuals appraised the faces as affectively relevant, the more easily and rapidly they should associate these faces with importance (vs. unimportance).

As learning biases are generally reflected by a faster acquisition of a CR and/or an enhanced resistance to extinction of that CR (e.g., Öhman & Mineka, 2001), we predicted that (a) the CR to angry faces would be more readily acquired and more resistant to extinction than the CR to both happy faces and neutral faces across participants, whereas (b) the CR to happy faces would be acquired more readily and more resistant to extinction than the CR to neutral faces. Moreover, we hypothesized that (c) participants' extraversion level, as well as the sensitivity and rapidity with which they associated happy faces with the attribute of importance versus unimportance, would predict the CR acquisition readiness and persistence to happy faces.

Method

Participants

One hundred and seventeen students from the University of Geneva participated in the experiment, which was approved by the Faculty of Psychology and Educational Sciences ethics committee at the University of Geneva. They provided informed consent and received partial course credit for their participation. Ten participants were excluded from the analyses because of technical problems (n = 2), for displaying virtually no SCR (n = 2), for failing to acquire a CR to at least one of the CSs+ (n = 5), or for withdrawing from the study early (n = 1). These exclusion criteria were determined prior to data collection (see Olsson et al., 2005; Olsson & Phelps, 2004; Stussi et al., 2015, 2018, 2019). The final sample size consisted of 107 participants (85 women, 22 men), aged between 19 and 34 years old (mean age = 21.85 ± 2.57 years). Two participants were further excluded from the computational modeling analyses because their individual parameters could not be estimated due to a lack of SCR to all the angry face CSs during the experiment (see supplemental materials). The sample size was established before data collection on the basis of the current heuristic suggesting a sample of at least 100 participants for studies considering inter-individual differences (see, e.g., Dubois & Adolphs, 2016). For counterbalancing purposes, we aimed to recruit a minimum sample size of 104 participants exhibiting differential conditioning to at least one of three CS categories. We stopped collecting data at the end of the academic year and ascertained that the established sample size had been reached. A sensitivity power analysis performed with G*Power 3 (Faul, Erdfelder, Lang, & Buchner, 2007) indicated that this sample size allowed for detecting a smallest population effect size of d_z = 0.242 with a power of 80% using a one-tailed paired-sample t test.

Apparatus and stimuli

The experiment took place in a sound-attenuated experimental chamber. The stimuli were presented using MATLAB (The MathWorks Inc., Natick, MA) with the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997) and displayed on a 23-inch LED monitor. Eight angry, eight happy, and eight neutral male face stimuli from the Karolinska Directed Emotional Faces (KDEF; Lundqvist, Flykt, & Öhman, 1998) were used either as targets or as distractors in the GNAT (see supplemental materials). Four word stimuli related to the attribute of importance (i.e., important words; "important", "relevant", "significant", "impactful") and four word stimuli related to the attribute of unimportance (i.e., unimportant words; "unimportant", "irrelevant", "insignificant", "secondary") were also used both as targets and distractors.

In the differential Pavlovian aversive conditioning procedure, the CSs consisted of two male angry (model numbers AM10ANS, AM29ANS), two male happy (AM07HAS, AM22HAS), and two male neutral (AM11NES, AM31NES) faces taken from the KDEF (Lundqvist et al., 1998). These faces were selected based on the correct identification (hit rate range: 89.06%-100%) and intensity ratings (mean intensity range: 5.73-7.63) of their respective emotional expression (Goeleven, De Raedt, Leyman, & Verschuere, 2008). Each face served both as a CS+ and as a CS-, counterbalanced across participants. Subjective ratings performed before the conditioning procedure (see supplemental materials) on a visual analog scale from 0 (*very unpleasant*) to 100 (*very pleasant*) indicated that the angry faces were evaluated as unpleasant (M = 15.29, SD = 15.76), the happy faces as pleasant (M = 68.28, SD = 20.39), and the neutral faces as relatively neutral (M = 43.47, SD = 13.07). The unconditioned stimulus (US) was a mild electric stimulation (200-ms duration) delivered to the participants' right wrist through a unipolar pulse electric stimulator (STM200; BIOPAC Systems Inc., Goleta, CA). 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' left hand. SCR was continuously recorded during the conditioning procedure with a sampling rate of 1000 Hz by means of a BIOPAC MP150 system (Santa Barbara, CA). The SCR data were analyzed offline with AcqKnowledge software (version 4.4; BIOPAC Sytems Inc., Goleta, CA).

Procedure

Between two to eight months prior to their participation in the study, participants completed the French version of the NEO Five-Factor Inventory (NEO-FFI; Costa & McCrae, 1992; Rolland, Parker, & Strumpf, 1998). Upon arrival at the laboratory, they were informed about the general layout of the experiment, provided written informed consent, and performed the GNAT. Participants were next asked to evaluate the to-be-CSs according to various dimensions (see supplemental materials) before undergoing the differential Pavlovian aversive conditioning procedure. Finally, they were asked again to provide subjective ratings of the CSs after conditioning (see supplemental materials) and were debriefed.

Differential Pavlovian aversive conditioning. Prior to conditioning, the electrodes for measuring SCR and delivering the electric stimulation were attached to participants. A work-up procedure was then performed to individually calibrate the electric stimulation intensity (M = 34.55 V, SD = 7.57, range = 20-50 V) to a level reported as "uncomfortable, but not painful". The differential Pavlovian aversive conditioning procedure (see Figure 1a,b) comprised three contiguous phases. In the initial habituation phase, the six CSs were each presented twice without being reinforced. During the subsequent acquisition phase, each CS was presented seven times. This phase always started with a reinforced CS+ trial. Each CS+ was paired with the US with a partial reinforcement schedule, five of the seven CS+ presentations co-terminating with the US delivery, whereas the CS- from each CS category was never associated with the US. The use of a partial reinforcement schedule aimed to

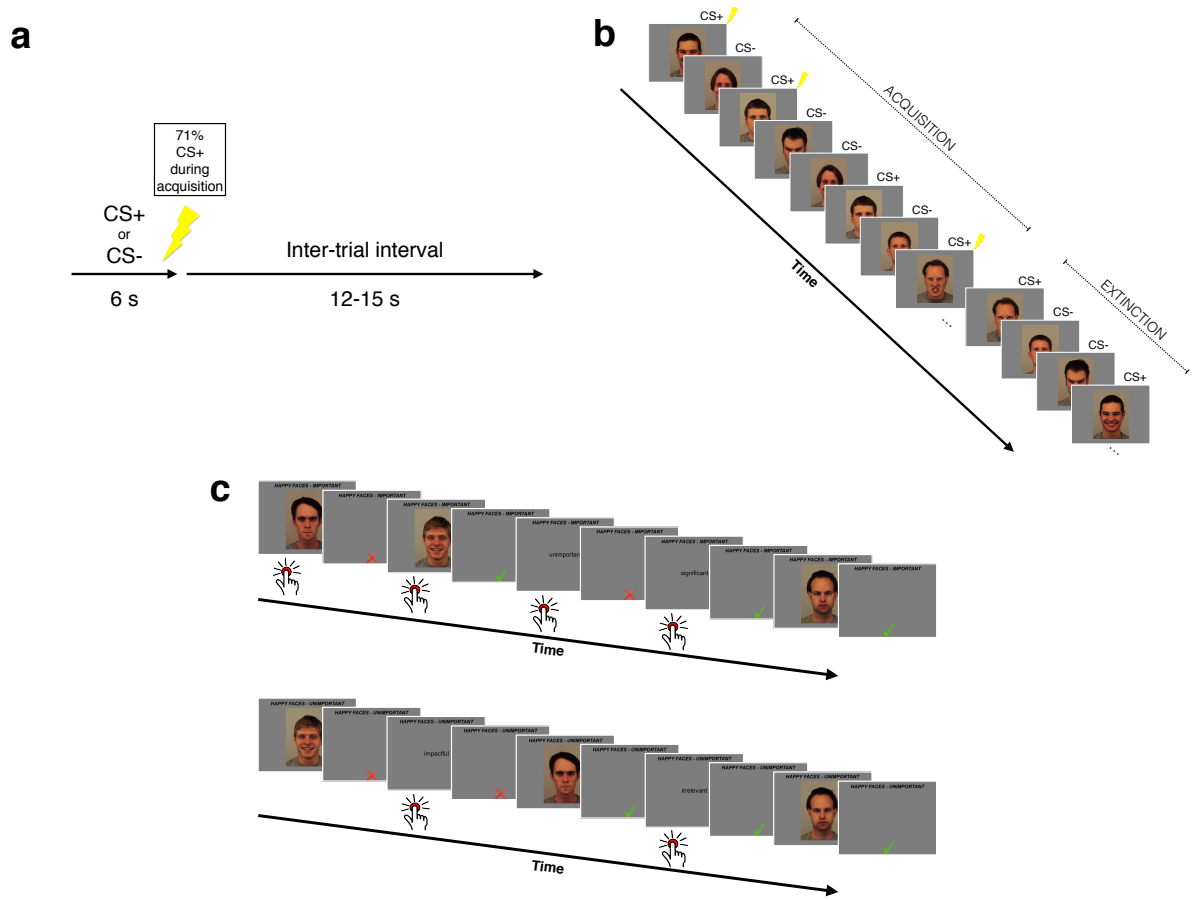

Figure 1. Schematic representation of the experimental procedures. (a) Within-trial structure during the differential Pavlovian aversive conditioning procedure: two angry, happy, and neutral faces were presented as conditioned stimuli (CSs) in a pseudorandom order for 6 s during three contiguous phases (habituation, acquisition, extinction). Five of the seven CS+ trials (71%) for each face category co-terminated with an electric stimulation during acquisition. Trials were separated by an inter-trial interval ranging from 12 to 15 s. (b) Illustration of the overall differential Pavlovian aversive conditioning structure during acquisition and extinction. Acquisition consisted of presentations of the six CSs on a partial reinforcement schedule, whereas extinction consisted of presentations of the same CSs while the electric stimulation was no longer delivered. (c) Illustration of the Go/No-go Association Task: examples of five trials in which participants had to detect whether the faces and the words belonged to the target categories "Happy faces" or "Important words" (upper panel), or to the target categories "Happy faces" or "Unimportant words" (lower panel). If the face or word belonged to one of the two target categories, the correct response was to press 'A' on the keyboard, but to withdraw from responding otherwise. After each response, participants received feedback consisting of either a green check or a red cross for correct and incorrect responses, respectively. The different faces shown (AM02NES, AM07HAS, AM10ANS, AM11NES, AM22HAS, AM23HAS, AM24ANS, AM29ANS, AM31NES) were taken with permission from the Karolinska Directed Emotional Faces database (Lundqvist et al., 1998), which allows their free use for scientific publication (see kdef.se).

potentiate the CR resistance to extinction, hence optimizing the examination of differences 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 inter-trial interval varying from 12 to 15 s. The CSs' presentation order was pseudorandomized into eight different orders to counterbalance the associations between the face stimuli and CS type (CS+ vs. CS-) across the three CS categories (angry vs. happy vs. neutral).

NEO Five-Factor Inventory (NEO-FFI). The NEO-FFI is a standard personality inventory measuring the Big Five personality traits consisting of neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness (Costa & McCrae, 1992). It comprises 60 items (12 per trait), each of which is measured on a 5-point Likert scale ranging from 0 (*strongly disagree*) to 4 (*strongly agree*). Given our a priori hypotheses, we focused here on extraversion ($M = 28.23$, $SD = 5.69$, range = 10-40, Cronbach's $\alpha = .76$; see Figure S1 in the supplemental materials). Exploratory analyses including the other personality traits are reported in the supplemental materials.

Go/No-go Association Task. In the GNAT, participants were presented with faces from three emotional categories (angry vs. happy vs. neutral) and words from two categories (important vs. unimportant). In each trial, a face or a word was displayed at the center of the screen. Participants were instructed to press as quickly and accurately as possible on the "A" key if the stimulus was a member of a target category (go trials), but to withdraw from responding otherwise (no-go trials). Throughout the task, the labels of the target categories were continuously displayed at the top of the screen as a reminder. After each trial, feedback about participants' response was displayed at the bottom of the screen (i.e., a green check for correct or a red cross for incorrect) during a 150-ms inter-trial interval (see Figure 1c).

The GNAT began with a practice session of five blocks in which there was only a single target category (see supplemental materials). The experimental session ensued and was composed of three parts, each divided into two blocks. Within each part, a specific face

category was one of the two target categories with “important” words being the other target category in block 1, and “unimportant” words the other target category in block 2. The order of the three parts as a function of the face categories was counterbalanced between participants. Each block consisted of 96 trials: 16 training trials and 80 critical trials. Four faces from the target face category and two faces from each distractor face category were presented intermixed with the four “important” and the four “unimportant” words in a pseudorandom order. The response deadline was idiosyncratically adapted to the participants’ reaction times and response accuracy (see, e.g., Coppin et al., 2016; Nosek & Banaji, 2001): When response was correct (for both go and no-go trials) and reaction time faster than the arbitrary response deadline (for go trials), the response deadline for the next trial was set as 500 ms or as 666 ms if reaction time was slower than 500 ms but faster than 666 ms (for go trials); otherwise, it was set as 800 ms.

Participants’ reaction times and response accuracy were recorded for each trial. All trials with reaction times faster than 100 ms were excluded from analysis. Data for all errors and distracter items were removed from the reaction times analysis. According to signal detection theory, we calculated a d’ score for each block within each part of the GNAT experimental session, considering only critical trials (Nosek & Banaji, 2001). We converted the proportions of hits (correct go-responses to targets) and false alarms (incorrect go-responses to distractors) to z scores before computing the difference between them, thereby obtaining d’. Hit and false-alarm rates equal to 0 or 1 were replaced with 1/(2N) and 1 – 1/(2N), respectively, where N is the number of trials (Macmillan & Creelman, 2005). A differential d’ index was then calculated by subtracting the d’ scores of the second block (target face category + unimportant words) from those of the first block (target face category + important words; see, e.g., Coppin et al., 2016). Higher values on this index indicated higher accuracy when faces from the target face category and “important” words were targets

in comparison with when faces from the target face category and "unimportant" words were targets. Additionally, we computed a differential index for reaction times by subtracting the mean reaction times of the first block to those of the second block, higher values thus reflecting faster responses when faces from the target face category and "important" words were targets relative to when faces from the target face category and "unimportant" words were targets. The differential d' and reaction times indices served as indicators of the strength of association between the faces categories and the attribute of importance versus that of unimportance (Nosek & Banaji, 2001). Although d' scores are usually used as the main dependent variable in the GNAT, we measured both indicators because reaction times have been suggested to be more reliable than d' scores due to their measurement on a continuous (vs. dichotomic) scale at the trial level (Nosek & Banaji, 2001).

Response definition

SCR was scored for each trial as the peak-to-peak amplitude difference in skin conductance of the largest response starting in the 0.5-4.5 s temporal window following CS onset. The minimal response criterion was 0.02 μS, and responses below this criterion were scored as zero and remained in the analysis. A low-pass filter (Blackman -92 dB, 1 Hz) was applied on the SCR data before analysis. SCRs were detected automatically with AcqKnowledge software and manually checked for artifacts and response detection. Trials containing artifacts affecting the scoring of event-related SCRs (0.17%) were removed from the subsequent analyses. The raw SCRs were scaled according to each participant's mean unconditioned response (UR), and square-root-transformed to normalize the distributions. The UR was scored as the peak-to-peak amplitude difference in skin conductance of the largest response starting in the 0.5-4.5 s temporal window after the US delivery, and the mean UR was calculated across all USs for each participant. The habituation means comprised the first two presentations of each CS (i.e., Trials 1 and 2). In order to tease apart

effects of faster conditioning from those of larger conditioning, the acquisition means were split into an early (i.e., the first three presentations of each CS following the first pairing between the CS+ of a given CS category and the US; Trials 4 to 6) and a late (i.e., the following three presentations of each CS; Trials 7 to 9) phase (see, e.g., Lonsdorf et al., 2017; Olsson, Carmona, Downey, Bolger, & Ochsner, 2013; Stussi et al., 2015, 2018, 2019). This allowed us to specifically examine the CR acquisition readiness during early acquisition. The first acquisition trial for each CS was removed from the CR analysis because the CSs+ became predictive of the US only after their first association therewith. The extinction means encompassed the last six presentations of each CS (i.e., Trials 10 to 15). The conditioning data analyses were performed on the CR, which was calculated as the SCR to the CS+ minus the SCR to the CS- from the same CS category (e.g., Olsson et al., 2005; Stussi et al., 2015, 2018, 2019). This procedure allows for reducing preexisting differences in emotional salience between the different CS categories (Olsson et al., 2005).

Computational modeling

Based on previous research (Stussi et al., 2018), we constructed a simple reinforcement learning model to characterize Pavlovian aversive conditioning to angry, happy, and neutral faces (for further details, see supplemental materials). We adapted the standard version of the Rescorla-Wagner model (Rescorla & Wagner, 1972) by implementing distinct learning rates for positive (i.e., when the outcome is not predicted or more than expected; excitatory learning) and negative (i.e., when the outcome is omitted or less than expected; inhibitory learning) prediction errors instead of a single learning rate (see Niv, Edlund, Dayan, & O'Doherty, 2012; Stussi et al., 2018). Excitatory and inhibitory learning rates exert an influence on associative learning by altering the impact of positive and negative prediction error signals, respectively, on the CS predictive value (see Niv & Schoenbaum, 2008). In the dual-learning-rate Rescorla-Wagner model, the predictive value (or associative strength) V of

a given CS *j* is updated based on the sum of the current predictive value V_j at trial *t*, and the prediction error between the predictive value V_j and the outcome *R* at trial *t*, weighted by different learning rates for positive and negative prediction errors 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 errors α^+ and the learning rate for negative prediction errors α^- are free parameters within the range [0, 1]. If the US was delivered on the current trial *t*, $R(t) = 1$, else $R(t) = 0$. This model allows for parsimoniously accounting for how specific stimulus categories can accelerate acquisition (through the excitatory learning rate) and enhance resistance to extinction (through the inhibitory learning rate) of the CR.

The learning-rate parameters were estimated, and the trial-by-trial CS values calculated, by fitting the model to the individual normalized (i.e., scaled and square-root-transformed) SCR data separately for each CS category. Model comparison indicated that the dual-learning-rate Rescorla-Wagner model provided the best fit to the SCR data relative to alternative models (see supplemental materials). Accordingly, we compared the estimated excitatory and inhibitory learning-rate parameters across the three different CS categories used (angry vs. happy vs. neutral).

Statistical analyses

The differential d' and the differential reaction time indices derived from the GNAT were each analyzed with a one-way repeated-measures analysis of variance (ANOVA) with face category (angry vs. happy vs. neutral) as a within-participant factor. Statistically significant main effects were followed up with a multiple comparison procedure using Tukey's HSD tests when applicable.

Following standard practice in the human conditioning literature (e.g., Lonsdorf et al., 2017; Olsson et al., 2005; Stussi et al., 2015, 2018, 2019), the SCR data was analyzed

separately for each conditioning phase. The habituation and extinction phases and the estimated learning rates were each analyzed with a one-way repeated-measures ANOVA with CS category (angry vs. happy vs. neutral) as a within-participant factor. The acquisition phase was analyzed with a two-way repeated-measures ANOVA with CS category (angry vs. happy vs. neutral) and time (early vs. late) as within-participant factors. One-sample *t* tests were additionally performed to test whether differential conditioning occurred for the CS categories across the entire acquisition phase. To specifically test our a priori hypotheses, we conducted planned contrast analyses comparing the CR during early acquisition and during extinction, as well as the estimated learning rates, to (a) angry versus neutral faces, (b) happy versus neutral faces, and (c) angry versus happy faces. As these contrasts were nonorthogonal, we applied a Holm-Bonferroni sequential procedure (Holm, 1979) to correct for multiple comparisons. The alpha level of the contrast with the lowest *p* value was set as $\alpha = .05/3 = .0167$, the alpha level with the second lowest *p* value as $\alpha = .05/2 = .025$, and the alpha level with the highest *p* value as $\alpha = .05$. For each planned contrast, we also calculated the Bayes factor (BF_{10}) quantifying the likelihood of the data under the alternative hypothesis compared with the likelihood of the data under the null hypothesis (e.g., Dienes, 2011; Rouder, Speckman, Sun, Morey, & Iverson, 2009). Because we expected moderate effects for angry faces and relatively small effects for happy faces, we used a noninformative Cauchy prior distribution with a width of 0.5 for the comparisons between angry and happy faces and between angry and neutral faces (see Stussi et al., 2018), and of 0.25 for the comparison between happy and neutral faces. When our theory-driven hypotheses clearly predicted the direction of the expected effects, we performed one-sided testing to test them (one-sample *t* tests, contrasts a, b, and c).

To assess our a priori hypotheses that extraversion, as well as the sensitivity and the rapidity with which happy faces were associated with the attribute of importance predicted

the CR acquisition readiness and persistence to these faces, we conducted multiple linear regression analyses. These analyses tested whether the CR acquisition readiness (i.e., during early acquisition) and persistence (i.e., during extinction), along with the excitatory and inhibitory learning-rate estimates, to happy faces were predicted by participants' (a) extraversion level, (b) differential d' index for happy faces, and (c) differential reaction time index for happy faces. Further exploratory multiple linear regression analyses carried out on the CR and the learning rates to angry and neutral faces to investigate the specificity of these predictive effects are reported in the supplemental materials.

All statistical analyses were performed with RStudio (RStudio Team, 2016). Huynh-Feldt adjustments of degrees of freedom were applied for repeated-measures ANOVAs when appropriate. Partial eta squared (η^2) or Hedges' g_{av} (or g_z) and their 90% or 95% confidence interval (CI) were used as estimates of effect sizes (see Lakens, 2013) for the repeated-measures ANOVAs and the planned contrasts analyses (or one-sample *t* tests), respectively, whereas the coefficient of determination R^2 along with its 90% CI was used for multiple linear regressions.

Results

Pavlovian aversive conditioning

Figure 2 depicts the mean SCR to angry, happy, and neutral faces across the habituation, acquisition, and extinction phases of the differential Pavlovian aversive conditioning separately for the CS+ and the CS-. In the habituation phase, no preexisting difference in differential SCR across the CS categories (angry vs. happy vs. neutral) was found, $F(2, 212) = 0.003$, $p = .997$, partial $\eta^2 = .00003$, 90% CI [.000, .0006].

Analysis of the acquisition phase revealed successful differential conditioning to all three CS categories, as reflected by larger SCRs to the CS+ than to the CS- for angry, $t(106) = 7.44$, $p < .001$ (one-tailed), $g_z = 0.714$, 95% CI [0.505, 0.931], happy, $t(106) = 8.10$, $p <$

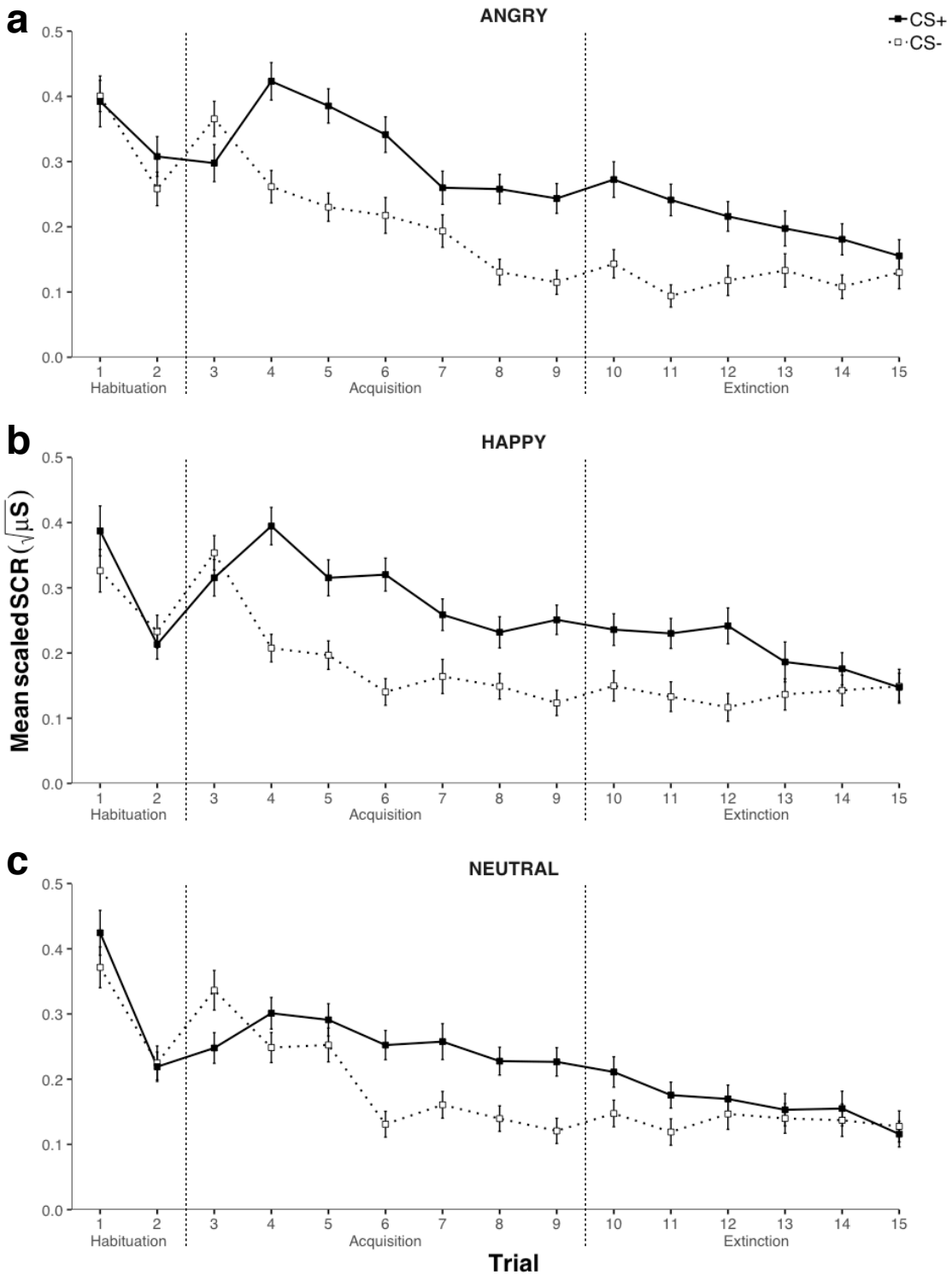

Figure 2. Mean scaled skin conductance response (SCR) to the conditioned stimuli as a function of the conditioned stimulus type (CS+ vs. CS-) across trials. Mean scaled SCR to (a) angry faces, (b) happy faces, and (c) neutral faces. Error bars indicate ± 1 *SEM* adjusted for within-participant designs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

.001 (one-tailed), $g_z = 0.777$, 95% CI [0.564, 0.998], and neutral faces, $t(106) = 5.97$, $p < .001$ (one-tailed), $g_z = 0.573$, 95% CI [0.372, 0.781]. The CS categories however differentially

influenced the CR acquisition as indicated by a main effect of CS category, $F(2, 212) = 3.27$, $p = .040$, partial $\eta^2 = .030$, 90% CI [.001, .071]. The interaction effect between CS category and time did not yield statistical significance, $F(2, 212) = 2.60$, $p = .076$, partial $\eta^2 = .024$, 90% CI [.000, .062]. Congruent with our a priori hypothesis, a planned contrast analysis showed that the CR to angry faces was more readily acquired than the CR to neutral faces during early acquisition, $t(106) = 2.60$, $p = .005$ (one-tailed), $g_{av} = 0.358$, 95% CI [0.084, 0.636], $BF_{10} = 6.642$ (see Figure 3). Importantly, the CR to happy faces was likewise more readily acquired than to neutral faces, $t(106) = 3.25$, $p < .001$ (one-tailed), $g_{av} = 0.442$, 95% CI [0.169, 0.720], $BF_{10} = 41.237$, whereas there was no statistical difference in CR acquisition readiness to angry faces compared with happy faces, $t(106) = -0.58$, $p = .717$ (one-tailed), $g_{av} = -0.073$, 95% CI [-0.324, 0.177], $BF_{10} = 0.101$ (see Figure 3). No statistical differences emerged between the three CS categories during late acquisition (all ps > .92, $0.02 < g_{av}$s < 0.05, all BFs$_{10} < 0.32$).

Critically, the CR persistence was also modulated by the CS categories during extinction, $F(2, 212) = 5.97$, $p = .003$, partial $\eta^2 = .053$, 90% CI [.011, .104]. As predicted, the CR to angry faces was more resistant to extinction than the CR to neutral faces, $t(106) = 3.69$, $p < .001$ (one-tailed), $g_{av} = 0.432$, 95% CI [0.196, 0.672], $BF_{10} = 133.200$. Similarly, the CR to happy faces was more persistent than to neutral faces, $t(106) = 2.01$, $p = .024$ (one-tailed), $g_{av} = 0.247$, 95% CI [0.003, 0.493], $BF_{10} = 2.777$ (see Figure 3). By comparison, we did not observe an enhanced CR persistence to angry faces relative to happy faces, $t(106) = 1.28$, $p = .102$ (one-tailed), $g_{av} = 0.133$, 95% CI [-0.072, 0.339], $BF_{10} = 0.573$.

Estimated learning rates

Analysis of the excitatory learning-rate estimates revealed no statistically significant main effect of CS category, $F(2, 208) = 2.50$, $p = .085$, partial $\eta^2 = .023$, 90% CI [.000, .061]. A more focused planned contrast analysis indicated that happy faces were associated with a

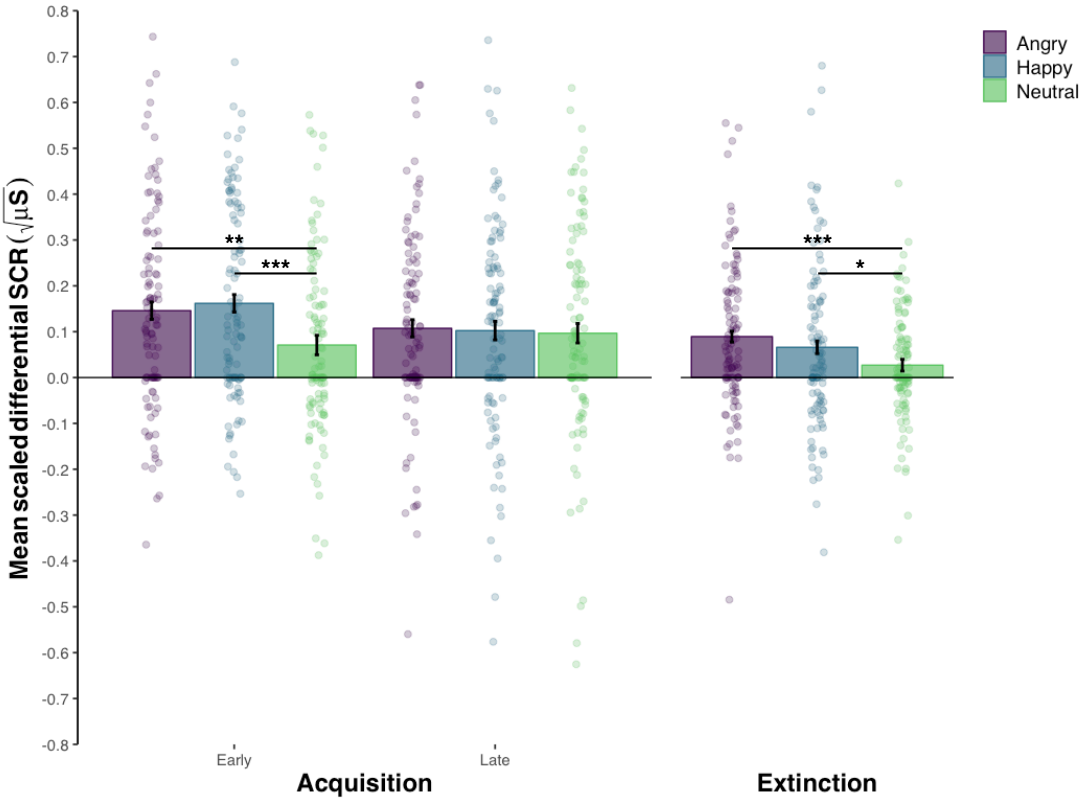

Figure 3. Mean conditioned response (scaled differential skin conductance response [SCR]) as a function of the conditioned stimulus category (angry vs. happy vs. neutral) during (early and late) acquisition and extinction. The dots indicated data for individual participants. Error bars indicated ± 1 *SEM* adjusted for within-participant designs (Morey, 2008). Asterisks indicate statistically significant differences between conditions (***$p < .001$, **$p < .01$, *$p < .05$, one-tailed, Holm-Bonferroni corrected).

higher excitatory learning rate than neutral faces, $t(104) = 2.05$, $p = .022$ (one-tailed), $g_{av} = 0.232$, 95% CI [0.007, 0.460], $BF_{10} = 2.986$ (see Figure 4a), but this difference was not statistically significant when correcting the alpha level for this contrast ($\alpha = .0167$). No statistical difference in excitatory learning rate was observed between angry and happy faces, $t(104) = -1.76$, $p = .959$ (one-tailed), $g_{av} = -0.205$, 95% CI [-0.438, 0.026], $BF_{10} = 0.058$, or between angry and neutral faces, $t(104) = 0.23$, $p = .410$ (one-tailed), $g_{av} = 0.027$, 95% CI [-0.203, 0.257], $BF_{10} = 0.181$. By contrast, the CS categories differentially affected the estimated inhibitory learning rates, $F(2, 208) = 5.95$, $p = .003$, partial $\eta^2 = .054$, 90% CI [.011, .106]. These estimates were lower for angry faces than for neutral faces, $t(104) = -3.52$, $p < .001$ (one-tailed), $g_{av} = -0.434$, 95% CI [-0.686, -0.186], $BF_{10} = 78.801$, and happy faces, $t(104) = -2.14$, $p = .017$ (one-tailed), $g_{av} = -0.242$, 95% CI [-0.468, -0.018], $BF_{10} = 2.477$,

whereas they were marginally lower for happy faces compared with neutral faces, $t(104) = -1.33$, $p = .093$ (one-tailed), $g_{av} = -0.164$, 95% CI [-0.409, 0.079], $BF_{10} = 1.015$ (see Figure 4b), although the latter difference did not yield statistical significance and the evidence for it remained inconclusive.

Go/No-go Association Task

The analysis of the differential d' index showed a statistically significant main effect of face category (angry vs. happy vs. neutral), $F(2, 212) = 15.46$, $p < .001$, partial $\eta^2 = .127$, 90% CI [.061, .193]. The differential d' index was higher for happy faces ($M = 0.15$, $SD = 0.55$) than for angry ($M = -0.20$, $SD = 0.46$; $p < .001$, $g_{av} = 0.683$, 95% CI [0.407, 0.965]) and neutral faces ($M = -0.10$, $SD = 0.44$; $p < .001$, $g_{av} = 0.493$, 95% CI [0.222, 0.769]), whereas there

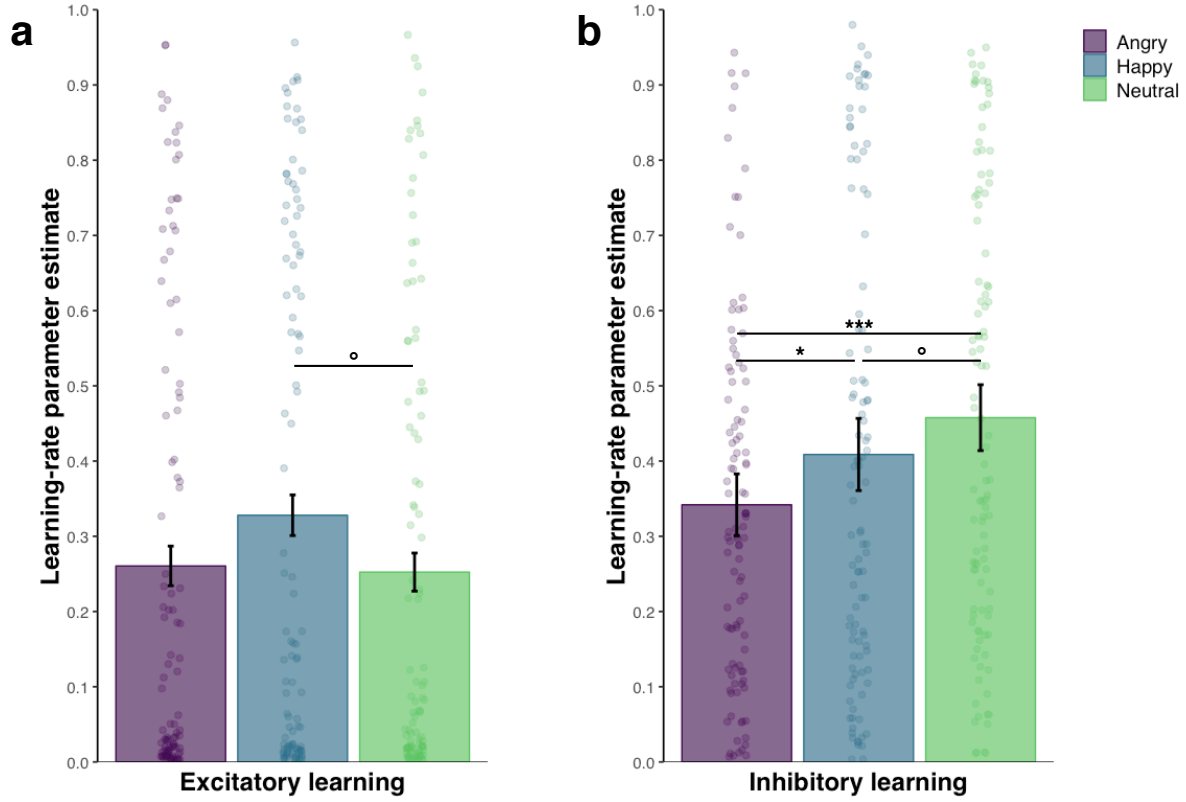

Figure 4. Learning-rate parameter estimates of the Rescorla-Wagner model implementing dual learning rates using the best-fitting parameters for positive predictions errors (excitatory learning) and negative prediction errors (inhibitory learning) as a function of the conditioned stimulus category (angry vs. happy vs. neutral). The dots indicate data for individual participants. Error bars indicate ± 1 *SEM* adjusted for within-participant designs (Morey, 2008). Asterisks indicate statistically significant differences between conditions (***$p < .001$, *$p < .05$, °$p < .10$, one-tailed, Holm-Bonferroni corrected).

was no statistical difference between angry and neutral faces (p = .273, g_{av} = 0.219, 95% CI [-0.030, 0.469]). These results suggest that participants exhibited a greater sensitivity to the association between the attribute of importance versus unimportance with happy faces than either angry or neutral faces. Conversely, the differential reaction time index did not differ statistically across the face categories, $F(2, 212) = 2.45$, p = .089, partial η^2 = .023, 90% CI [.000, .059].

Regression analyses

The multiple linear regression analyses on the CR to happy faces (see Table 1) showed that participants' extraversion level, differential d' index for happy faces, and differential reaction time index for happy faces did not predict the CR to happy faces during early acquisition (all ps > .34) where they only explained 1.51% of its variance (R^2 = .015, 90% CI [.000, .048], adjusted R^2 = -.014, $F(3, 103) = 0.53$, p = .664). However, these three predictors explained 13.06% of the variance of the CR to happy faces during extinction (R^2 = .131, 90% CI [.031, .224], adjusted R^2 = .105, $F(3, 103) = 5.16$, p = .002). Whereas extraversion and the differential d' index for happy faces did not predict the CR to happy faces (both ps > .38), the CR to happy faces was predicted by the differential reaction time index for these faces, b = 0.002, 95% CI [0.001, 0.003], β = .360, $t(103) = 3.83$, p < .001, reflecting that participants who were faster to associate happy faces with the attribute of importance than that of unimportance exhibited a larger CR to happy faces during extinction (see Figure 5). Regarding the excitatory and inhibitory learning rates (see Table 1), participants' extraversion level, differential d' index for happy faces, and differential reaction time index for happy faces explained 4.09% (R^2 = .041, 90% CI [.000, .100], adjusted R^2 = .012, $F(3, 101) = 1.44$, p = .236) and 4.71% (R^2 = .047, 90% CI [.000, .110], adjusted R^2 = .019, $F(3, 101) = 1.66$, p = .180) of their variance, respectively. No significant relationship emerged between the predictors and the excitatory and inhibitory learning-rate estimates (all ps > .05).

Table 1

Results for the multiple linear regression analyses

	Conditioned response to happy faces during early acquisition ($N = 107$)					Conditioned response to happy faces during extinction ($N = 107$)					Estimated excitatory learning rate to happy faces ($N = 105$)					Estimated inhibitory learning rate to happy faces ($N = 105$)				
	b	*SE*	*β*	*t* (103)	*p*	*b*	*SE*	*β*	*t* (103)	*p*	*b*	*SE*	*β*	*t* (101)	*p*	*b*	*SE*	*β*	*t* (101)	*p*
Intercept	0.073	0.106		0.69	.494	0.027	0.087		0.31	.759	0.069	0.169		0.41	.685	0.446	0.150		2.97	.004**
Extraversion	0.003	0.004	.085	0.87	.388	0.002	0.003	.046	0.50	.621	0.009	0.006	.146	1.49	.140	-0.002	0.005	-.031	-0.32	.750
Differential d' index	-0.005	0.039	-.013	-0.13	.896	-0.028	0.032	-.082	-0.87	.386	0.083	0.062	.133	1.33	.187	0.076	0.055	.137	1.37	.173
Differential reaction time index	-0.001	0.001	-.096	-0.96	.341	0.002	0.0005	.360***	3.83	< .001	-0.000	0.001	-.019	-0.19	.852	-0.002	0.001	-.195	-1.95	.054
R^2			.015					.131					.041					.047		

Note. ***$p < .001$, **$p < .01$.

For angry and neutral faces, no statistically significant relationship was observed between participants' extraversion level, differential d' index, and differential reaction time index, and the CR during early acquisition and extinction as well as the learning-rate estimates (all *p*s > .08; see supplemental materials).

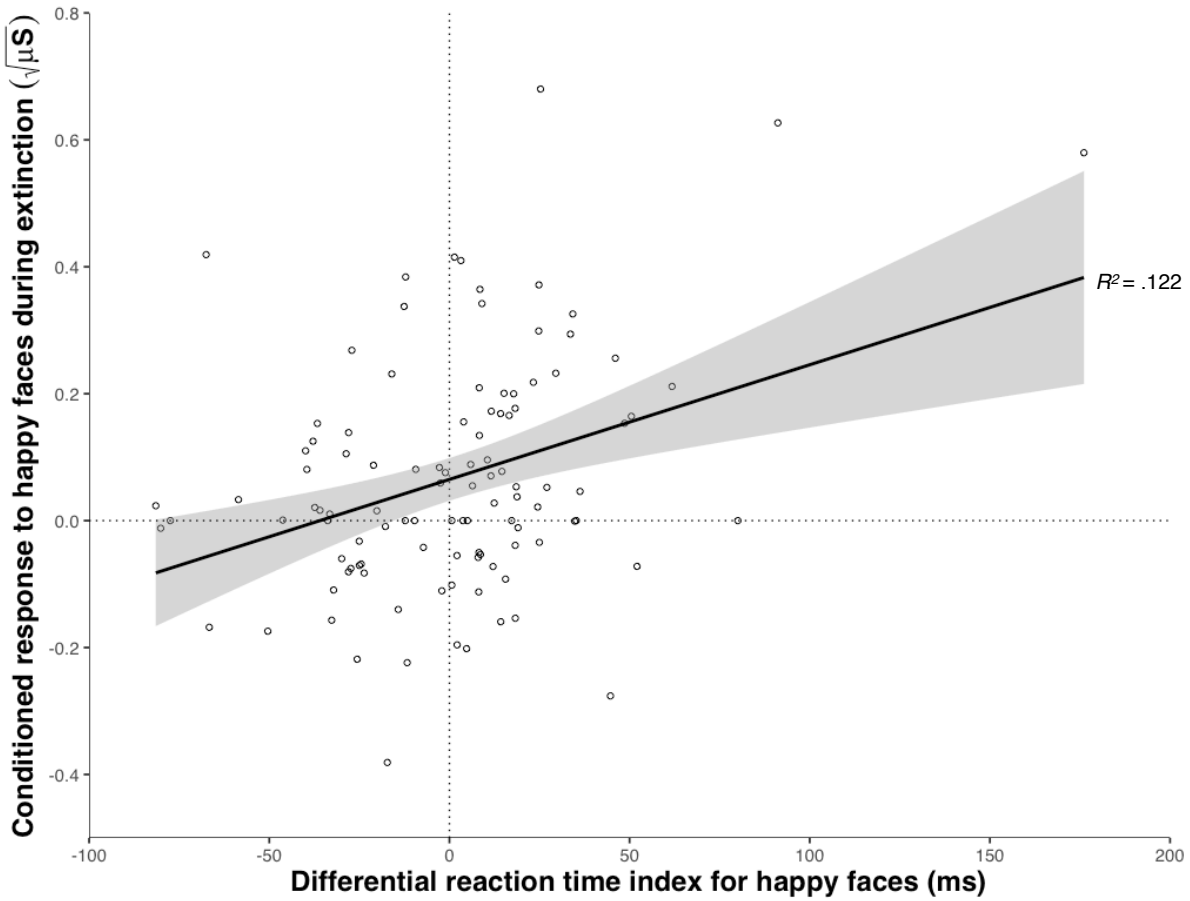

Figure 5. Relationship between the differential reaction time index for happy faces in the Go/No-go Association Task (mean reaction times in the block where happy faces and the attribute of importance were target categories minus mean reaction times in the block where happy faces and the attribute of unimportance were target categories) and the conditioned response to happy faces during extinction. The line represents the fitted regression line using least squares estimation and 95% confidence interval.

Discussion

In this study, we aimed to test the predictions of two competing theoretical approaches of emotional learning. More particularly, we tested the hypothesis deriving from appraisal

theories that enhanced emotional learning is driven by a relevance detection mechanism that is not specific to threat, and depends on individual differences in affective relevance appraisal. This hypothesis departs from the preparedness and fear module theories, according to which enhanced emotional learning is selective to threat. To that end, we compared Pavlovian aversive conditioning to threat-related (angry faces), positive (happy faces), and neutral (neutral faces) social stimuli and investigated the influence of inter-individual differences in affective evaluation on this process. Altogether, our results showed that both angry and happy faces were preferentially associated with an aversive outcome during Pavlovian conditioning relative to neutral faces, and that the persistence of this association for happy faces was related to inter-individual differences in their affective evaluation.

The conditioned response to angry and happy faces was more readily acquired and more persistent than the conditioned response to neutral faces, thus reflecting learning biases associated with these stimuli. Moreover, the conditioned response to happy faces during extinction was greater in participants who were faster to associate them with the attribute of importance (vs. unimportance) in the Go/No-go Association Task. In comparison, no such relationship was found for angry and neutral faces (see supplemental materials). Whereas the results obtained for angry faces align with well-established findings in the human conditioning literature (e.g., Öhman & Dimberg, 1978; Rowles et al., 2012; see also Dimberg & Öhman, 1996; Mallan et al., 2013; Öhman & Mineka, 2001), the occurrence of learning biases to happy faces challenges the view that enhanced Pavlovian aversive conditioning is selective to threat-relevant stimuli (Öhman & Mineka, 2001; Seligman, 1971). Conversely, our results indicate that positive stimuli with moderate affective relevance can also be rapidly and persistently associated with an aversive event, with these effects being moderate to small. They further show that individual differences in affective evaluation may affect the emergence of learning biases. In this respect, our findings replicate and expand recent

evidence supporting the appraisal-based predictions according to which preferential Pavlovian aversive learning is driven by affective relevance without being bound to a specific valence or inherent threat value, and can be modulated by individual differences in the way the stimulus is appraised in relation to the individual's concerns (Stussi et al., 2018, 2019).

At the computational level, the effects of greater persistence of the conditioned response to angry faces was characterized by a lower inhibitory learning rate. More specifically, the learning rate for negative prediction errors was lower to angry faces than to happy and neutral faces. This lower inhibitory learning altered the impact of negative prediction error signals, which likely contributed to weakening inhibitory learning underlying extinction (Dunsmoor, Niv, Daw, & Phelps, 2015). The observation that angry faces were associated with a lower inhibitory learning rate than happy faces additionally suggests that angry faces led to more persistent Pavlovian aversive conditioning, even though this difference was not visible when using conventional summary statistics on the conditioned response during extinction. This finding dovetails with the notion that happy faces hold a generally lower level of relevance to the organism than angry faces (Brosch et al., 2008, 2010; Pool et al., 2016), hence entailing smaller learning biases than angry faces. Happy faces were associated with a marginally lower inhibitory learning rate relative to neutral faces, but only inconclusive evidence was observed for this difference. Further evidence is thus required to determine whether the heightened conditioned response persistence to happy compared with neutral faces could be underlain by a lower inhibitory learning rate. In comparison, we did not find strong evidence that faster acquisition of the conditioned response to angry and happy faces than to neutral faces was driven by a higher excitatory learning rate. These results are partially inconsistent with previous studies using reward learning paradigms (Watanabe & Haruno, 2015; Watanabe, Sakagami, & Haruno, 2013), which reported that threat-related (i.e., fearful) faces not only accelerated learning in

comparison to neutral faces, but also increased the associated excitatory learning rate. Tentatively, this discrepancy may be due to habituation effects in the skin conductance response in the present case, which could have biased the estimation of the excitatory learning rates and mitigated the emergence of robust differences between the face categories.

The fact that happy faces led to a relatively small learning bias during extinction could potentially account for failures to report a resistance-to-extinction effect for this specific emotional category in prior research (see, e.g., Bramwell et al., 2014; Esteves et al., 1994; Mazurski et al., 1996; Öhman & Dimberg, 1978; Rowles et al., 2012; see also Dimberg & Öhman, 1996; Öhman & Mineka, 2001). Indeed, past studies have generally used between-participant designs (but see Bramwell et al., 2014) that are less sensitive than within-participant designs (see, e.g., Ho & Lipp, 2014), and importantly, often with modest sample sizes, typically varying from 15 to 25 participants by group. These two methodological factors likely contributed to hindering the possibility to reveal the existence of learning biases to happy faces given that, as our results suggest here with the use of a larger sample and stringent within-participant design, this bias has a small effect size[2]. It is therefore highly desirable in future research to set up adequately-powered experiments when the goal is to explore differences in Pavlovian aversive learning to happy compared with neutral or angry faces.

Although our study shows that inter-individual differences in stimulus affective evaluation can exert an effect on learning biases in Pavlovian conditioning, we only found a clear relationship between the conditioned response to happy faces during extinction and the differential reaction time index, but not with the differential d' index–this dissociation likely stemming from the putative lower reliability of this latter index (Nosek & Banaji, 2001)–or

[2] Additional post-hoc power analyses corroborated this assumption in indicating that achieved power to detect a small effect as reported in the present study (g_{av} = 0.247) using a one-tailed *t* test and an alpha level of .05 with a sample size ranging from 15 to 25 participants per group would vary between 23.14% and 32.83% for a within-participant design, and between 16.24% and 21.66% for a between-participant design.

during early acquisition. In addition, we found no evidence that inter-individual differences in extraversion affected the conditioned response to happy faces during either early acquisition or extinction, which is at odds with our predictions. Speculatively, this null result might arise from a relative lack of heterogeneity in the current sample's extraversion scores (see Figure S1; see Rolland et al., 1998, for a comparison with normative data from a similar student population). For these reasons, caution is warranted in the interpretation of the specific dimensions that underlain the impact of individual differences in affective evaluation on the conditioned response to happy faces during extinction, and these findings await replication in future studies before stronger conclusions might be drawn.

Another caveat pertains to the Go/No-go Association Task that we used in the sense that it probably did not provide a direct and pure measure of the affective relevance or importance value of the face categories. Results of this task showed that participants more easily associated happy faces with importance (vs. unimportance) than they did for angry and neutral faces. This suggests that the Go/No-go Association task rather captured the stimuli's valence and may have reflected participants' preferences or liking toward the face categories (Nosek & Banaji, 2001). Accordingly, it is possible that differential preferences toward happy faces actually drove the conditioned response persistence to these faces in the present study.

As angry and happy faces are usually considered as more arousing than neutral faces, it could be argued that these faces induced enhanced Pavlovian aversive conditioning because of their higher arousal value rather than, or in addition to, their affective relevance. Appraisal theories (e.g., Sander et al., 2003, 2005, 2018) suggest that stimuli appraised as relevant to the organism's concerns often trigger a physiological state of arousal that can be felt consciously as a consequence of the elicitation of a motivational state (see Montagrin & Sander, 2016; Pool et al., 2016), hence rendering it difficult to disentangle the specific

contributions of affective relevance and arousal from one another. Although we cannot rule out that arousal contributed to our findings, it seems unlikely that they were solely determined by felt and/or physiological arousal (see Stussi et al., 2018, for a related discussion). In fact, previous studies (Hamm, Greenwald, Bradley, & Lang, 1993; Hamm & Stark, 1993; Hamm & Vaitl, 1996) have reported that highly arousing negative and positive stimuli, without taking into account their affective relevance to the organism, did not produce preferential Pavlovian aversive conditioning relative to less arousing stimuli. Moreover, supplementary analysis of the habituation phase[3] revealed that (a) angry faces elicited larger skin conductance responses than happy faces before conditioning, whereas no difference emerged between angry and neutral faces, and between happy and neutral faces, and (b) the skin conductance responses to the various face categories during habituation did not correlate with the conditioned response to these stimuli during early acquisition and extinction. These considerations suggest that an explanation in terms of arousal alone does not satisfactorily account for the occurrence of differential learning biases to both angry and happy faces.

Alternatively, our results could also be interpreted as reflecting the involvement of two different mechanisms instead of a single relevance detection mechanism: a specialized mechanism selectively acting on threat-related stimuli that is consistently engaged across individuals, and a more general one acting on affectively relevant stimuli that is more sensitive to individual differences. Future research is needed to disentangle these two competing explanations, for instance by investigating at the neurobiological level whether

[3] A repeated-measures ANOVA with CS type (CS+ vs. CS-) and CS category (angry vs. happy vs. neutral) as within-participant factors performed on the skin conductance response data during habituation revealed a main effect of CS category, $F(2, 212) = 4.20$, $p = .016$, partial $\eta^2 = .038$, 90% CI [.004, .083]. Further post-hoc comparisons using Tukey's HSD tests indicated that angry faces elicited larger skin conductance responses than happy faces ($p = .012$, $g_{av} = 0.215$, 95% CI [0.064, 0.369]), whereas no statistically significant difference was found between angry and neutral faces ($p = .190$, $g_{av} = 0.129$, 95% CI [-0.019, 0.279]) or between happy and neutral faces ($p = .497$, $g_{av} = -0.088$, 95% CI [-0.239, 0.062]). Pearson's correlation analyses moreover showed no statistically significant relationship between the skin conductance responses to the different faces during habituation and the conditioned response to these faces during the early acquisition phase (-.129 < all rs(105) < .100, all ps > .18) or during the extinction phase (.001 < all rs(105) < .129, all ps > .18). Of note, computational learning models incorporating a Pavlovian bias to account for possible differences in inherent responding to the various CS categories did not provide a better fit to the normalized SCR data than the modified Rescorla-Wagner model implementing dual learning rates (see supplemental materials).

learning biases in Pavlovian aversive conditioning occurring in response to threat-relevant stimuli are underpinned by a threat-specific mechanism that is functionally distinct from a mechanism of relevance detection.

In conclusion, the present study highlights that positive stimuli with a relatively moderate level of relevance can be readily and persistently associated with an aversive outcome as is the case for threat-relevant stimuli, thus replicating and extending recent work showing that learning biases in Pavlovian aversive conditioning are not specific to threat-related stimuli, but can likewise occur for positive emotional stimuli (Stussi et al., 2018). Our results furthermore suggest that inter-individual differences may play a key role in the development of these learning biases (Stussi et al., 2019; see also Lonsdorf & Merz, 2017). In this context, our study suggests that the determinants of Pavlovian aversive conditioning are more flexible than previously thought and may adaptively rely on the interaction between the stimulus at play and the individuals' current concerns. These findings thereby contribute to further advancing and refining our understanding of the basic mechanisms underlying emotional learning in humans, and could ultimately provide insights into impairments in this process that are typically associated with specific emotional disorders, including anxiety, phobia, or addictions.

References

Åhs, F., Rosén, J., Kastrati, G., Fredrikson, M., Agren, T., & Lundström, J. N. (2018). Biological preparedness and resistance to extinction of skin conductance responses conditioned to fear relevant animal pictures: A systematic review. *Neuroscience and Biobehavioral Reviews, 95*, 430-437. http://dx.doi.org/10.1016/j.neubiorev.2018.10.017

Ambadar, Z., Cohn, J. F., & Reed, L. I. (2009). All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous. *Journal of Nonverbal Behavior, 33*, 17-34. http://dx.doi.org/10.1007/s10919-008-0059-5

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*, 289-300. http://dx.doi.org/10.2307/2346101

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. http://dx.doi.org/10.1111/ejn.12094

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10*, 433-436. http://dx.doi.org/10.1163/156856897x00357

Bramwell, S., Mallan, K. M., & Lipp, O. V. (2014). Are two threats worse than one? The effects of face race and emotional expression on fear conditioning. *Psychophysiology, 51*, 152-158. http://dx.doi.org/10.1111/psyp.12155

Brosch, T., Pourtois, G., & Sander, D. (2010). The perception and categorisation of emotional stimuli: A review. *Cognition and Emotion, 24*, 377-400. http://dx.doi.org/10.1080/02699930902975754

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

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. http://dx.doi.org/10.1126/science.1068749

Coppin, G., Pool, E., Delplanque, S., Oud, B., Magot, C., Sander, D., & Van Bavel, J. J. (2016). Swiss identity smells like chocolate: Social identity shapes olfactory judgments. *Scientific Reports, 6*, 34979. http://dx.doi.org/10.1038/srep34979

Costa, P. T. Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Critcher, C. R., & Ferguson, M. J. (2016). "Whether I like it or not, it's important": Implicit importance of means predicts self-regulatory persistence and success. *Journal of Personality and Social Psychology, 110*, 818-839. http://dx.doi.org/10.1037/pspa0000053

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

de Berker, A. O., Tirole, M., Rutledge, R. B., Cross, G. F., Dolan, R. J., & Bestmann, S. (2016). Acute stress selectively impairs learning to act. *Scientific Reports, 6*, 29816. http://dx.doi.org/10.1038/srep29816

Delgado, M. R., Olsson, A., & Phelps, E. A. (2006). Extending animal models of fear conditioning to humans. *Biological Psychology, 73*, 39-48. http://dx.doi.org/10.1016/j.biopsycho.2006.01.006

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6*, 274-290. http://dx.doi.org/10.1177/1745691611406220

Dimberg, U., & Öhman, A. (1996). Behold the wrath: Psychophysiological responses to facial stimuli. *Motivation and Emotion, 20*, 149-182. http://dx.doi.org/10.1007/BF02253869

Dubois, J., & Adolphs, R. (2016). Building a science of individual differences from fMRI. *Trends in Cognitive Sciences, 20*, 425-443. http://dx.doi.org/10.1016/j.tics.2016.03.014

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

Esteves, F., Parra, C., Dimberg, U., & Öhman, A. (1994). Nonconscious associative learning: Pavlovian conditioning of skin conductance responses to masked fear-relevant facial stimuli. *Psychophysiology, 31*, 375-385. http://dx.doi.org/10.1111/j.1469-8986.1994.tb02446.x

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. http://dx.doi.org/10.3758/BF03193146

Frijda, N. H. (1986). *The emotions*. London, UK: Cambridge University Press.

Garcia, J., & Koelling, R. A. (1966). Relation of cue to consequence in avoidance learning. *Psychonomic Science, 4*, 123-124. http://dx.doi.org/10.3758/BF03342209

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

Goeleven, E., De Raedt, R., Leyman, L., & Verschuere, B. (2008). The Karolinska Directed Emotional Faces: A validation study. *Cognition and Emotion, 22*, 1094-1118. http://dx.doi.org/10.1080/02699930701626582

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

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. http://dx.doi.org/10.1037/0021-843X.102.3.435

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. http://dx.doi.org/10.1007/BF02691223

Hamm, A. O., & Vaitl, D. (1996). Affective learning: Awareness and aversion. *Psychophysiology, 33*, 698-710. http://dx.doi.org/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. http://dx.doi.org/10.1111/psyp.12223

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

https://doi.org/10.17605/OSF.IO/DK2NP

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. http://dx.doi.org/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. http://dx.doi.org/10.3389/fpsyg.2013.00863

LeDoux, J. E., & Daw, N. D. (2018). Surviving threats: Neural circuit and computational implications of a new taxonomy of defensive behaviour. *Nature Reviews Neuroscience, 19*, 269-282. http://dx.doi.org/10.1038/nrn.2018.22

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

Lindström, B., Golkar, A., & Olsson, A. (2015). A clash of values: Fear-relevant stimuli can enhance or corrupt adaptive behavior through competition between Pavlovian and instrumental valuation systems. *Emotion, 15*, 668-676. http://dx.doi.org/10.1037/emo0000075

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. http://dx.doi.org/10.1016/j.neubiorev.2017.02.026

Lonsdorf, T. B., & Merz, C. J. (2017). More than just noise: Inter-individual differences in fear acquisition, extinction and return of fear in humans – Biological, experiential, temperamental factors, and methodological pitfalls. *Neuroscience and Biobehavioral Reviews, 80*, 703-728. http://dx.doi.org/10.1016/j.neurobiorev.2017.07.007

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

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. New York, NY: Psychology Press.

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. http://dx.doi.org/10.1080/02699931.2013.778195

Martin, J., Rychlowska, M., Wood, A., & Niedenthal, P. (2017). Smiles as multipurpose social signals. *Trends in Cognitive Sciences, 21*, 864-877. http://dx.doi.org/10.1016/j.tics.2017.08.007

Mazurski, E. J., Bond, N. W., Siddle, D. A. T., & Lovibond, P. F. (1996). Conditioning with facial expressions of emotion: Effects of CS sex and age. *Psychophysiology, 33*, 416-425. http://dx.doi.org/10.1111/j.1469-8986.1996.tb01067.x

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

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology, 4*, 61-64. http://dx.doi.org/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. *The Journal of Neuroscience, 32*, 551-562. http://dx.doi.org/10.1523/JNEUROSCI.5498-10.2012

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

Nosek, B. A., & Banaji, M. R. (2001). The Go/No-go Association Task. *Social Cognition, 19*, 625-666. http://dx.doi.org/10.1521/soco.19.6.625.20886

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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/10.1037/0033-295X.108.3.483

Olsson, A., Carmona, S., Downey, G., Bolger, N., & Ochsner, K. N. (2013). Learning biases underlying individual differences in sensitivity to social rejection. *Emotion, 13*, 616-621. http://dx.doi.org/10.1037/a0033150

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. http://dx.doi.org/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. http://dx.doi.org/10.1111/j.0956-7976.2004.00762.x

Pauli, W. M., Larsen, T., Collette, S., Tyszka, J. M., Seymour, B., & O'Doherty, J. P. (2015). Distinct contributions of ventromedial and dorsolateral subregions of the human substantia nigra to appetitive and aversive learning. *The Journal of Neuroscience, 35*, 14220-14233. http://dx.doi.org/10.1523/JNEUROSCI.2277-15.2015

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. http://dx.doi.org/10.1037/0033-295X.87.6.532

Pelli, D. G. (1997). The VideoToolbox software for psychophysics: Transforming numbers into movies. *Spatial Vision, 10*, 437-442. http://dx.doi.org/1163/156856897x00366

Phelps, E. A., & LeDoux, J. E. (2005). Contributions of the amygdala to emotion processing: From animal models to human behavior. *Neuron, 48*, 175-187. http://dx.doi.org/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. http://dx.doi.org/ 10.1037/bul0000026

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. http://dx.doi.org/10.1371/journal.pcbi.1002918

RStudio Team (2016). *RStudio: Integrated development for R*. RStudio, Inc., Boston, MA. Retrieved from: https://www.rstudio.com/

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist, 43*, 151-160. http://dx.doi.org/10.1037/0003-066X.43.3.151

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.

Rolland, J. P., Parker, W. D., & Strumpf, H. (1998). A psychometric examination of the French translations of the NEO-PI-R and NEO-FFI. *Journal of Personality Assessment, 71*, 269-291. http://dx.doi.org/10.1207/s15327752jpa7102_13

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* test for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225-237. http://dx.doi.org/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. http://dx.doi.org/10.1111/j.1469-8986.2011.01308.x

Sander, D., Grafman, J., & Zalla, T. (2003). The human amygdala: An evolved system for relevance detection. *Reviews in the Neurosciences, 14*, 303-316. http://dx.doi.org/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. http://dx.doi.org/10.1016/j.neunet.2005.03.001

Sander, D., Grandjean, D., & Scherer, K. R. (2018). An appraisal-driven componential approach to the emotional brain. *Emotion Review, 10*, 219-231. http://dx.doi.org/10.1177/1754073918765653

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

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

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

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. http://dx.doi.org/10.1016/j.biopsycho.2015.06.008

Stussi, Y., Ferrero, A., Pourtois, G., & Sander, D. (2019). Achievement motivation modulates Pavlovian aversive conditioning to goal-relevant stimuli. *npj Science of Learning, 4*, 4. http://dx.doi.org/10.1038/s41539-019-0043-3

Stussi, Y., Pourtois, G., & Sander, D. (2018). Enhanced Pavlovian aversive conditioning to positive emotional stimuli. *Journal of Experimental Psychology: General, 147*, 905-923. http://dx.doi.org/10.1037/xge0000424

Watanabe, N., & Haruno, M. (2015). Effects of subconscious and conscious emotions on human cue-reward association learning. *Scientific Reports, 5*, 8478. http://dx.doi.org/10.1038/srep08478

Watanabe, N., Sakagami, M., & Haruno, M. (2013). Reward prediction error signal enhanced by striatum-amygdala interaction explains the acceleration of probabilistic reward learning by emotion. *The Journal of Neuroscience, 33*, 4487-4493. http://dx.doi.org/10.1523/JNEUROSCI.3400-12.2013

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

Supplemental Materials

Learning biases to angry and happy faces during Pavlovian aversive conditioning

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

Supplemental Method and Results

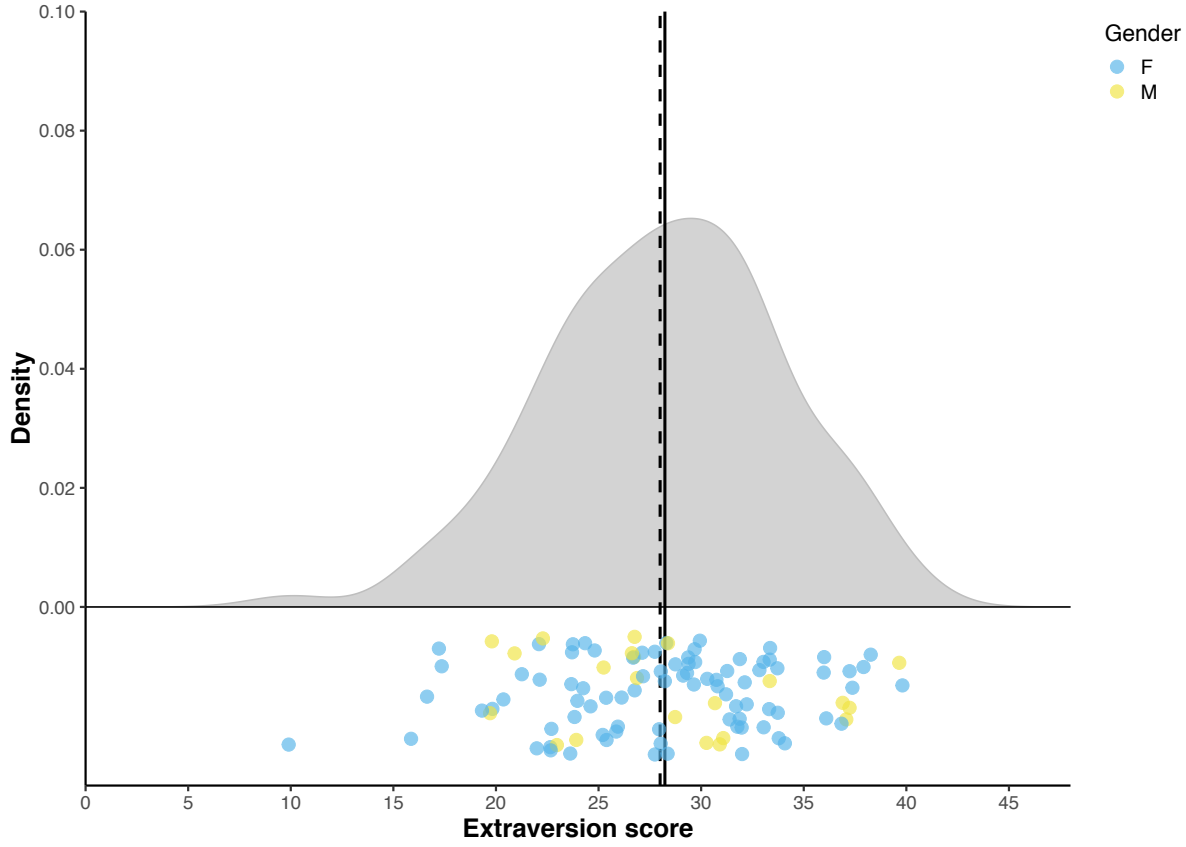

Figure S1. Distribution of extraversion scores as measured with the NEO-FFI (Costa & McCrae, 1992; Rolland, Parker, & Strumpf, 1998). The dots indicate data for individual participants. The solid line indicates the mean extraversion score and the dashed line the median extraversion score.

Go/No-go Association Task

Face stimuli from the Karolinska Directed Emotional Faces (KDEF; Lundqvist, Flykt, & Öhman, 1998) were used either as targets or as distractors in the Go/No-go Association Task (GNAT; Nosek & Banaji, 2001). They consisted of eight angry faces (model numbers for targets: AM05ANS, AM09ANS, AM17ANS, AM30ANS; model numbers for distractors: AM14ANS, AM19ANS, AM21ANS, AM24ANS), eight happy faces (model numbers for targets: AM20HAS, AM23HAS, AM25HAS, AM26HAS; model number for distractors: AM04HAS, AM12HAS, AM16HAS, AM32HAS), and eight neutral faces (models numbers for targets: AM01NES, AM06NES, AM08NES, AM13NES; model numbers for distractors: AM02NES, AM18NES, AM28NES, AM35NES).

The practice session of the GNAT included five blocks in which there was only a single target category. In the first three blocks, participants learned to discriminate between the different face categories, each of them being the target category in one of the blocks whereas the two other categories were distractors; the order being counterbalanced across participants. In these blocks, four faces from the target face category and two faces from each distractor face category were each presented twice in a pseudorandom order. In the last two practice blocks, participants were presented with the four "important" and the four "unimportant" words, which were each presented twice in a pseudorandom order. The "important" words were targets and the "unimportant" words distractors in the fourth block, which was reversed in the last block. Each practice block consisted of 16 trials and used a 666-ms response deadline.

To assess whether a trade-off relationship occurred between reaction times (RTs) and accuracy in the GNAT, we conducted point-biserial correlations between individual participants' trial-by-trial RTs and accuracy (0 [incorrect] vs. 1 [correct]) for each face category. To do so, we calculated a correlation coefficient for each participant separately

before averaging these coefficients. These analyses showed a weak positive relationship between participants' RTs and accuracy for angry (mean r = .197, SD = 0.12, range = -.201–.486), happy (mean r = .203, SD = 0.13, range = -.184–.477]), and neutral faces (mean r = .256, SD = 0.11, range = -.155–.517). In addition, we tested whether participants were overall slower for correct trials relative to incorrect trials across the three face categories by means of a two-way repeated-measures ANOVA on the mean RT data. We observed significant main effects of face category, $F(2, 212) = 16.29$, $p < .001$, partial $\eta^2 = .133$, 90% CI [.066, .200], and of accuracy, $F(1, 106) = 696.48$, $p < .001$, partial $\eta^2 = .868$, 90% CI [.830, .891]. These main effects were qualified by their interaction, $F(2, 212) = 6.77$, $p = .001$, partial $\eta^2 = .060$, 90% CI [.015, .113]. Follow-up comparisons using Tukey's HSD confirmed that the RTs in correct trials were slower than in incorrect trials for angry ($p < .001$, $g_{av} = 1.569$, 95% CI [1.291, 1.867]), happy ($p < .001$, $g_{av} = 1.520$, 95% CI [1.243, 1.814]), and neutral faces ($p < .001$, $g_{av} = 1.859$, 95% CI [1.565, 2.176]). Additional post-hoc comparisons further revealed faster RTs in correct trials for angry ($p < .001$, $g_{av} = 0.432$, 95% CI [0.249, 0.619]) and happy ($p < .001$, $g_{av} = 0.620$, 95% CI [0.437, 0.809]) faces compared with neutral faces, whereas there was no statistical difference in RTs in correct trials between happy and angry faces ($p = .085$, $g_{av} = 0.207$, 95% CI [0.016, 0.401]), or across the face categories in incorrect trials (all ps > .06, 0.02 < g_{av}s < 0.27).

Subjective ratings

Subsequent to the GNAT but before the differential Pavlovian aversive conditioning procedure, participants provided subjective ratings of the two angry face conditioned stimuli (CSs), the two happy face CSs, and the two neutral face CSs as a function of their pleasantness, subjective arousal, and subjective relevance. In this procedure, the faces were presented to participants along with a visual analog scale (VAS). For the pleasantness ratings, participants were asked to rate the degree to which the face was unpleasant or pleasant from 0

(*very unpleasant*) to 100 (*very pleasant*). For the arousal ratings, they were asked to rate the degree to which the face was arousing from 0 (*not at all arousing*) to 100 (*very arousing*). For the relevance ratings, participants were asked to rate the degree to which the face was important to them from 0 (*not at all important*) to 100 (*very important*). After the end of the conditioning procedure, participants completed again pleasantness, arousal, and relevance ratings of the CSs using the same procedure as for the preconditioning ratings. In addition, they were asked to rate how many electric stimulations they received in response to each CS on a Likert scale from 0 to 9 to assess their explicit awareness of the CS-US contingencies. The order of the CS presentations and the questions was randomized between participants for both the preconditioning and postconditioning ratings.

The pleasantness, arousal, and relevance ratings were analyzed with separate three-way repeated-measures analyses of variance (ANOVAs) with time (pre vs. post), CS category (angry vs. happy vs. neutral), and CS type (CS+ vs. CS-) as within-participant factors, whereas the CS-US contingency ratings were analyzed with a two-way repeated-measures ANOVA with CS category (anger vs. happy vs. neutral) and CS type (CS+ vs. CS-) as within-participant factors. Statistically significant effects were followed up with more focused repeated-measures ANOVAs and/or a multiple comparison procedure using Tukey's HSD tests when applicable.

Analysis of the pleasantness ratings (see Figure S2a) showed a three-way interaction between time, CS category, and CS type, $F(2, 212) = 5.29$, $p = .006$, partial $\eta^2 = .048$, 90% CI [.008, .096]. A follow-up 3 (CS category: angry vs. happy vs. neutral) × 2 (CS type: CS+ vs. CS-) repeated-measures ANOVA for the preconditioning ratings indicated that the CS categories modulated the CSs' rated pleasantness before conditioning, $F(1.73, 183.12) = 323.15$, $p < .001$, partial $\eta^2 = .753$, 90% CI [.707, .785]. As expected, happy faces were deemed more pleasant than angry faces ($p < .001$, $g_{av} = 2.887$, 95% CI [2.432, 3.378]) and

neutral faces ($p < .001$, $g_{av} = 1.438$, 95% CI [1.149, 1.743]), whereas neutral faces were evaluated as more pleasant than angry faces ($p < .001$, $g_{av} = 1.933$, 95% CI [1.590, 2.298]). The follow-up repeated-measures ANOVA for the postconditioning ratings revealed an interaction effect between CS category and CS type, $F(2, 212) = 6.40$, $p = .002$, partial $\eta^2 = .057$, 90% CI [.013, .109], reflecting that the difference in rated pleasantness between the CS+ and the CS- was higher for happy faces than angry and neutral faces. The CS+ was evaluated as less pleasant than the CS- for angry faces ($p = .015$, $g_{av} = 0.404$, 95% CI [0.160, 0.652]), happy faces ($p < .001$, $g_{av} = 0.811$, 95% CI [0.546, 1.085]), and neutral faces ($p < .001$, $g_{av} = 0.662$, 95% CI [0.401, 0.929]). Furthermore, happy faces were rated as more pleasant than angry faces (all ps $< .001$, $0.80 < g_{av}$s < 2.26), and neutral faces were deemed more pleasant than angry faces (all ps $< .04$, $0.39 < g_{av}$s < 1.59). The happy face CS- was likewise evaluated as more pleasant than the neutral face CS+ and CS- ($p < .001$, $g_{av} = 1.462$, 95% CI [1.133, 1.808], and $p < .001$, $g_{av} = 0.950$, 95% CI [0.669, 1.241], respectively), and the happy face CS+ as more pleasant than the neutral face CS+ ($p < .001$, $g_{av} = 0.479$, 95% CI [0.253, 0.710]), whereas there was no statistical difference in rated pleasantness between the happy face CS+ and the neutral face CS- ($p = .999$, $g_{av} = -0.044$, 95% CI [-0.326, 0.238]).

The arousal ratings analysis (see Figure S2b) revealed an interaction between time and CS type, $F(1, 106) = 87.23$, $p < .001$, partial $\eta^2 = .451$, 90% CI [.335, .541]. Before conditioning, the CSs+ and the CSs- did not statistically differ in felt arousal ($p > .99$, $g_{av} = 0.004$, 95% CI [-0.155, 0.163]); by contrast, the CSs+ were rated as more arousing than the CSs- after conditioning ($p < .001$, $g_{av} = 1.149$, 95% CI [0.874, 1.436]). The CSs+ were also deemed more arousing after conditioning than before it ($p < .001$, $g_{av} = 0.872$, 95% CI [0.645, 1.108]), whereas the CSs- were deemed less arousing after than before conditioning ($p < .001$, $g_{av} = 0.382$, 95% CI [0.188, 0.581]). Moreover, the interaction between time and CS category yielded statistical significance, $F(2, 212) = 22.81$, $p < .001$, partial $\eta^2 = .177$, 90%

CI [.101, .248]. Angry and happy faces were evaluated as more arousing than neutral faces both in the preconditioning and postconditioning ratings (all ps < .001, 0.86 < g_{av}s < 1.86), but did not differ statistically between each other (all ps > .55, 0.03 < g_{av}s < 0.18). In addition, neutral faces were evaluated as more arousing after than before conditioning (p < .001, g_{av} = 0.837, 95% CI [0.559, 1.123]), which was not the case for angry faces (p = .949, g_{av} = 0.070, 95% CI [-0.086, 0.226]) and happy faces (p = .953, g_{av} = -0.077, 95% CI [-0.253, 0.098]).

For the relevance ratings (see Figure S2c), the analysis showed an interaction effect of time and CS type, F(1, 106) = 38.56, p < .001, partial η^2 = .267, 90% CI [.153, .371]. Whereas there was no statistical difference in relevance ratings between the CSs+ and the CSs- prior to conditioning (p = .843, g_{av} = 0.050, 95% CI [-0.070, 0.171]), the CSs+ were deemed more relevant than the CSs- after conditioning (p < .001, g_{av} = 0.786, 95% CI [0.539, 1.042]). Furthermore, the CSs+ were rated as more relevant after than before conditioning (p < .001, g_{av} = 0.627, 95% CI [0.421, 0.840]), which was not the case for the CSs- (p = .489, g_{av} = -0.133, 95% CI [-0.319, 0.052]). We also observed an interaction between time and CS category, F(2, 212) = 28.41, p < .001, partial η^2 = .211, 90% CI [.131, .283]. Happy faces were evaluated as more relevant than angry and neutral faces both before and after conditioning (all ps < .001, 0.57 < g_{av}s < 1.82), and angry faces as more relevant than neutral faces (all ps < .003, 0.39 < g_{av}s < 0.98). Neutral faces were additionally rated as higher in relevance after conditioning relative to before conditioning (p < .001, g_{av} = 0.897, 95% CI [0.633, 1.171]), whereas there was no statistical difference in preconditioning and postconditioning relevance ratings for angry faces (p = .876, g_{av} = 0.086, 95% CI [-0.067, 0.240]) and happy faces (p = .786, g_{av} = -0.110, 95% CI [-0.278, 0.057]).

The postconditioning ratings of CS-US contingency (see Figure S2d) revealed an interaction between the CS categories and the CS types, $F(2, 212) = 3.35$, $p = .037$, partial $\eta^2 = .031$, 90% CI [.001, .072]. Follow-up analyses indicated that the CS+ was rated to be associated with the delivery of more electric stimulations than the CS- for angry ($p < .001$, $g_{av} = 1.876$, 95% CI [1.495, 2.278]), happy ($p < .001$, $g_{av} = 2.345$, 95% CI [1.933, 2.784]), and neutral ($p < .001$, $g_{av} = 1.817$, 95% CI [1.468, 2.188]) faces. Additionally, participants evaluated the happy face CS+ as paired with more electric stimulations than the neutral face CS+ ($p = .010$, $g_{av} = 0.369$, 95% CI [0.155, 0.587]), whereas the difference between the angry face CS+ and the neutral face CS+ did not yield statistical significance ($p = .087$, $g_{av} = 0.312$, 95% CI [0.080, 0.547]). No difference was found between the angry and the happy face CSs+ ($p = .985$, $g_{av} = 0.071$, 95% CI [-0.136, 0.278]) or between the CSs- among the three CS categories (all ps > .16, 0.03 < g_{av}s < 0.25).

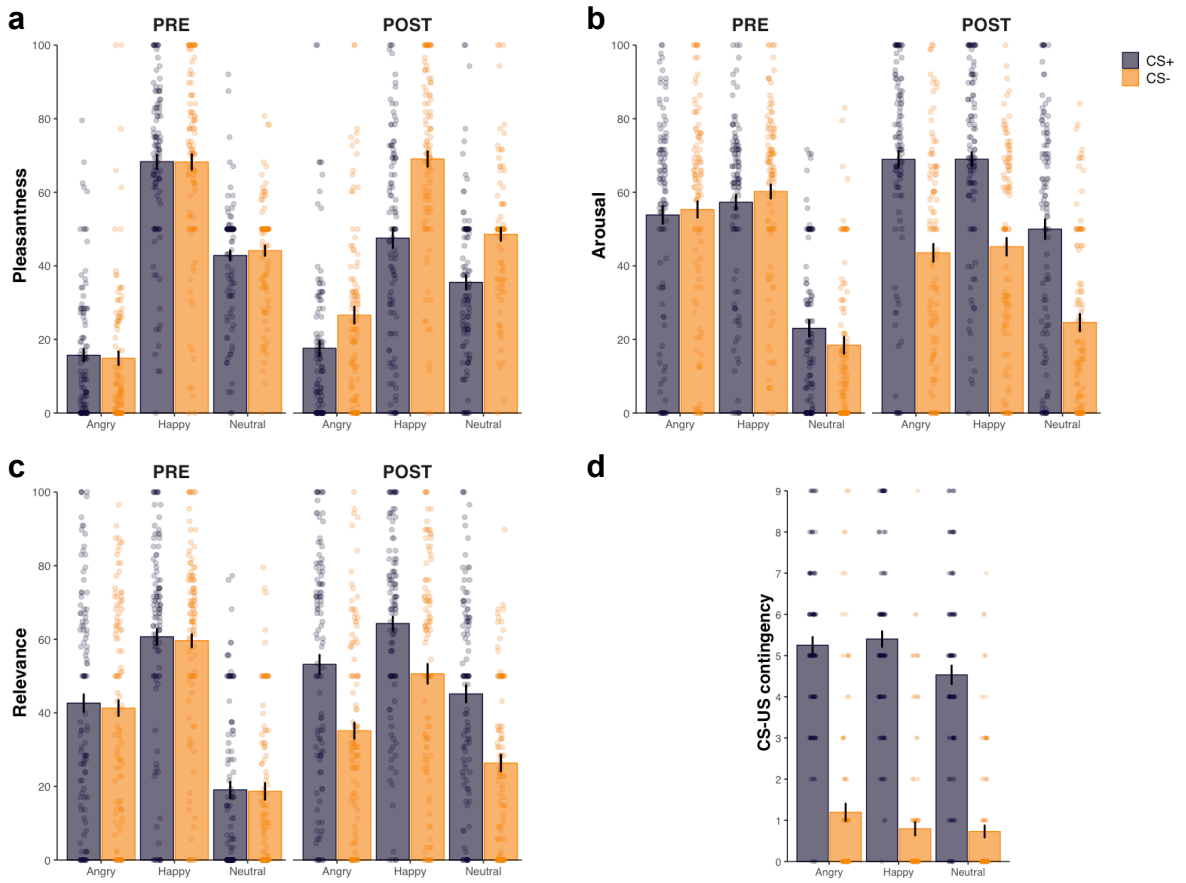

Figure S2. Subjective ratings before (pre) and after (post) the conditioning procedure as a function of conditioned stimulus type (CS+ vs. CS-) and stimulus category (angry vs. happy vs. neutral). Mean (a) pleasantness ratings, (b) arousal ratings, (c) relevance ratings, and (d) CS-US contingency ratings. The dots indicate data for individual participants. Error bars indicate ± 1 *SEM* adjusted for within-participant designs (Morey, 2008).

Computational modeling

To characterize and provide insights into the computations underlying the influence of angry and happy faces, as opposed to neutral faces, on Pavlovian aversive conditioning, we constructed simple reinforcement learning models (Li, Schiller, Schoenbaum, Phelps, & Daw, 2011; Pearce & Hall, 1980; Rescorla & Wagner, 1972; see also Stussi, Pourtois, & Sander, 2018) and fitted them to the trial-by-trial normalized (i.e., scaled and squared-root-transformed) skin conductance response (SCR) data for each CS category separately in order to estimate the models' free parameters and to identify the best-fitting model. After selection of the best-fitting model, its parameter estimates were subsequently compared across angry, happy, and neutral face CSs. In addition to the modified Rescorla-Wagner model implementing dual learning rates reported in the main text, we considered the following alternative models.

Rescorla-Wagner model. According to the Rescorla-Wagner model (Rescorla & Wagner, 1972), learning occurs when events deviate from expectations and correspondingly serves to update future expectations (Niv & Schoenbaum, 2008). It formalizes the notion of prediction error by stating that associative learning is directly driven by the discrepancy between the actual and the expected outcome. In this model, the predictive value (or associative strength) V at trial $t + 1$ of a given CS j is updated on the basis of the sum of the

current predictive value V_j at trial t and the prediction error between the predictive 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 (US) was delivered on the current trial t, $R(t) = 1$, else $R(t) = 0$.

Rescorla-Wagner model with a Pavlovian bias. Because we used CSs having a preexisting emotional value prior to conditioning, we tested a modified version of the Rescorla-Wagner model including a Pavlovian bias (see, e.g., de Berker et al., 2016; Guitart-Masip et al., 2012), which considered the influence of the CS categories' inherent features on learning (i.e., through the update of the CS predictive values indexed by the normalized SCR) beyond the CS-US contingency. In this model, the Pavlovian bias aims to capture inherent responding to the various CS categories by (a) affecting the impact of the CS current value on the computation of the updated CS value, with stimuli associated with a higher Pavlovian bias assigning a greater weight to the CS current value, and (b) setting the CS initial value V_0 at the estimated value of the Pavlovian bias for each CS category separately, thereby modeling initial responding to the CSs before conditioning. In accordance, the Rescorla-Wagner model with a Pavlovian bias updates the predictive value V of a given CS j as follows:

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

where the learning rate α and the Pavlovian bias b are free parameters within the range [0, 1]. As for the standard Rescorla-Wagner model, if the US was delivered on the current trial t, $R(t) = 1$, else $R(t) = 0$. According to this model, a higher Pavlovian bias leads to greater responding before conditioning (i.e., during habituation) as well as during acquisition and extinction relative to a lower Pavlovian bias. This adapted version of the Rescorla-Wagner model thus allows for accommodating how certain stimulus categories can be associated with

enhanced initial responding during habituation, along with a heightened CR during acquisition and extinction.

Rescorla-Wagner with dual learning rates and a Pavlovian bias. To determine whether the CS categories could differentially influence excitatory and inhibitory learning processes independently of, or in combination with, their inherent responding, we adapted the Rescorla-Wagner model by incorporating both dual learning rates and a Pavlovian bias. In this model, the predictive value V of a given CS j is updated as indicated below:

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

where the learning rate for positive prediction errors α^+, the learning rate for negative prediction errors α^-, and the Pavlovian bias b are free parameters within the range [0, 1]. If the US was delivered on the current trial t, $R(t) = 1$, else $R(t) = 0$. As for the previous model, the initial CS values V_0 were set at the value of the estimated Pavlovian bias for each CS category.

Hybrid model. In addition to maintaining the basic assumption that learning is directly driven by prediction errors as stated in the Rescorla-Wagner model, the hybrid model proposed by Li et al. (2011) incorporates the Pearce-Hall associability mechanism (Pearce & Hall, 1980). The Pearce-Hall model specifically asserts that the CS associability determines the learning rate and is dynamically modulated on each trial as a function of unsigned past prediction errors. According to the Pearce-Hall algorithm, the CS associability decreases when the CS accurately and reliably predicts the actual outcome, whereas it increases when the CS is an unreliable predictor of the actual outcome. In the hybrid model, the predictive value V and the associability α of a given CS j are 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 US was delivered on the current trial t, $R(t) = 1$, else $R(t) = 0$.

Hybrid model with dual learning rates. Similar to the modified Rescorla-Wagner model implementing different learning rates for positive and negative prediction errors, we also constructed a modified hybrid model with dual learning rates. In this modified version of the hybrid model, the predictive value V and the associability α of a given CS j are 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 errors κ^+, the learning rate for negative prediction errors κ^-, and the weighting factor η are free parameters within the range [0, 1]. If the US was delivered on the current trial t, $R(t) = 1$, else $R(t) = 0$.

Hybrid model with a Pavlovian bias. We additionally considered a hybrid model including a Pavlovian bias modulating the CS current value. According to this model, the predictive value V and the associability α of a given CS j are updated as shown below:

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

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

where the initial associability α_0, the learning rate κ, the Pavlovian bias b, and the weighting factor η are free parameters within the range [0, 1]. If the US was delivered on the current trial t, $R(t) = 1$, else $R(t) = 0$. Similarly to the versions of the Rescorla-Wagner model implementing a Pavlovian bias, the Pavlovian bias values likewise determined the initial CS values V_0.

Hybrid model with dual learning rates and a Pavlovian bias. The last model that we considered consisted of an adapted version of the hybrid model implementing both dual learning rates and a Pavlovian bias. In this model, the predictive value *V* and the associability α of a given CS *j* are updated as follows:

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

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

where the initial associability α_0, the learning rate for positive prediction errors κ^+, the learning rate for negative prediction errors κ^-, the Pavlovian bias *b*, and the weighting factor η are free parameters within the range [0, 1]. If the US was delivered on the current trial *t*, $R(t) = 1$, else $R(t) = 0$. The initial CS values V_0 were also set at the estimated Pavlovian bias values to account for possible differences in initial responding to the various CS categories prior to conditioning.

Model and parameter fitting. We fitted and optimized the models' free parameters using maximum a posteriori estimation, which consisted in finding the set of parameters maximizing the likelihood of each participant's trial-by-trial normalized SCRs to the CS given the model, constrained by a regularizing prior (Gershman, 2016; Niv et al., 2012). The free parameters were constrained with a Beta (1.2, 1.2) prior distribution that favors a normal distribution of the estimated parameters. Fifty random initializations were performed to obtain maximum likelihood estimates for each parameter in order to avoid local optima. We used the trial-by-trial timeseries of CS predictive values $V(t)$ to optimize the free parameters for the Rescorla-Wagner model (RW[*V*]), the modified Rescorla-Wagner model with dual learning rates (dual RW[*V*]), and their version implementing a Pavlovian bias (RW *b*[*V*] and dual RW *b*[*V*], respectively). For the hybrid model, the hybrid model with dual learning rates, and the hybrid models incorporating a Pavlovian bias, we optimized the free parameters

separately for each possible combination based on the trial-by-trial timeseries of CS values $V(t)$ (Hybrid[V], dual Hybrid[V], Hybrid b[V], and dual Hybrid b[V]), the trial-by-trial timeseries of CS associabilities $\alpha(t)$ (Hybrid[α], dual Hybrid[α], Hybrid b[α], and dual Hybrid b[α]), or the combination of both (Hybrid[$V+\alpha$], dual Hybrid[$V+\alpha$], Hybrid b[$V+\alpha$], and dual Hybrid b[$V+\alpha$]; see Li et al., 2011; Zhang, Mano, Ganesh, Robbins, & Seymour, 2016). Given that participants were expecting to receive electric stimulations at the outset of the Pavlovian aversive conditioning procedure because of the work-up procedure and the instructions, we set each CS initial predictive value V_0 to 0.5 for the models that did not incorporate a Pavlovian bias. We also conducted further analyses that modeled the initial values as free parameters to capture potential differential responding to the CSs prior to conditioning, but these analyses did not provide a better fit to the data. We fitted the various models using a separate set of free parameters for each participant (a) across all trials, and (b) separately for each CS category (Boll, Gamer, Gluth, Finsterbusch, & Büchel, 2013). This allowed for comparing the parameter estimates that best fitted to the normalized SCR data between the three different CS categories. Two participants were excluded from the computational analyses because their individual parameters could not be estimated due to a lack of SCR to all the angry face CSs during the experiment. The final sample size for the computational analyses included 105 participants (83 women, 22 men; mean age = 21.79 ± 2.46 years).

Model comparison. We performed model comparison with the Bayesian information criterion (BIC; Schwarz, 1978; see also Stussi et al., 2018; Zhang et al., 2016). In addition to providing a quantitative measure of the models' goodness of fit, the BIC considers and penalizes for the number of free parameters that the model includes. For each model, the mean BIC value was computed using the average of individual participant's estimated parameters. The models were additionally compared against a random model, in which the

predictive value $V_f(t)$ and the prediction errors 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). This allowed us to confirm that the reinforcement learning models that we used outperformed a model implementing random predictions. The mean BIC values for each model are reported in Table S1.

Relationship between modeled learning signals and participants' normalized skin conductance responses. We further assessed whether, and the extent to which, modeled predictive value and prediction error signals from the best-fitting model (i.e., the dual-learning-rate Rescorla-Wagner model; see Table S1) were predictive of the participants' trial-by-trial normalized SCRs (see Li et al., 2011; Pauli et al., 2015). To do so, we performed a multiple linear regression in which we regressed predictive value and prediction error timeseries generated with the individual parameter estimates from the dual-learning-rate Rescorla-Wagner model and averaged across participants against the averaged trial-by-trial normalized SCRs. This analysis revealed that predictive value and prediction error signals explained a statistically significant portion of variance of trial-by-trial normalized SCRs (R^2 = .638, 90% CI [.518, .725], adjusted R^2 = .630, $F(2, 87) = 76.84$, $p < .001$). Predictive value signals predicted trial-by-trial normalized SCRs, b = 0.442, 95% CI [0.371, 0.513], β = .799, $t(87) = 12.39$, $p < .001$ (see Figure S3), which was not the case for prediction error signals, b = 0.014, 95% CI [-0.024, 0.053], β = .048, $t(87) = 0.74$, p = .462.

Table S1

Goodness of fit to normalized skin conductance responses for individual models using the mean Bayesian information criterion (N = 105)

	Model																
CS category	RW(*V*)	Dual RW(*V*)	RW *b*(*V*)	Dual RW *b*(*V*)	Hybrid (*V*)	Hybrid (*α*)	Hybrid (*V*+*α*)	Dual Hybrid (*V*)	Dual Hybrid (*α*)	Dual Hybrid (*V*+*α*)	Hybrid *b*(*V*)	Hybrid *b*(*α*)	Hybrid *b*(*V*+*α*)	Dual Hybrid *b*(*V*)	Dual Hybrid *b*(*α*)	Dual Hybrid *b*(*V*+*α*)	Random
All	-0.59	-4.86	-3.57	-0.15	6.38	-1.17	-1.64	2.15	1.78	1.37	3.80	0.97	1.47	7.35	5.04	5.57	13.17
Angry	-0.32	-1.26	-0.24	2.22	5.46	1.07	0.93	4.36	3.07	3.36	5.41	2.07	3.08	7.89	5.74	6.33	7.31
Happy	-1.52	-2.96	-1.71	0.90	4.37	0.05	-0.26	2.83	1.92	1.93	3.85	1.81	2.05	6.39	4.81	4.52	5.17
Neutral	-3.83	-5.03	-5.05	-2.23	2.25	-2.09	-2.23	0.93	0.08	0.17	0.83	0.11	-0.43	3.65	3.24	2.76	2.09

Note. RW = Rescorla-Wagner model, *V* = predictive values, *α* = associabilities, Dual = dual-learning-rate, *b* = Pavlovian bias.

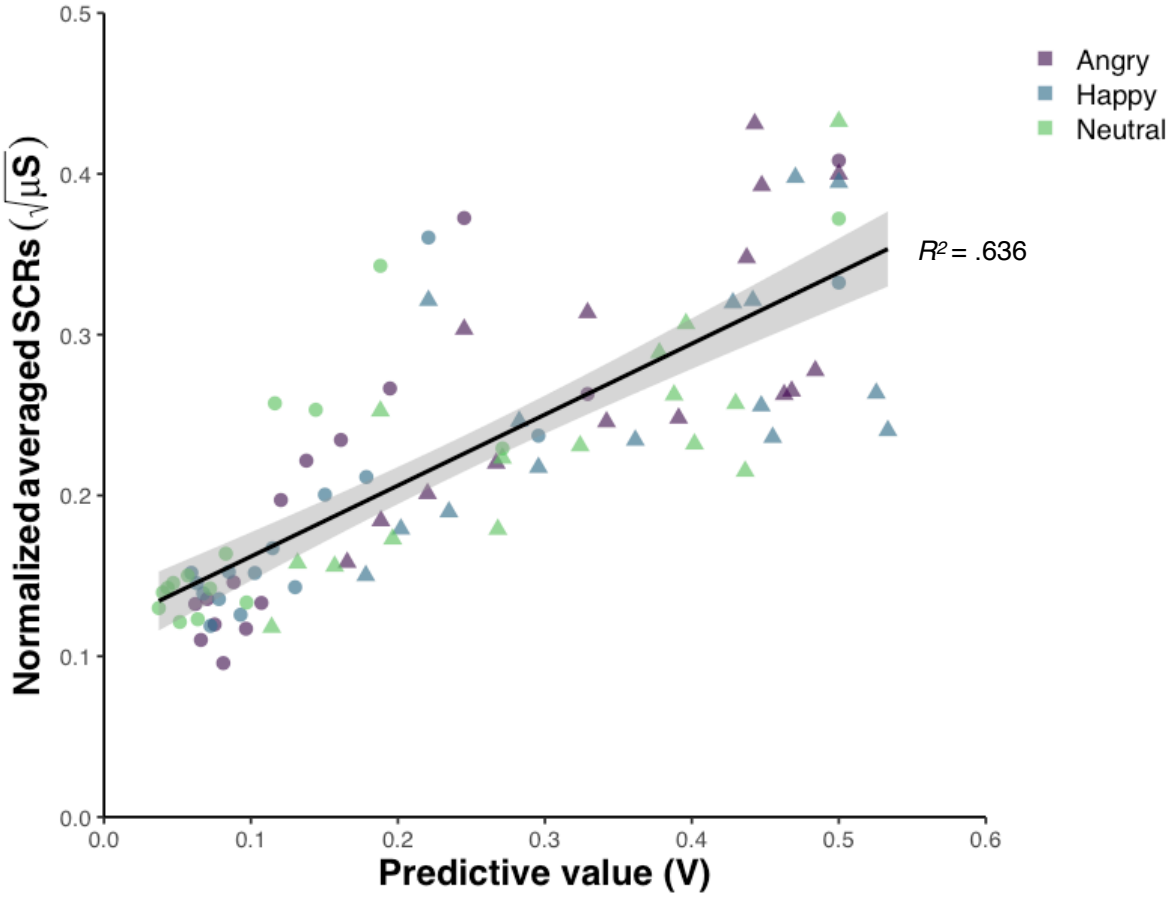

Figure S3. Relationship between modeled predictive values (*V*) and trial-by-trial normalized skin conductance responses (SCRs) averaged across participants using the individual best-fitting parameters for the Rescorla-Wagner model implementing dual learning rates. Triangles represent reinforced conditioned stimuli (CSs+) and circles represent unreinforced conditioned stimuli (CSs-). The line represents the fitted regression line using least squares estimation and 95% confidence interval.

Exploratory analyses

We carried out exploratory analyses to investigate whether the conditioned response to happy faces during early acquisition and extinction, as well as the estimated learning rates for positive and negative prediction errors to happy faces, were predicted by personality traits besides extraversion. To do so, we performed hierarchical multiple linear regressions to examine whether the addition of neuroticism (M = 24.07, SD = 9.10, range = 3-46,

Cronbach's α = .89), openness (*M* = 29.71, *SD* = 6.08, range = 16-42, Cronbach's α = .70), agreeableness (*M* = 33.33, *SD* = 6.57, range = 13-46, Cronbach's α = .78), and conscientiousness (*M* = 31.59, *SD* = 7.64, range = 0-46, Cronbach's α = .83) scores improved prediction of the various dependent variables relative to a model including only extraversion, differential d' index for happy faces, and differential reaction time index for happy faces as predictors. We adjusted the alpha level of significance α to correct for multiple testing using false discovery rate (FDR) with the following formula (Benjamini & Hochberg, 1995):

$$\alpha = \frac{i}{m} \cdot Q$$

where i corresponds to the individual p value's rank (in an ascending order), m is the total number of tests (i.e., the sum of the number of predictors in each model multiplied by the number of dependent variables), and Q is the false discovery rate (here Q = .05). Results of these analyses are displayed in Table S2. Altogether, the inclusion of participants' neuroticism, openness, agreeableness, and conscientiousness scores did not statistically significantly improve prediction of the conditioned response, as well as the estimated learning rates for negative prediction errors, to happy faces (all *F*s < 2.24, all *p*s > .07). Neuroticism was negatively associated with the conditioned response to happy faces during early acquisition, *b* = -0.005, 95% CI [-0.010, -0.0004], β = -.224, *t*(99) = -2.15, *p* = .034; however, this association did not survive correction of the significance level for multiple comparisons using FDR (α = 4/40 · .05 = .005). By contrast, the inclusion of these predictors improved prediction of the estimated learning rates for positive prediction errors, *F*(4, 97) = 2.52, *p* = .046, and explained an additional 9.03% of the variation thereof. Neuroticism negatively predicted the estimated learning rates for positive prediction errors to happy faces (see Figure S4), *b* = -0.011, 95% CI [-0.019, -0.004], β = -.307, *t*(97) = -2.98, *p* = .0037; this association remaining statistically significant after correcting for multiple testing with FDR (α = 3/40 · .05 = . 0038). This exploratory result suggests that happy faces were associated

with lower excitatory learning rates in participants high in neuroticism than in those lower in this trait. None of the other personality traits were found to be statistically significant predictors of the conditioned response to happy faces during early acquisition or extinction and of the learning rates for positive and negative prediction errors to happy faces, even without correcting for multiple comparisons (all *p*s > .05).

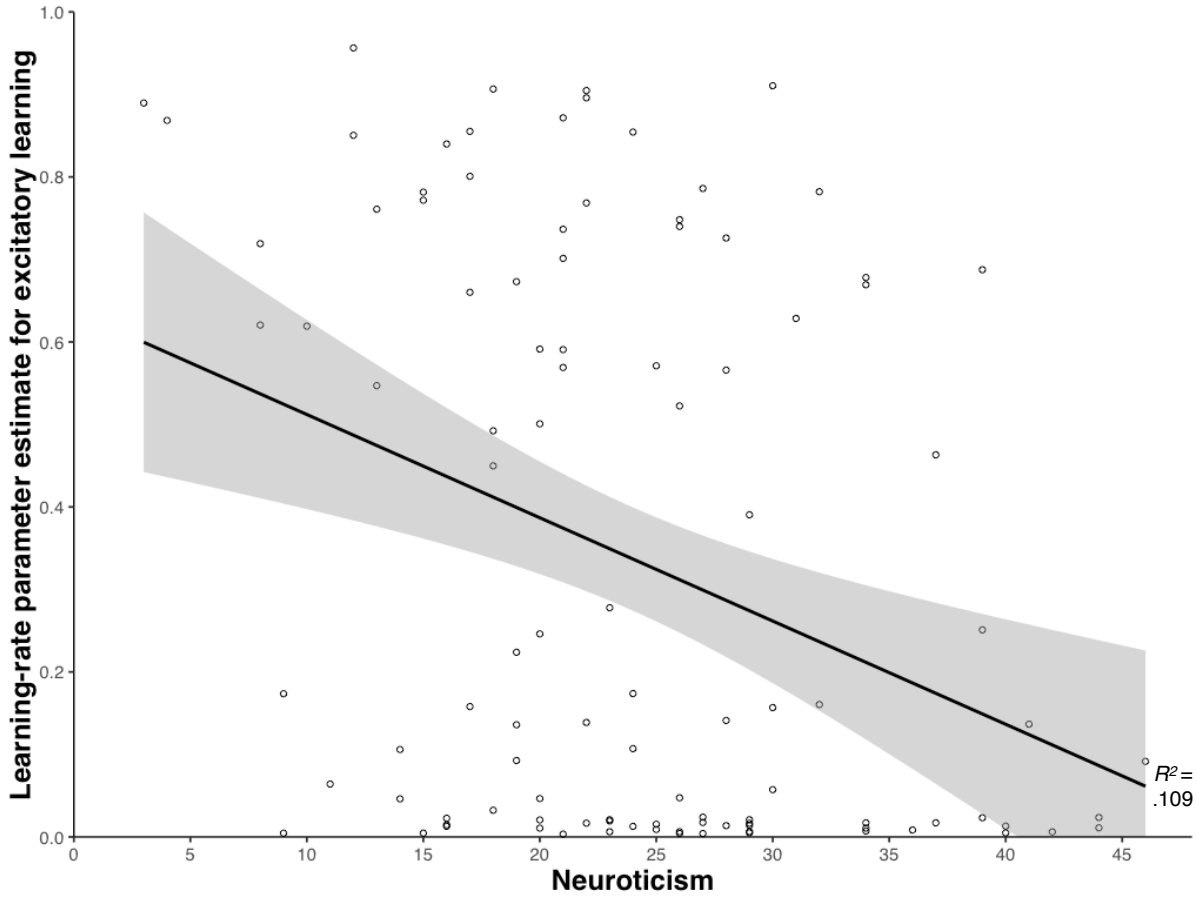

Figure S4. Relationship between neuroticism and the learning rates for positive prediction errors (excitatory learning) for happy faces. The line represents the fitted regression line using least squares estimation and 95% confidence interval.

We conducted additional exploratory analyses to investigate whether personality traits and differential d' and reaction time indices derived from the GNAT were related to the conditioned response and the learning rates to angry and neutral faces. We ran multiple linear regressions with extraversion, neuroticism, openness, agreeableness, conscientiousness,

differential d' index, and differential reaction time index as predictors of the conditioned response during (a) early acquisition and (b) extinction, and of the learning-rate estimates for (c) positive and (d) negative prediction errors to both angry and neutral faces. Results of the analyses for angry faces and neutral faces are shown in Table S3 and Table S4, respectively. Conscientiousness was positively associated with the conditioned response to neutral faces during extinction, $b = 0.003$, 95% CI [0.00005, 0.006], $\beta = .196$, $t(99) = 2.02$, $p = .047$, and with the inhibitory learning rate to neutral faces, $b = 0.007$, 95% CI [0.00008, 0.015], $\beta = .198$, $t(97) = 2.01$, $p = .048$; however, these relationships were no longer statistically significant after adjusting the significance level for multiple comparisons using FDR ($\alpha = 1/28 \cdot .05 = .0018$ and $\alpha = 2/28 \cdot .05 = .0036$, respectively). No other statistically significant relationship was observed between the various predictors and the dependent variables, even without correcting for multiple testing (all ps > .06).

Table S2

Results for the exploratory hierarchical multiple linear regression analyses for happy faces

	Conditioned response to happy faces during early acquisition ($N = 107$)					Conditioned response to happy faces during extinction ($N = 107$)					Estimated excitatory learning rate to happy faces ($N = 105$)					Estimated inhibitory learning rate to happy faces ($N = 105$)				
	b	*SE*	*β*	*t*	*p*	*b*	*SE*	*β*	*t*	*p*	*b*	*SE*	*β*	*t*	*p*	*b*	*SE*	*β*	*t*	*p*
Model 1	$R^2 = .015$					$R^2 = .131$					$R^2 = .041$					$R^2 = .047$				
Intercept	0.073	0.106		0.69	.494	0.027	0.087		0.31	.759	0.069	0.169		0.41	.685	0.446	0.150		2.97	.004
Extraversion	0.003	0.004	.085	0.87	.388	0.002	0.003	.046	0.50	.621	0.009	0.006	.146	1.49	.140	-0.002	0.005	-.031	-0.32	.750
Differential d' index	-0.005	0.039	-.013	-0.13	.896	-0.028	0.032	-.082	-0.87	.386	0.083	0.062	.133	1.33	.187	0.076	0.055	.137	1.37	.173
Differential reaction time index	-0.001	0.001	-.096	-0.96	.341	0.002	0.0005	.360***	3.83	< .001	-0.000	0.001	-.019	-0.19	.852	-0.002	0.001	-.195	-1.95	.054
Model 2	$R^2 = .097$, $\Delta R^2 = .082$, $F(4, 99) = 2.24$, $p = .070$					$R^2 = .149$, $\Delta R^2 = .019$, $F(4, 99) = 0.55$, $p = .699$					$R^2 = .131$, $\Delta R^2 = .090$, $F(4, 97) = 2.52$, $p = .046$					$R^2 = .058$, $\Delta R^2 = .011$, $F(4, 97) = 0.28$, $p = .890$				
Intercept	0.080	0.222		0.36	.720	0.031	0.188		0.17	.868	0.270	0.350		0.77	.443	0.302	0.324		0.93	.355
Extraversion	0.002	0.004	.055	0.51	.612	-0.000	0.003	-.008	-0.08	.936	0.002	0.006	.040	0.37	.710	-0.003	0.006	-.049	-0.44	.695
Differential d' index	-0.023	0.039	-.059	-0.59	.558	-0.033	0.033	-.097	-1.01	.317	0.059	0.062	.095	0.95	.343	0.070	0.058	.125	1.21	.230

Differential reaction time index	-0.000	0.001	-.082	-0.83	.409	0.002	0.0005	.359**	3.74	< .001	-0.000	0.001	-.009	-0.09	.927	-0.002	0.001	-.202	-1.98	.051
Neuroticism	-0.005	0.002	-.224	-2.15	.034	-0.003	0.002	-.127	-1.26	.211	-0.011	0.004	-.307*	-2.98	.004	0.001	0.004	.038	0.35	.728
Openness	0.006	0.003	.165	1.70	.093	0.002	0.003	.055	0.59	.559	0.006	0.005	.100	1.03	.303	-0.002	0.005	-.039	-0.39	.699
Agreeableness	-0.003	0.003	-.106	-1.02	.312	0.001	0.003	.042	0.41	.681	0.002	0.005	.031	0.30	.767	0.003	0.005	.060	0.56	.580
Conscien-tiousness	0.003	0.003	.110	1.10	.273	0.001	0.002	.025	0.26	.797	0.001	0.004	.028	0.29	.775	0.003	0.004	.086	0.83	.407

Note. ***$p < .001$, **$p < .01$, *$p < .05$ (FDR-corrected).

Table S3

Results for the exploratory multiple linear regression analyses for angry faces

	Conditioned response to angry faces during early acquisition ($N = 107$)					Conditioned response to angry faces during extinction ($N = 107$)					Estimated excitatory learning rate to angry faces ($N = 105$)					Estimated inhibitory learning rate to angry faces ($N = 105$)				
	b	*SE*	*β*	*t* (99)	*p*	*b*	*SE*	*β*	*t* (99)	*p*	*b*	*SE*	*β*	*t* (97)	*p*	*b*	*SE*	*β*	*t* (97)	*p*
Intercept	-0.170	0.245		-0.69	.490	0.118	0.181		0.65	.518	-0.171	0.341		-0.50	.617	0.533	0.273		1.95	.054
Extraversion	-0.000	0.004	-.008	-0.07	.943	-0.004	0.003	-.134	-1.22	.225	0.010	0.006	.192	1.77	.080	-0.005	0.005	-.113	-1.00	.320
Differential d' index	-0.017	0.051	-.036	-0.34	.733	0.000	0.037	.000	0.00	.999	-0.018	0.070	-.026	-0.25	.802	-0.012	0.056	-.023	-0.22	.829
Differential reaction time index	0.000	0.001	.002	0.02	.983	-0.001	0.001	-.132	-1.24	.219	-0.000	0.001	-.022	-0.21	.835	-0.000	0.001	-.055	-0.50	.618
Neuroticism	-0.003	0.003	-.109	-1.01	.317	-0.001	0.002	-.081	-0.75	.454	-0.003	0.004	-.098	-0.92	.361	-0.002	0.003	-.066	-0.60	.551
Openness	0.003	0.004	.081	0.80	.426	-0.002	0.003	-.081	-0.79	.430	-0.001	0.005	-.026	-0.26	.798	0.001	0.004	.023	0.21	.830
Agreeableness	0.005	0.004	.150	1.40	.166	0.002	0.003	.093	0.87	.389	0.003	0.005	.065	0.61	.542	-0.002	0.004	-.062	-0.57	.572
Conscien-tiousness	0.004	0.003	.142	1.36	.178	0.003	0.002	.146	1.40	.164	0.005	0.004	.117	1.12	.266	0.001	0.003	.030	0.28	.780
R^2			.061					.066					.098					.027		

Table S4

Results for the exploratory multiple linear regression analyses for neutral faces

	Conditioned response to neutral faces during early acquisition ($N = 107$)					Conditioned response to neutral faces during extinction ($N = 107$)					Estimated excitatory learning rate to neutral faces ($N = 105$)					Estimated inhibitory learning rate to neutral faces ($N = 105$)				
	b	*SE*	*β*	*t* (99)	*p*	*b*	*SE*	*β*	*t* (99)	*p*	*b*	*SE*	*β*	*t* (97)	*p*	*b*	*SE*	*β*	*t* (97)	*p*
Intercept	-0.258	0.203		-1.27	.206	-0.089	0.123		-0.72	.471	-0.104	0.319		-0.33	.745	0.357	0.296		1.21	.231
Extraversion	0.005	0.004	.144	1.31	.193	0.000	0.002	.001	0.01	.993	0.003	0.006	.053	0.47	.640	-0.005	0.005	-.108	-0.99	.325
Differential d' index	-0.025	0.044	-.057	-0.58	.565	0.025	0.026	.093	0.96	.342	0.017	0.069	.025	0.24	.807	0.028	0.064	.044	0.45	.656
Differential reaction time index	0.000	0.001	.010	0.10	.924	0.000	0.0003	.027	0.27	.788	0.000	0.001	.054	0.53	.598	-0.000	0.001	-.038	-0.38	.703
Neuroticism	0.000	0.002	.008	0.07	.942	-0.001	0.001	-.040	-0.39	.699	0.000	0.004	.011	0.11	.916	-0.002	0.003	-.055	-0.52	.603
Openness	0.000	0.003	.007	0.07	.944	-0.003	0.002	-.134	-1.38	.171	-0.001	0.005	-.018	-0.17	.862	0.008	0.005	.169	1.71	.090
Agreeableness	0.005	0.003	.181	1.72	.089	0.003	0.002	.184	1.78	.078	0.009	0.005	.200	1.85	.067	-0.005	0.005	-.119	-1.13	.259
Conscien-tiousness	-0.000	0.003	-.003	-0.03	.978	0.003	0.002	.196	2.02	.047	-0.000	0.004	-.003	-0.03	.976	0.007	0.004	.198	2.01	.048
R^2			.075					.106					.051					.099		

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*, 289-300. http://dx.doi.org/10.2307/2346101

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. http://dx.doi.org/10.1111/ejn.12094

Costa, P. T. Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

de Berker, A. O., Tirole, M., Rutledge, R. B., Cross, G. F., Dolan, R. J., & Bestmann, S. (2016). Acute stress selectively impairs learning to act. *Scientific Reports, 6*, 29816. http://dx.doi.org/10.1038/srep29816

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

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

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

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. http://dx.doi.org/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. http://dx.doi.org/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. *The Journal of Neuroscience, 32*, 551-562. http://dx.doi.org/10.1523/JNEUROSCI.5498-10.2012

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

Nosek, B. A., & Banaji, M. R. (2001). The Go/No-go Association Task. *Social Cognition, 19*, 625-666. http://dx.doi.org/10.1521/soco.19.6.625.20886

Pauli, W. M., Larsen, T., Collette, S., Tyszka, J. M., Seymour, B., & O'Doherty, J. P. (2015). Distinct contributions of ventromedial and dorsolateral subregions of the human substantia nigra to appetitive and aversive learning. *The Journal of Neuroscience, 35*, 14220-14233. http://dx.doi.org/10.1523/JNEUROSCI.2277-15.2015

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. http://dx.doi.org/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. http://dx.doi.org/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.

Rolland, J. P., Parker, W. D., & Strumpf, H. (1998). A psychometric examination of the French translations of the NEO-PI-R and NEO-FFI. *Journal of Personality Assessment, 71*, 269-291. http://dx.doi.org/10.1207/s15327752jpa7102_13

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

Stussi, Y., Pourtois, G., & Sander, D. (2018). Enhanced Pavlovian aversive conditioning to positive emotional stimuli. *Journal of Experimental Psychology: General, 147*, 905-923. http://dx.doi.org/10.1037/xge0000424

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