Research

Evidence for a minimal role of stimulus awareness in reversal of threat learning

Philipp Homan,[1] H. Lee Lau,[2,3,4] Ifat Levy,[5,6,7] Candace M. Raio,[8] Dominik R. Bach,[9,10] David Carmel,[11,12] and Daniela Schiller[2,3,4,12]

[1]*Psychiatric University Hospital Zurich, University of Zurich, 8032 Zurich, Switzerland;* [2]*Department of Psychiatry,* [3]*Department of Neuroscience,* [4]*Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA;* [5]*Department of Comparative Medicine,* [6]*Department of Neuroscience,* [7]*Department of Psychology, Yale University, New Haven, Connecticut 06520, USA;* [8]*New York University Grossman School of Medicine, New York, New York 10016, USA;* [9]*Computational Psychiatry Research, Department of Psychiatry, Psychotherapy, and Psychosomatics, University of Zurich, 8032 Zurich, Switzerland;* [10]*Wellcome Centre for Human Neuroimaging, London WC1N 3BG, United Kingdom;* [11]*School of Psychology, Victoria University of Wellington, Kelburn Parade, Wellington 6012, New Zealand*

In an ever-changing environment, survival depends on learning which stimuli represent threat, and also on updating such associations when circumstances shift. It has been claimed that humans can acquire physiological responses to threat-associated stimuli even when they are unaware of them, but the role of awareness in updating threat contingencies remains unknown. This complex process—generating novel responses while suppressing learned ones—relies on distinct neural mechanisms from initial learning, and has only been shown with awareness. Can it occur unconsciously? Here, we present evidence that threat reversal may not require awareness. Participants underwent classical threat conditioning to visual stimuli that were suppressed from awareness. One of two images was paired with an electric shock; halfway through the experiment, contingencies were reversed and the shock was paired with the other image. Despite variations in suppression across participants, we found that physiological responses reflected changes in stimulus-threat pairings independently of stimulus awareness. These findings suggest that unconscious affective processing may be sufficiently flexible to adapt to changing circumstances.

[Supplemental material is available for this article.]

Flexible responses to environmental threats are essential for adaptive behavior. Cues that predict threat constantly change—new threats may arise while old ones cease to pose a risk. When consciously perceiving such cues, we are able to flexibly update and shift threat responses from one cue to another (Morris and Dolan 2004; Schiller et al. 2008; Fleming et al. 2012). But can we update our reaction to stimuli that predict danger when we are not aware of them?

There is some evidence that threat-conditioned stimuli that are perceived without awareness can still elicit defensive physiological reactions (Ohman and Soares 1994; Morris et al. 1998; Whalen et al. 1998; Critchley et al. 2002). Additionally, although current evidence is inconsistent and controversial (Mertens and Engelhard 2020), there have also been reports that new threat associations can be formed through classical conditioning even without any awareness of the conditioned stimuli (Katkin et al. 2001; Manns et al. 2002; Raio et al. 2012), and that such unconscious learning correlates negatively with anxiety (Raio et al. 2012).

Updating threat associations when contingencies change, however, is an entirely different matter: It involves a complex process of creating novel responses while simultaneously suppressing acquired ones. To date, such updating has only been shown in humans who were aware of the stimuli (Schiller et al. 2008), and in animals under conditions where stimuli were fully available for perceptual processing (Izquierdo et al. 2017); these studies have shown, furthermore, that the neural substrates of threat updating differ from those of the initial learning. It is thus unknown whether the sophisticated reevaluation involved in such affective flexibility requires awareness, or can be accomplished without it. Here we show that it can, and furthermore, that stimulus awareness does not seem to play a substantial role in such affective flexibility.

To examine this, we used the reversal paradigm, a laboratory model that requires flexible updating of threat contingencies (Schiller et al. 2008). In an initial acquisition phase, participants encounter two conditioned stimuli (CSs) and learn that only one of them predicts an electric shock. Halfway through the experiment, with no warning, these contingencies flip, initiating the reversal phase: Participants must flexibly learn that the formerly safe CS now predicts the shock and that the old one no longer does. Appropriate response reversal requires a sophisticated form of updating (Costa et al. 2014), in that one must learn to respond to a cue that now predicts threat while simultaneously inhibiting responses to the previously threatening cue that is now safe. Although it is not necessary (Schiller et al. 2008), explicit instruction can also lead to reversal (Atlas et al. 2016), indicating the involvement of high-level functions and brain regions (Atlas 2019) that may be independent of the more automatic processes that underlie classic conditioning.

[12]**These authors contributed equally to this work.**
Corresponding author: daniela.schiller@mssm.edu; david.carmel@vuw.ac.nz
Article is online at http://www.learnmem.org/cgi/doi/10.1101/lm.050997.119.

To assess learning, participants' physiological arousal is recorded throughout the experiment, typically by measuring their skin conductance responses (SCRs); in the present study, we used a computational approach that uses SCRs to generate estimates of sudomotor nerve activity (SNA; the neural driver of sweat-gland activity that produces SCRs). This approach has been shown to have better sensitivity than traditional SCR peak-scoring methods for discriminating aversive and neutral stimuli, both naturally occurring and fear-conditioned (see "Model based analysis" in the Materials and Methods for further details; Bach et al. 2009, 2010; Bach and Friston 2013).

To see whether reversal of conditioned threat requires awareness, we had a large group of participants ($N = 86$) undergo reversal learning with the CSs suppressed from awareness by continuous flash suppression (CFS), a technique commonly used to examine unconscious perception (Tsuchiya and Koch 2005; Carmel et al. 2010; Stein et al. 2011; Raio et al. 2012): The CSs were visual images presented monocularly, while the other eye was shown a high-contrast, dynamic image (the CFS mask) at the corresponding retinal location (See Fig. 1 for a description of the design and procedure).

CFS can suppress images from awareness for several seconds. However, it is also known that its effectiveness may vary across trials and individuals, and the suppressed stimulus may "break through" the suppression (Gayet and Stein 2017). Over the last decade, a growing body of work has raised concerns that the standard approach—removing from analysis data (participants and trials) in which breakthrough had occurred—may bias the findings (see the Supplemental Material for further details of these issues; Stein and Sterzer 2014; Shanks 2016). Here, we adopt a number of methodological approaches to ensure our results are robust to these potential concerns. In the interest of clarity, we will now introduce these approaches briefly; detailed explanations are provided in the Materials and Methods section.

Specifically, we remove no data and instead incorporate individual levels of reported stimulus awareness, as well as response patterns that might reflect residual awareness, into a regression model accounting for physiological responses. The model also adjusts for baseline anxiety (which, as mentioned above, has been previously shown to correlate with unconscious learning) (Raio et al. 2012). Additionally, we use a Bayesian approach to establish that a model in which participants were updating their learning provides a better account for the findings than models in which they were simply (and independently of the stimulus) predicting the probability of a shock on the next trial (Wiens et al. 2003). Finally, if we found no learning or reversal under CFS, this may be simply due to an ineffective procedure; as a sanity-check—to rule this out and verify that our procedure is able to induce reversal learning when participants are aware of the stimuli—we ran a no-CFS group ($N = 12$) (see /"Participants" in the Materials and Methods for details on sample size determination), in which participants also viewed the CSs monocularly (as the CFS group did), but were aware of them as no CFS masks were presented to their other eye.

We hypothesized that physiological responses to threat can be flexibly reversed without perceptual awareness. As detailed below, we find that reversal indeed occurs independently of CS awareness, and that there is evidence for the reversal of threat learning even in its complete absence.

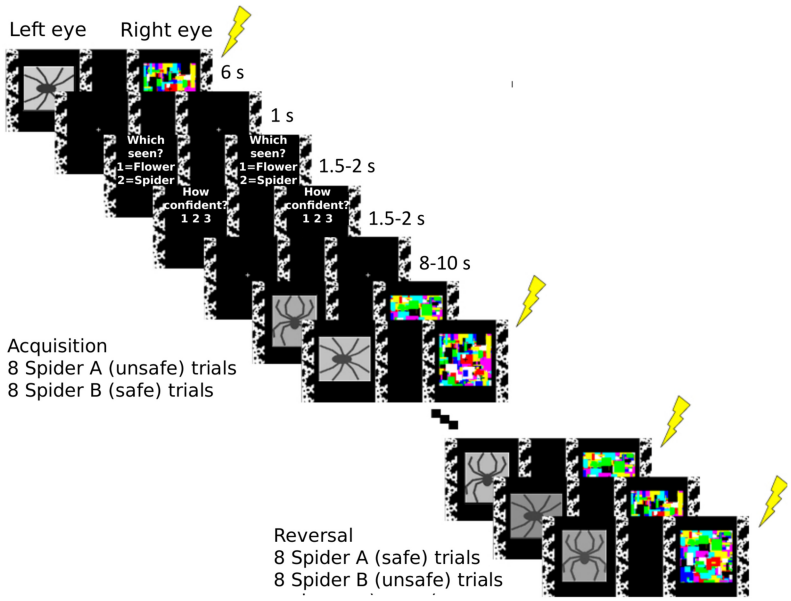

Figure 1. Schematic description of experimental design and procedure. In each trial of the acquisition phase, participants were presented with one of two stimuli (schematic pictures of spiders, presented monocularly for 6 sec and suppressed from awareness by a CFS mask shown to the other eye). One image (spider A) always terminated with a mild electric shock to the wrist, whereas the other (spider B) never did. Halfway through the experiment, with no warning, the contingencies flipped and the reversal phase began: The formerly safe stimulus (spider B) now predicted the shock, and the old threat-associated one (spider A) was now safe. Each spider was shown eight times in each phase. Trial order was pseudorandomized (see the Materials and Methods) and spider identity (A or B) was counterbalanced across participants. To assess the success of the awareness manipulation, participants answered the questions "Which seen?" (1 = flower, 2 = spider) and "How confident?" (1 = guess to 3 = sure), presented binocularly (1.5–2 sec each), beginning 1 sec after the offset of every CS, and followed by an 8- to 10-sec intertrial interval (the questions are only shown here for the first depicted trial, but were repeated in all trials). Participants who underwent the same procedure without CFS were shown identical CSs, but the CFS mask was absent.

Results

Overall assessment of physiological reversal learning

To assess the physiological arousal evoked by CSs, we used a model-based approach (Bach et al. 2010) to estimate the amplitude of anticipatory sudomotor nerve activity (SNA) from skin conductance data recorded during stimulus presentation. A variational Bayes approximation was used to invert a forward model that describes how hidden SNA translates into observable SCRs (see the Materials and Methods). Previous work has shown that this approach is more sensitive than conventional SCR peak scoring analysis (Bach et al. 2010; Bach 2014; Staib et al. 2015). Figure 2A shows the time course of evoked SNA to spiders A and B, separately for the CFS and no-CFS groups. In both groups, responses to spider A relative to spider B were larger during the acquisition phase and smaller during the reversal phase [CFS acquisition: $\beta = 0.14$, $t(341.88) = 3.02$, $P = 0.003$; CFS reversal: $\beta = 0.13$, $t(341.88) = 2.82$, $P = 0.005$; no-CFS acquisition: $\beta = 1.06$, $t(201.15) = 4.59$, $P < 0.001$; no-CFS reversal: $\beta = 0.44$, $t(341.88) = 3.62$, $P < 0.001$].

To quantify the magnitude of physiological reversal learning, we calculated a

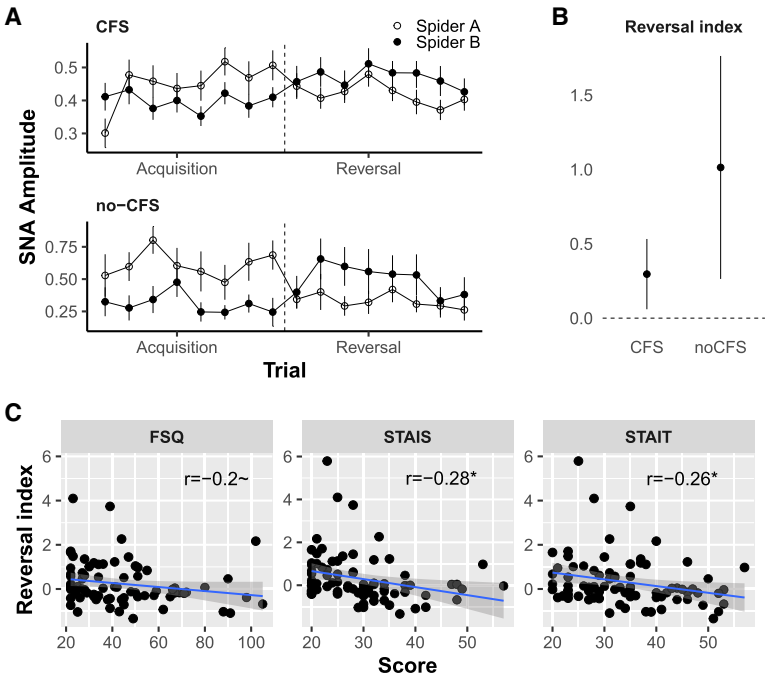

Figure 2. Physiological reversal learning. (*A*) Time courses reveal reversal of threat responses with and without continuous flash suppression. Data points represent trial-wise mean responses to spider A (the CS+ during acquisition) and spider B (the CS− during acquisition). Both groups showed reversal learning, as indicated by greater responses to spider A during the acquisition phase and greater responses to Spider B during the reversal phase. Error bars represent standard errors. (*B*) Mean reversal learning index for each group. Error bars represent 95% confidence intervals, indicating that the interaction of stage and stimulus and thus the magnitude of reversal learning in both groups was significantly greater than zero. (*C*) Heightened anxiety is associated with impaired reversal learning under CFS. A negative correlation between baseline anxiety measures and the strength of threat reversal learning is evident for state and trait anxiety. Blue lines show linear fits of each score to the reversal index, and ribbons around lines indicate bootstrapped 95% confidence intervals around the estimate. Note that the participant with the highest reversal index provided data for the STAIS and STAIT, but not FSQ. (STAIS/STAIT) state/trait anxiety subscale of the Spielberger State-Trait Anxiety Inventory, (FSQ) Fear of Spider Questionnaire, (∼) $P<0.1$, (*) $P<0.05$.

reversal learning index for each participant (see the Materials and Methods). The reversal learning index was positive and significantly greater than zero for both the CFS and no-CFS groups (Fig. 2B) as evidenced by a linear mixed model with SNA as the dependent variable, which revealed a significant interaction of stage and spider in both groups [CFS: $\beta=0.27$, $t(2839)=4.23$, $P<0.001$; no-CFS: $\beta=1.23$, $t(2839)=7.29$, $P<0.001$]. Note that a significant interaction of stage and spider is formally equivalent to a significant reversal learning index; finding a significant interaction for each group (separately) means each group had a significant reversal learning index. On its own, however, this simply reveals a difference in the comparative magnitude of responses to the two CSs across the two halves of the experiment; follow-up tests show that this difference is indeed due to reversal: Spider A evoked greater responses than spider B in the acquisition phase [CFS: $t(341.9)=3.0$, $P=0.003$; no-CFS: $t(201.1)=4.6$, $P<0.001$] and the pattern was reversed in the reversal phase [CFS: $t(341.9)=2.8$, $P=0.005$; no-CFS: $t(341.9)=3.6$, $P=0.0003$]. These results indicate that reversal learning was evident in both groups. Although Figure 2 suggests that it was more pronounced in the no-CFS group, we note that this difference did not reach statistical significance in a Welch two sample *t*-test (accounting for unequal variances) of the reversal index between groups [$t(13.28)=-1.79$, $P=0.097$]. A group difference would also not be straightforwardly interpretable because, as addressed in detail below, suppression from awareness was very heterogenous in the CFS group.

As previous work has found a negative association between anxiety and threat acquisition with and without awareness (Raio et al. 2012), we also calculated correlations between the CFS group's baseline anxiety measures (STAIT, STAIS, and FSQ) and the reversal learning index. Overall, reversal learning decreased significantly with increasing levels of state and trait anxiety, and to a lesser but nonsignificant extent for spider phobia (Fig. 2C).

Reversal learning and perceptual awareness

The CFS manipulation reduced awareness of the CSs; as expected, however, it was differentially effective in doing so across participants, precluding an overall conclusion that all learning under CFS happened nonconsciously. The CFS group showed significantly lower accuracy in response to the "which seen?" question ($M=0.46$, $SD=0.29$) compared with the no-CFS group [$M=0.86$, $SD=0.16$; $t(22.77)=-7.24$, $P<0.001$], and accuracy in the CFS group was not significantly different from the 50% random-response level [$t(85)=-1.21$, $P=0.229$]. The CFS group also showed lower confidence ($M=1.73$, $SD=0.65$) than the no-CFS group [$M=2.83$, $SD=0.08$; $t(95.38)=-15.05$, $P<0.001$].

However, group differences in accuracy and confidence, and even random-level response accuracy, are not sufficient to establish an absence of perceptual awareness in the CFS group. Notably, average confidence of correct responses in this group was low but significantly greater than the minimum value of 1 [$t(77)=10.79$, $P<0.001$], suggesting that at least some participants were aware of some of the CSs; learning might thus have arisen from a subset of trials and/or participants where such awareness occurred. To address this, we quantified CS awareness by calculating an awareness index for each participant, ranging in possible values from 0 for no awareness to 1 for full awareness (see the Materials and Methods). Although the awareness index of the CFS group ($M=0.28$, $SD=0.34$) was significantly lower than the no-CFS group's [$M=0.92$, $SD=0.18$; $t(23.93)=-10.19$, $P<0.001$], it was still significantly higher than zero [$t(85)=7.59$, $P<0.001$], and was also higher for reinforced trials compared with nonreinforced trials [$M_{difference}=0.04$, $SD=0.14$; $t(85)=2.51$, $P=0.014$], suggesting that residual awareness was higher for trials with a shock. Note that we did not see a significant association between baseline state or trait anxiety (indexed by STAIS and STAIT scores) and the awareness index [STAIS: $\beta=0.04$, $t(78)=0.39$, $P=0.699$; STAIT: $\beta=0.04$, $t(78)=0.32$, $P=0.752$], indicating that anxiety was not related to CSs breaking through suppression.

Therefore, in order to test our main hypothesis that the reversal of acquired threat responses can be achieved without perceptual awareness, we characterized the quantitative relation between the level of awareness and the magnitude of reversal learning in the

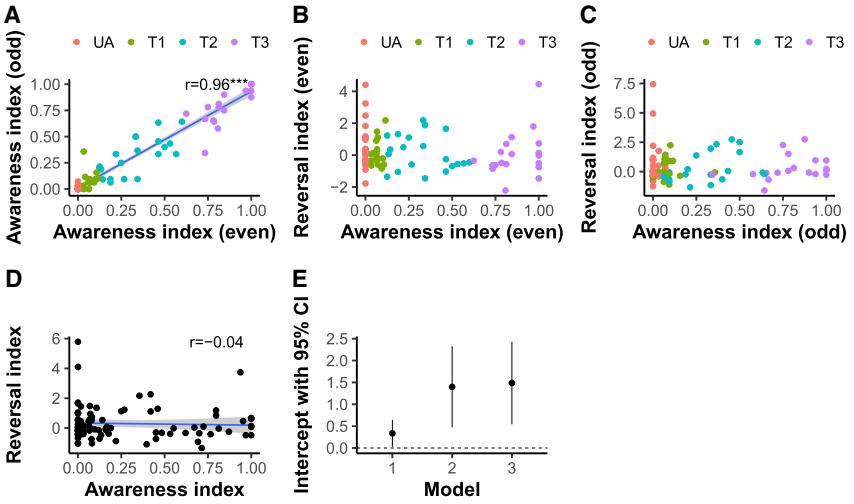

Figure 3. Characterizing the relation between perceptual awareness and reversal learning in the CFS group. (*A*) Correlation between the awareness index of even and odd-numbered trials. Each data point represents an individual participant. The strong positive correlation between these independent measures of awareness demonstrates that individual participants' awareness ratings—even those with extreme values of zero or one—are unlikely to be due to measurement noise. For illustrative purposes, the color scheme marks all participants with an awareness index of 0 in even trials in red ([UA] unaware, $N = 27$) and classifies the rest of the CFS group in three tertiles (T1–T3). Note that some data points overlap. (*B*) Reversal learning plotted against perceptual awareness for individual participants, for data obtained from even-numbered trials. The color scheme is the same as in *A*. (*C*) Reversal learning plotted against perceptual awareness for individual participants, for data obtained from odd-numbered trials. Individual participants are marked with the same color as in the previous panels; the overall distribution of participants is highly similar across panels. (*D*) Reversal learning as a function of perceptual awareness in the CFS group, using data pooled from all trials. The intercept, indicating the magnitude of reversal learning in the absence of awareness, is positive and significantly different from zero. (*E*) Reversal Index intercepts and their 95% confidence intervals in a series of regression models. Model 1 depicts the intercept (the value of the reversal index when the awareness index equals zero) shown in *D*. Model 2 shows the intercept when the regression model includes STAIT scores in addition to the perceptual awareness index. Model 3 regresses the reversal index onto the perceptual awareness index, STAIT and tracking scores. (Excluding the potential outlier in the *top left* corner of *D* weakens significance of the intercept in model 1, $P = 0.07$; the intercepts of models 2 and 3 remain significant after removal of this outlier.) Blue lines show linear fits, and ribbons around lines indicate bootstrapped 95% confidence intervals around the estimate.

CFS group. To control for possible artifacts of regression to the mean (see the Supplemental Material), we followed the recommendation (Shanks 2016) to first examine the correlation between two independent estimates of the awareness index, one calculated from even-numbered trials, the other from odd-numbered trials. Because noise at the measurement level might occasionally yield extreme (i.e., very low or very high) awareness index scores, an association of such randomly extreme scores with reversal learning (specifically, low awareness with intact learning) could be an artifact.

However, it is highly unlikely that across participants, random noise would yield consistent (and similarly extreme) measurements in separate estimates. Due to regression to the mean, if random extreme values occur in one of the two estimates, they are less likely to occur in the other, resulting in a considerable attenuation of any correlation between the two. We found, however, that the two measures were strongly correlated [$r(84) = 0.96$, $P < 0.001$] (Fig. 3A); participants' awareness level in one set of trials was overwhelmingly predictive of their awareness in the other set, confirming their reliability as estimates of awareness.

Next, we examined the association between the awareness index and the reversal learning index, using values of both indices obtained separately from even (Fig. 3B) and odd (Fig. 3C) trials. As the color-coding of Figure 3 shows, the relation between individual participants' awareness and their reversal learning was highly consistent across these separate measurements. In light of this, we pooled the data from all trials and regressed the reversal learning index on the perceptual awareness index (Fig. 3D). The parameter of interest was the intercept, that is, the magnitude of reversal learning at zero perceptual awareness. The intercept was positive and significantly different from zero. Furthermore, the awareness index regressor did not contribute significantly to prediction of reversal learning; importantly, this finding was even stronger in models that accounted for STAIT scores and a binary factor indicating whether participants were tracking the stimuli with their responses (see the Materials and Methods; Fig. 3E; Table 1).

Comparing learning and expectation-based accounts

Well-controlled laboratory-based conditioning procedures require strict constraints that preclude complete randomization of the number and order of different CSs; this comes with a cost: Participants are able to develop expectations with above-chance validity, based on the sequence of trials so far, about the likelihood of a shock on any upcoming trial (Wiens et al. 2003). Even without any awareness of the CSs, a participant should have been able to distinguish two types of trials: reinforced (with shock) and nonreinforced (no-shock). In a study with two CSs and a 100% reinforcement rate like ours, such expectations would correspond to an anticipation based on the experienced pattern of trial-types (shock/no-shock or vice versa), with an increase in shock anticipation after every no-shock trial. The question, therefore, was whether the physiological responses we had measured might simply reflect participants' pattern-based anticipation of shock, rather than learning of the contingencies associated with the CSs.

The simplest way of addressing this question was to examine whether participants might use a trial-sequence heuristic that

Table 1. Regression coefficients for all awareness index models

Model	Predictor	β	SE	*t*	*P*
1	Intercept	0.3	0.2	2.1	0.035
1	Awareness index	−0.1	0.4	−0.4	0.692
2	Intercept	1.4	0.5	3	0.004
2	STAIT	0	0	−2.3	0.024
2	Awareness index	−0.2	0.4	−0.5	0.596
3	Intercept	1.5	0.5	3.1	0.003
3	STAIT	0	0	−2.4	0.021
3	Tracking score	−0.3	0.3	−1	0.318
3	Awareness index	−0.2	0.4	−0.5	0.597

Reversal learning was the dependent variable in all models. Model 1 included an intercept and the perceptual awareness index, model 2 additionally included STAIT scores, and model 3 additionally included STAIT and tracking scores.

would allow them to discriminate spider A from spider B (and to achieve apparent reversal learning) by assuming that a shock was more likely after a no-shock trial. In the SNA data, this would be indicated by better discrimination for alternating trials (where the present trial's stimulus differs from the previous trial's) than nonalternating trials (where the stimuli of the present and previous trial are the same). In a first step, we thus tested whether there was evidence that the interaction of stage and spider was modulated by trial-order effects (i.e., alternating and nonalternating trials). Using a linear mixed model with SNA as the dependent variable and trial number, stage (acquisition or reversal), spider (spider A or spider B), and trial-type (alternating or nonalternating) as predictors, we tested for a three-way interaction of stage, spider, and trial-type. This interaction tested whether the reversal learning effect (the two-way interaction between stage and spider) was modulated by the trial-type. If we found a three-way interaction that was significantly different from zero, this would indicate that the reversal learning effect was different for alternating versus nonalternating trials. However, the interaction was not significantly different from zero [$\beta = -0.13$, $t(2582.49) = -0.86$, $P = 0.391$]. Thus, our data do not provide any support for the idea that the reversal learning effect was influenced by trial-order effects. We did, however (perhaps more convincingly), find clear evidence for reversal learning in nonalternating trials [indicated by a two-way interaction of stage × spider in these trials, $\beta = 0.33$, $t(2574.6) = 4.57$, $P < 0.001$].

To further address the same question, we also used a Bayesian approach to compare the probability of our findings being accounted for by a classic Rescorla–Wagner learning model (Rescorla and Wagner 1972) versus two different trial-sequence expectation models. We hypothesized that successful threat reversal without perceptual awareness should be better explained by the Rescorla–Wagner learning model, compared with a model informed either by trial alternation (where participants simply expect an alternating pattern of shock/no-shock trials) or by pattern-based expectation (in which the expectation of shock on the next trial increased after every nonshock trial, and accounting for consecutive trials of the same type). We used maximum likelihood estimation to assess the log likelihood and calculate the Bayesian information criterion (BIC) of each model (See Materials and Methods for details of each model and calculation of the BIC). A smaller BIC indicates a better model, and BIC values can thus be compared by calculating the difference between them and interpreting the resulting Δ BIC as providing evidence against the higher BIC.

As a validation, we first tested these two models in the no-CFS group, where we expected to find that the Rescorla–Wagner model would fit the data better than the trial alternation model. We found that the data of the no-CFS group was indeed more in line with a Rescorla–Wagner model (BIC: 200.98) than a trial switch model (BIC: 224.28), with the difference (Δ BIC: 23.3) being >10 and thus—by widely accepted convention—large enough (Raftery 1995) to conclude that the Rescorla–Wagner model fit the data significantly better.

For the CFS group, the Rescorla–Wagner model (BIC: 1019.93) also outperformed the trial alternation expectation model (BIC: 1098.86), with the difference (Δ BIC: 78.93) >10, suggesting that the evidence against the trial switch model is very strong (Raftery 1995). (Repeating this comparison for just the participants with zero mean awareness confirmed the lower BIC for the Rescorla–Wagner model [BIC: 263.99] compared with the pattern-based expectation model [BIC: 300.34], with the difference again >10 [Δ BIC: 36.36] [see also Supplemental Figure S2].) The pattern-based expectation model (BIC: 1390.48) was even less successful at accounting for the data than the trial alternation model.

Finally, an extended Rescorla–Wagner model assuming different learning rates for acquisition and reversal (BIC: 1291.84) did not fit the data better than the simpler one. This model comparison confirms that a classical Rescorla–Wagner learning model fits our data better than alternative expectation-based models.

Discussion

These results indicate that participants updated their defensive physiological responses independently of their awareness of threat-related cues. The findings therefore suggest that the complex process of threat reversal—shifting reactions from a stimulus that no longer predicts danger to one that now does—may be accomplished independently of perceptual awareness, and thus that dissociable processes might underlie affective flexibility and conscious processing (Lau and Rosenthal 2011). Conversely, the negative correlation between reversal learning and anxiety suggests that the various impairments caused by anxiety are not limited to the systems underlying conscious processes.

The present findings add to the growing literature on threat processing outside awareness. Several previous studies have reported evidence that new threat associations can be formed without perceptual awareness of the conditioned stimuli (Katkin et al. 2001; Manns et al. 2002; Raio et al. 2012), but a recent meta-analysis (Mertens and Engelhard 2020) has indicated that such reports often suffer from various methodological issues, and furthermore, found evidence for publication bias. Of course, no single study can conclusively resolve the discussion on a topic that presents multiple difficulties; we believe, however, that our attempt to address methodological issues through rigorous testing and analyses provides a useful addition to the literature by examining both initial conditioning and reversal of threat responses. Previous studies have pointed out the limitations of using accuracy and confidence measures to assess perceptual awareness, and suggested remedies including the calculation of metacognitive sensitivity measures (Fleming and Lau 2014), Bayesian statistics (Dienes 2015), or parametric variation of the experimental manipulation (Schmidt 2015). The present study addresses an issue not covered in previous discussions, by showing that a trial-wise analysis may reveal hints for incomplete suppression that analyses relying on average measures might easily miss. Future studies that rely on forced-choice questions for awareness assessment should thus examine response patterns across trials in addition to collecting aggregate measures.

Notably, a previous study (Raio et al. 2012) that used CFS to investigate acquisition of threat responses without awareness of the stimuli found that such acquisition can occur, but is rapidly forgotten. The present study again showed that such acquisition can occur (and, additionally, be reversed), but did not find the same rapid forgetting. The reasons for this are unclear, but we speculate that the difference may be due to specific aspects of the stimuli, design and procedure: Our use of pictures of spiders (rather than the faces used in the previous study) and a 100% (rather than 50%) reinforcement protocol may have altered the temporal characteristics of acquisition. Similarly, the temporal profile of reversal may change if the stimuli and reinforcement regime are different.

The present results add to a growing body of findings distinguishing functions that do and do not require awareness. Such distinctions are important in guiding research into the neural mechanisms of conscious and nonconscious processing. Previous research hints at the mechanism underlying the nonconscious affective flexibility reported here, although it remains to be elucidated: The ability to reverse conditioned responses depends on the integrity of circuitry spanning several neural

regions, particularly the ventromedial prefrontal cortex (vmPFC) and its connections with the amygdala (Morris and Dolan 2004) where threat associations are formed (Roy et al. 2012). Consistent with this, it is known that patients with anxiety disorders often show rigid and inflexible threat responses in conjunction with prefrontal cortex dysfunction (Rauch et al. 2006; Ressler and Mayberg 2007).

Indeed, the real-life settings that people with anxiety disorders find challenging often require the updating and shifting of threat responses. Deficits in affective flexibility may thus explain the threat learning and extinction deficits seen in such disorders (Duits et al. 2015): Compared with healthy controls, patients are less able to distinguish between safe and unsafe stimuli in threat learning (when it is adaptive to do so), and distinguish between them to a greater extent during extinction (when it is nonadaptive). Recent findings of an association between prediction error weighting during reversal learning (with awareness) and the severity of posttraumatic stress disorder (Homan et al. 2019) demonstrate the usefulness of the reversal learning paradigm in studying disorders characterized by impaired threat inhibition (Jovanovic and Norrholm 2011). Our new findings—that baseline anxiety is not correlated with stimulus awareness under CFS, but is negatively correlated with affective flexibility—augment the emerging picture by showing that the association between reversal and anxiety may not depend on awareness.

Materials and Methods

Participants

Ninety-eight healthy participants (mean age = 29.97; range 18–65) were assigned to one of the two groups: reversal learning with CFS (CFS group; $N = 86$, 48 female) or without CFS (no-CFS group; $N = 12$, 5 female). The sample size for the CFS group was based on the strength of the effects found in our previous study (Raio et al. 2012), where effect sizes (Cohen's d) were ~1.6 in early conditioning and 0.5 in late conditioning. Using a conservative effect size estimate of $d = 0.5$, we would have needed a sample size of 43 or 44 to detect this effect with 90% power; because our effect of interest (reversal) required a significant interaction between stage (acquisition vs. reversal) and stimulus (spider A vs. B), we doubled this estimate. For our no-CFS group, we based our sample size on the strong late conditioning effect ($d = 1.7$) in the aware group of our previous study (Raio et al. 2012), as well as previous literature on reversal without suppression—under sufficiently similar conditions to the present no-CFS group (Schiller et al. 2008)—which also suggested that the effect of reversal would be similar to that of the initial conditioning. We used a slightly more conservative effect size estimate of $d = 1.5$ for the present no-CFS group, which required in a sample size of $N = 7$ to detect this effect with 90% power. Because we intended to test for an interaction between stage and stimulus in this group as well, we increased the sample to 12.

Assignment was random until each group reached a size of 12; subsequent participants were assigned to the CFS group. Measures of trait and state anxiety (Spielberger Trait-State Anxiety Inventory [STAIT and STAIS, respectively] [Spielberger 1983]) and spider phobia (Fear of Spider Questionnaire [FSQ] [Szymanski and O'Donohue 1995]) were taken prior to participation and did not differ between the groups (Supplemental Table S1). The experiment was approved by the Institutional Review Board of the Icahn School of Medicine at Mount Sinai. All participants provided written informed consent and were financially compensated for their participation.

Experimental procedures

Participants viewed the stimuli monocularly, through a mirror stereoscope (StereoAids, Australia) placed at a distance of 45 cm from a 17-in Dell monitor. The CSs (schematic low-contrast images of spiders), presented to the left eye only, were suppressed from awareness in the CFS group: While the left eye saw them, the right eye was presented with "Mondrians"—arrays of high-contrast, multicolored, randomly generated rectangles alternating at 10 Hz. Both the CSs and the CFS masks were flanked by identical textured black and white bars, to facilitate stable ocular vergence. The no-CFS group viewed identical CSs (also presented monocularly), but with no Mondrians presented to the other eye.

The experiment consisted of 16 acquisition trials followed by 16 reversal trials. One of two spider images was presented on each trial. The spider images were schematic and had similar low-level features. During acquisition, spider A always terminated with a shock and spider B never did. Reversal occurred halfway through the experiment: Spider B now terminated with a shock and spider A did not. The spider stimuli were presented for 6 sec each in pseudorandomized order. One of four possible trial orders was used for each participant. Orders were generated by imposing specific constraints on the trial order, such that the first trial was always reinforced and no more than two trials of the same type ever occurred consecutively.

Trial order and spider identity were counterbalanced across participants. To assess the effectiveness of the awareness manipulation (44), 1 sec after the offset of every CS participants were shown the question "Which seen?" (1 = flower, 2 = spider; notably, flowers were never shown, meaning the question addressed detection rather than discrimination as it could be answered correctly even with a brief glimpse). This was followed by the question "How confident?" (1 = guess to 3 = sure; participants were instructed to indicate how confident they were of the flower/spider answer they had just given). Both questions were presented binocularly (1.5–2 sec each, during which responses had to be given by pressing number keys on a standard keyboard). The second question was followed by an 8- to 10-sec intertrial interval. We did not ask participants about their awareness of CS–US contingencies.

Psychophysiological stimulation and measurement

Mild electric shocks were delivered using a Grass Medical Instruments SD9 stimulator and stimulating bar electrode attached to the participant's right wrist. Shocks (200 msec; 50 pulse/sec) were delivered at a level determined individually by each participant as "uncomfortable but not painful" (maximum of 60 V), during a work-up procedure prior to the experiment.

Skin conductance responses (SCR) were measured with Ag–AgCl electrodes, filled with standard isotonic NaCl electrolyte gel, and attached to the middle phalanges of the second and third fingers of the left hand. SCR signals were sampled continuously at a rate of 200 Hz, amplified and recorded with a MP150 BIOPAC Systems skin conductance module connected to a PC.

Analysis of physiological responses

Model–based analysis

To quantify the expression of CS–US memory on each trial, we used a model-based approach to quantify Sudomotor Nerve Activity (SNA). Sudomotor nerves are the nerves that control sweat glands, whose activity in turn produces skin conductance responses. Because sweat (and the resulting skin conductance) is an indirect measure of the underlying nerve activity, methods that characterize the nerve activity itself provide a better assessment of the neural processing that led to the observed response. This approach uses a psychophysiological model that describes skin conductance as the convolution of a Gaussian-shaped sudomotor nerve (SN) input with a canonical skin conductance response model (Bach et al. 2010). This model is inverted to yield the most likely SNA amplitude given the data. This approach is conceptually similar to the standard approach used in fMRI dynamic causal modeling analysis (where the BOLD signal is convolved with a canonical hemodynamic response function, and the model is inverted to yield sources of neural activation). Compared with peak-scoring methods for SCR, this approach has been shown to better

discriminate between responses to aversive and neutral stimuli (Bach et al. 2009; Bach and Friston 2013) and to better discriminate between responses to CS+ and CS− in fear conditioning (Bach et al. 2010). This better discrimination implies better accuracy and precision than standard SCR analysis in inferences on the latent CS–US association (Bach et al. 2020).

Specifically, we used a model that describes, for each trial, a CS-related SN burst at some point during CS presentation, for which amplitude, onset latency and duration are estimated from the data; and an additional burst related to the US (or its omission), for which timings are known and only amplitude is estimated. The model also captures spontaneous fluctuations and baseline changes during intertrial-intervals (Bach et al. 2010). This nonlinear model is inverted using a Variational-Bayes algorithm (for further details on the computational aspects of this algorithm, see "Model-based SNA analysis of SCR data" in the Supplemental Material). The SNA estimates were computed using the PsPM software package (version 3.0; http://bachlab.org/pspm; Bach et al. 2010) implemented in MATLAB R2016b (The Mathworks, Inc.). The statistical analyses were conducted with R software (R version 3.6.1 [2019-07-05]; R Core Team 2016) and the libraries lme4 (Bates 2005) and lsmeans (Lenth 2016).

Reversal learning index

An estimate of SNA was obtained for each trial. We expected Spider A to evoke greater SNA than Spider B during the acquisition phase, and Spider B to evoke greater SNA than Spider A during the reversal phase. The strength of reversal learning can thus be quantified by calculating, separately for the acquisition and reversal phases, the difference between the average SNA evoked by each spider. To quantify the degree of reversal (which is formally equivalent to the interaction of phase and stimulus), the reversal learning index was calculated by subtracting the difference between mean SNAs evoked by each spider during reversal from the difference during acquisition (the larger the index, the greater the magnitude of reversal learning):

$$\begin{aligned} &\textit{Reversal learning index} = \Delta \textit{Acquistion} - \Delta\text{Reversal} \\ &\Delta \textit{Acquistion} = [\text{mean(Spider A)} - \text{mean(Spider } B)]_{\textit{Acquistion}} \\ &\Delta\text{Reversal} = [\text{mean(Spider A)} - \text{mean(Spider } B)]_{\text{Reversal}}. \end{aligned} \quad (1)$$

To formally test for group differences in the strength of reversal learning, we computed a linear mixed model using the lme4 library in R. We used the skin conductance response (converted to a model-based measure of sudomotor nerve activity, SNA) as the dependent variable and entered group (CFS, no-CFS), stage (acquisition, reversal), and spider (spider A, spider B) as well as a continuous variable for trial (to account for habituation) as predictors. The random structure of the model included an intercept and slopes for stage and spider.

Assessments of perceptual awareness

Perceptual awareness index

To characterize participants' reported awareness of CSs, each trial was assigned a perceptual awareness score, defined by a combination of detection and confidence responses: Correct answers with a confidence rating of 1 (guess) and incorrect answers irrespective of confidence were assigned an awareness score of 0; correct answers with a confidence rating of 2 (medium) were assigned a score of 0.5, and correct answers with a confidence rating of 3 (high) were assigned an awareness score of 1. A perceptual awareness index was calculated for each participant by averaging awareness scores across all trials.

Stimulus–response association patterns ('tracking')

We also assessed response patterns across trials, to see whether participants were able to track stimuli with their responses, accurately discriminating the images despite not being able to label them. We plotted individual trial-by-trial responses to the question "Which seen?", overlaid on the trial-by-trial presentation of spiders (spider A or spider B) (Supplemental Fig. S1A). We then calculated the number of consecutive "hits," defined as the number of consecutive trials where these two time-courses were either identical or consistently in opposition, suggesting that there was a possible association between the stimulus and the response during those trials. The probability of such consecutive hits occurring by chance alone can be derived as follows:

Let $P = 0.5$ be the probability of a hit, k the number of consecutive hits, n the number of trials left, i the number of consecutive hits already observed; the chance of observing k consecutive hits for the remaining n trials can then be formulated as a recursive problem:

$$f_{p,k}(i, n) = pf_{p,k}(i + 1,\ n - 1) + (1 - p)f_{p,k}(0, n - 1), \quad (2)$$

which can be solved analytically with dynamic programming or recursion. Trivially, $f_{p,k}(k, n) = 1$ for $n \geq 0$ since k consecutive hits have already been observed, and $f_{p,k}\ (i, n) = 0$ for $k - i > n$ since there are not enough trials left to observe k consecutive hits.

For example, assuming we want to know how likely it is to observe $k = 8$ consecutive hits within $n = 32$ trials given $P = 0.5$, i.e., $f0.5, 8\ (0, 32)$, we find that this yields a probability of 0.050.

Alternatively, the probability can be derived by simulation for all possible numbers of consecutive hits within 32 trials (i.e., from 1 to 31). For each possible number, we thus also simulated 10^5 draws of a binomial distribution and calculated the average probability of that number of hits being consecutive. As can be seen in Supplemental Figure S1B, the result for eight consecutive hits (0.04991) was very close to the analytical solution. Fifteen participants showed evidence of tracking the spiders or the shocks with their responses (eight or more consecutive hits); notably, three of these participants appeared to have a perceptual awareness index of zero. We thus adjusted our subsequent analysis with an additional binary covariate, indicating whether participants did or did not show eight or more consecutive hits.

Comparing learning and expectation-based models

The Rescorla–Wagner (RW) model (30) describes how the prediction for each trial is updated according to a prediction error and learning rate:

$$\begin{aligned} V_{n+1}(x_n) &= V_n(x_n) + \alpha\delta_n \\ \delta_n &= r_n - V_n(x_n), \end{aligned} \quad (3)$$

where x_n is the conditioned stimulus on trial n (spider A or spider B), and δ_n is the punishment prediction error that measures the difference between the expected and the actual shock (r_n) on trial n. The learning rate α for the value update is a constant free parameter. The value for the CS not observed on trial n remains unchanged. To derive the best fits for the Rescorla–Wagner model, we assumed that $V0 = 1$, reflecting an assumption that participants expected to get a shock on the first trial. We also used an extended version of the RW model that included an additional weight parameter ρ for the reversal phase to account for a potential change in the learning rate during reversal compared with acquisition. For acquisition, we thus used the classical RW model, and used the extended model for reversal:

$$\begin{aligned} V_{n+1}(x_n) &= V_n(x_n) + \rho\alpha\delta_n \\ \delta_n &= r_n - V_n(x_n). \end{aligned} \quad (4)$$

For the alternative trial-sequence learning model, we assumed that a participant expecting a strict sequence of alternating trial types (shock/no shock or vice versa) would update this expectation according to the actually encountered trial types and a constant learning rate:

$$
\begin{aligned}
V'_{n+1} &= V'_n + \alpha' \delta'_n \\
\delta'_n &= r'_n - V'_n \\
\tau_n &= |(r'_{n-1} - 1)|,
\end{aligned} \tag{5}
$$

where $V'n+1$ is the expected trial type switch at trial $n+1$ (if $V'n+1$ is >0.5, a trial switch is expected), α' is the learning rate, and $\delta'n$ is the prediction error. The prediction error corresponds to the difference between the actual trial type switch for trial n ($r'n$; coded as one for a trial type switch and zero for an equal trial type) and the expectation for trial n. A changing trial type for trial n was tracked by τn, which was one if the preceding trial was zero and zero if the preceding trial type was one. To map these expectations onto expected values, we assumed that

$$
V_{n+1} = \begin{cases} V'_{n+1} \cdot \tau_n (1 - V'_{n+1})(1 - \tau_n), & \text{if } V' > 0.5 \\ V'_{n+1}, & \text{otherwise,} \end{cases} \tag{6}
$$

where the expected value for trial $n+1$ was calculated according to whether a trial type switch was expected ($V'>0.5$) or not.

To account for instances in which two consecutive nonreinforced trials (two consecutive spider B trials in acquisition or Spider A trials in reversal) might impact the expected values, we also tested an extended, pattern-based version of the trial-sequence learning model that included an additional weight parameter, ξ, for the learning rate α' under these circumstances:

$$
V'_{n+1} = \begin{cases} V'_n + \xi \alpha' \delta'_n, & \text{if two consecutive neutral trials} \\ V'_n + \alpha' \delta'_n, & \text{otherwise.} \end{cases} \tag{7}
$$

We performed formal model comparisons using maximum likelihood estimation and nonlinear optimization (implemented with the fmincon function in MATLAB R2016b (The Mathworks, Inc). Using the log likelihood, we calculated the Bayesian Information Criterion (BIC) to compare the two models as follows:

$$
\mathrm{BIC} = \log(n)k - 2 \cdot \log(\hat{L}), \tag{8}
$$

where n is the number of data points, k is the number of regressors, and $\hat{K}$ is the maximized value of the likelihood function. The conventional Rescorla–Wagner model provided the best account of the data (lowest BIC), and the model with the closest BIC was the simple trial alternation model. Supplemental Figure S2 therefore shows the direct comparison between these two models.

Acknowledgments

We thank Patrik Vuilleumier who created and shared the spider stimuli. This work was supported in part through the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai. Funding was provided by National Institute of Mental Health grant MH105515 and a Klingenstein-Simons Fellowship Award in the Neurosciences to D.S., ERC Advanced Grant XSPECT-DLV-692739 to D.C. (Co-I), and Swiss National Science Foundation grant SNF 161077 to P.H.

Author contributions: P.H. carried out the computational modeling and statistical analysis, interpreted the results, and drafted the manuscript. H.L.L. prepared materials, collected the data, and critically revised the manuscript. I.L. contributed to the conception of the study, the computational modeling, and the interpretation of the results, and critically revised the manuscript. C.M.R. contributed to the interpretation of the results and critically revised the manuscript. D.R.B. contributed to the computational modeling and critically revised the manuscript. D.S. conceived, designed, and coordinated the study; contributed to data analysis and interpretation; and critically revised the manuscript. D.C. contributed to the conception of the study, data analysis, and interpretation of results, and drafted the manuscript in its final form. All authors gave final approval for publication.

References

Atlas LY. 2019. How instructions shape aversive learning: higher order knowledge, reversal learning, and the role of the amygdala. *Curr Opin Behav Sci* **26:** 121–129. doi:10.1016/j.cobeha.2018.12.008

Atlas LY, Doll BB, Li J, Daw ND, Phelps EA. 2016. Instructed knowledge shapes feedback driven aversive learning in striatum and orbitofrontal cortex, but not the amygdala. *Elife* **5:** e15192. doi:10.7554/eLife.15192

Bach DR. 2014. A head-to-head comparison of scralyze and ledalab, two model-based methods for skin conductance analysis. *Biol Psychol* **103:** 63–68. doi:10.1016/j.biopsycho.2014.08.006

Bach DR, Friston KJ. 2013. Model-based analysis of skin conductance responses: towards causal models in psychophysiology. *Psychophysiology* **50:** 15–22. doi:10.1111/j.1469-8986.2012.01483.x

Bach DR, Flandin G, Friston KJ, Dolan RJ. 2009. Time-series analysis for rapid event-related skin conductance responses. *J Neurosci Methods* **184:** 224–234. doi:10.1016/j.jneumeth.2009.08.005

Bach DR, Daunizeau J, Friston KJ, Dolan RJ. 2010. Dynamic causal modelling of anticipatory skin conductance responses. *Biol Psychol* **85:** 163–170. doi:10.1016/j.biopsycho.2010.06.007

Bach D, Melinscak F, Fleming S, Voelkle M. 2020. Calibrating the experimental measurement of psychological attributes. *Nat Hum Behav* **4:** 1229–1235. doi:10.1038/s41562-020-00976-8

Bates DM. 2005. Fitting linear mixed models in R. *R News* **5:** 27–30.

Carmel D, Arcaro M, Kastner S, Hasson U. 2010. How to create and use binocular rivalry. *J Vis Exp* **45:** 2030. doi:10.3791/2030

Costa VD, Bradley MM, Lang PJ. 2014. From threat to safety: instructed reversal of defensive reactions. *Psychophysiology* **52:** 325–332. doi:10.1111/psyp.12359

Critchley HD, Mathias CJ, Dolan RJ. 2002. Fear conditioning in humans: the influence of awareness and autonomic arousal on functional neuroanatomy. *Neuron* **33:** 653–663. doi:10.1016/S0896-6273(02)00588-3

Dienes Z. 2015. How Bayesian statistics are needed to determine whether mental states are unconscious. In *Behavioural methods in consciousness research* (ed. Overgaard M), pp. 199–221. Oxford University Press, Oxford.

Duits P, Cath DC, Lissek S, Hox JJ, Hamm AO, Engelhard IM, van den Hout MA, Baas JM. 2015. Updated meta-analysis of classical fear conditioning in the anxiety disorders. *Depress Anxiety* **32:** 239–253. doi:10.1002/da.22353

Fleming SM, Lau HC. 2014. How to measure metacognition. *Front Hum Neurosci* **8:** 443. doi:10.3389/fnhum.2014.00443

Fleming SM, Dolan RJ, Frith CD. 2012. Metacognition: computation, biology and function. *Philos Trans R Soc Lond B Biol Sci* **367:** 1280–1286. doi:10.1098/rstb.2012.0021

Gayet S, Stein T. 2017. Between-subject variability in the breaking continuous flash suppression paradigm: potential causes, consequences, and solutions. *Front Psychol* **8:** 437. doi:10.3389/fpsyg.2017.00437

Homan P, Levy I, Feltham E, Gordon C, Hu J, Li J, Pietrzak RH, Southwick S, Krystal JH, Harpaz-Rotem I, et al. 2019. Neural computations of threat in the aftermath of combat trauma. *Nat Neurosci* **22:** 470–476. doi:10.1038/s41593-018-0315-x

Izquierdo A, Brigman J, Radke A, Rudebeck P, Holmes A. 2017. The neural basis of reversal learning: an updated perspective. *Neuroscience* **345:** 12–26. doi:10.1016/j.neuroscience.2016.03.021

Jovanovic T, Norrholm SD. 2011. Neural mechanisms of impaired fear inhibition in posttraumatic stress disorder. *Front Behav Neurosci* **5:** 44. doi:10.3389/fnbeh.2011.00044

Katkin ES, Wiens S, Ohman A. 2001. Nonconscious fear conditioning, visceral perception, and the development of gut feelings. *Psychol Sci* **12:** 366–370. doi:10.1111/1467-9280.00368

Lau H, Rosenthal D. 2011. Empirical support for higher-order theories of conscious awareness. *Trends Cogn Sci* **15:** 365–373. doi:10.1016/j.tics.2011.05.009

Lenth RV. 2016. Least-squares means: the R package lsmeans. *J Stat Softw* **69:** 1–33. doi:10.18637/jss.v069.i01

Manns JR, Clark RE, Squire LR. 2002. Standard delay eyeblink classical conditioning is independent of awareness. *J Exp Psychol Anim Behav Processes* **28:** 32–37. doi:10.1037/0097-7403.28.1.32

Mertens G, Engelhard IM. 2020. A systematic review and meta-analysis of the evidence for unaware fear conditioning. *Neurosci Biobehav Rev* **108:** 254–268. doi:10.1016/j.neubiorev.2019.11.012

Morris JS, Dolan RJ. 2004. Dissociable amygdala and orbitofrontal responses during reversal fear conditioning. *Neuroimage* **22:** 372–380. doi:10.1016/j.neuroimage.2004.01.012

Morris JS, Ohman A, Dolan RJ. 1998. Conscious and unconscious emotional learning in the human amygdala. *Nature* **393:** 467–470. doi:10.1038/30976

Ohman A, Soares JJ. 1994. 'Unconscious anxiety': phobic responses to masked stimuli. *J Abnorm Psychol* **103:** 231–240. doi:10.1037/0021-843X.103.2.231

Raftery AE. 1995. Bayesian model selection in social research. *Sociol Methodol* **25:** 111–163. doi:10.2307/271063
Raio CM, Carmel D, Carrasco M, Phelps EA. 2012. Nonconscious fear is quickly acquired but swiftly forgotten. *Curr Biol* **22:** R477–R479. doi:10.1016/j.cub.2012.04.023
Rauch SL, Shin LM, Phelps EA. 2006. Neurocircuitry models of posttraumatic stress disorder and extinction: human neuroimaging research—past, present, and future. *Biol Psychiatry* **60:** 376–382. doi:10.1016/j.biopsych.2006.06.004
R Core Team. 2016. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
Rescorla R, Wagner A. 1972. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In *Classical conditioning: current research and theory* (ed. Black A, Prokasy W), pp. 64–99. Appleton-Century-Crofts, New York.
Ressler KJ, Mayberg HS. 2007. Targeting abnormal neural circuits in mood and anxiety disorders: from the laboratory to the clinic. *Nat Neurosci* **10:** 1116–1124. doi:10.1038/nn1944
Roy M, Shohamy D, Wager TD. 2012. Ventromedial prefrontal-subcortical systems and the generation of affective meaning. *Trends Cogn Sci* **16:** 147–156. doi:10.1016/j.tics.2012.01.005
Schiller D, Levy I, Niv Y, LeDoux JE, Phelps EA. 2008. From fear to safety and back: reversal of fear in the human brain. *J Neurosci* **28:** 11517–11525. doi:10.1523/JNEUROSCI.2265-08.2008
Schmidt T. 2015. Invisible stimuli, implicit thresholds: why invisibility judgments cannot be interpreted in isolation. *Adv Cogn Psychol* **11:** 31–41. doi:10.5709/acp-0169-3
Shanks DR. 2016. Regressive research: the pitfalls of post hoc data selection in the study of unconscious mental processes. *Psychon Bull Rev* **24:** 752–775. doi:10.3758/s13423-016-1170-y
Spielberger C. 1983. *Manual for the State-Trait Inventory STAI (Form Y)*. Mind Garden, Palo Alto, CA.
Staib M, Castegnetti G, Bach DR. 2015. Optimising a model-based approach to inferring fear learning from skin conductance responses. *J Neurosci Methods* **255:** 131–138. doi:10.1016/j.jneumeth.2015.08.009
Stein T, Sterzer P. 2014. Unconscious processing under interocular suppression: getting the right measure. *Front Psychol* **5:** 387. doi:10.3389/fpsyg.2014.00387
Stein T, Hebart MN, Sterzer P. 2011. Breaking continuous flash suppression: a new measure of unconscious processing during interocular suppression? *Front Hum Neurosci* **5:** 167. doi:10.3389/fnhum.2011.00167
Szymanski J, O'Donohue W. 1995. Fear of spiders questionnaire. *J Behav Ther Exp Psychiatry* **26:** 31–34. doi:10.1016/0005-7916(94)00072-t
Tsuchiya N, Koch C. 2005. Continuous flash suppression reduces negative afterimages. *Nat Neurosci* **8:** 1096–1101. doi:10.1038/nn1500
Whalen PJ, Rauch SL, Etcoff NL, McInerney SC, Lee MB, Jenike MA. 1998. Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. *J Neurosci* **18:** 411–418. doi:10.1523/JNEUROSCI.18-01-00411.1998
Wiens S, Katkin ES, Ohman A. 2003. Effects of trial order and differential conditioning on acquisition of differential shock expectancy and skin conductance conditioning to masked stimuli. *Psychophysiology* **40:** 989–997. doi:10.1111/1469-8986.00117

Received September 29, 2019; accepted in revised form December 2, 2020.