Affective flexibility without perceptual awareness

Philipp Homan,[1] H. Lee Lau,[1] Ifat Levy,[2] Candace M. Raio,[3] Dominik R. Bach,[4,5]
David Carmel,[6*] Daniela Schiller,[1*]

[1]Department of Psychiatry, Department of Neuroscience, and Friedman Brain Institute,
Icahn School of Medicine at Mount Sinai
1470 Madison Ave, NY, NY 10029, USA.
[2]Departments of Comparative Medicine, Neuroscience and Psychology
Yale University, New Haven, CT, USA.
[3]Department of Psychology, New York University
6 Washington Place, NY, NY 10003, USA.
[4]Clinical Psychiatry Research
Department of Psychiatry, Psychotherapy, and Psychosomatics
University of Zurich, Zurich, Switzerland
[5]Wellcome Centre for Human Neuroimaging,
12 Queen Square, London WC1N 3BG, UK
[6]Psychology Department, University of Edinburgh
7 George Square, Edinburgh EH8 9JZ, UK.

[*]These authors contributed equally
[*]Correspondence: daniela.schiller@mssm.edu or dave.carmel@ed.ac.uk

2018-07-11 21:30

Abstract

Can we change our reactions to threat without being aware of it? Environmental 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. Threat-conditioned stimuli that are perceived without awareness can still elicit physiological defensive reactions, but it remains unknown whether the complex learning involved in affective flexibility - creating new threat associations while suppressing old ones - can be accomplished without awareness. Here, participants underwent classical threat conditioning, in which 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. The images were presented under continuous flash suppression, a technique allowing suppression of stimuli from awareness for several seconds. Importantly, although the trial-by-trial effectiveness of suppression varied across participants, parametric and computational analyses show that this was independent of the reversal learning evident in physiological (skin conductance) responses: dynamic updating - shifting reactions from a stimulus that no longer predicts threat to one that now does - was accomplished regardless of individual levels of perceptual awareness. In addition, baseline anxiety levels scaled negatively with the strength of this updating, consistent with previous reports of heightened anxiety impeding learning. Our findings provide convincing evidence that the successful updating of acquired threat responses does not require perceptual awareness.

Keywords: fear conditioning; perceptual awareness; reversal learning; unconscious processing; anxiety; continuous flash suppression

Significance statement

In an ever-changing environment, survival depends on learning which stimuli represent threat, and on updating such associations when circumstances shift. This is a complex process: novel responses must be generated while simultaneously suppressing learned ones. The role of awareness in this affective flexibility remains unknown, and has major implications for theories of both learning and consciousness. Humans can acquire physiological responses to threat-associated stimuli even when they are unaware of them; but we don't know whether the task of updating such learning, previously shown only with awareness, can be accomplished without it. Here, we show that it can, and furthermore, that human physiological responses reflect changes in stimulus-threat pairings independently of stimulus awareness, demonstrating the sophistication of unconscious affective flexibility.

Introduction

Flexible responses to environmental threats are essential for adaptive behavior. This involves learning to respond to stimuli that signal threat, but also the ability to update such responses if environmental contingencies change - new threats may arise while old ones cease to pose a risk. It is well established that learned defensive responses can be updated when danger is consciously perceived [1–3] - physiological reactions shift from a stimulus that no longer predicts danger to one that now does. But can we update our reaction to stimuli that predict danger when we are not even aware of them? Stimuli previously associated with an aversive outcome can elicit physiological arousal when they are suppressed from awareness [4–7], and new threat associations can be formed through classical conditioning even without any awareness of the conditioned stimuli [8–11]; however, it remains unknown whether the complex process of updating threat learning requires awareness of the stimuli, 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 employed the reversal paradigm, a laboratory model that requires flexible updating of threat contingencies [2]. 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. To assess learning, participants' physiological arousal is recorded throughout the experiment, typically (and here) by measuring their skin conductance responses. Appropriate response reversal requires a sophisticated form of updating, 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.

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 [10, 12–14]: 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 Figure 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 [15]. 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 break trough had occurred - may bias the findings ([16,17]; See Supplementary Methods for further details of these issues.) Here, we adopt a number of methodological approaches to ensure our results are robust to these potential concerns.

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 has been previously shown to correlate with unconscious learning; [10]). 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 a model in which they were simply (and independently of the stimulus) predicting the probability of a shock on the next trial. Finally, to verify that reversal learning can be induced with awareness of our stimuli, we ran a no-CFS group (N = 12), in which participants also viewed the CSs monocularly, 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. We find that CS awareness is indeed very weakly associated with reversal, and that there is strong evidence for the reversal of threat learning even in its complete absence.

Results and Discussion

Overall assessment of physiological reversal learning

To assess the physiological arousal evoked by CSs, we used a model-based approach [18] to estimate the amplitude of anticipatory sudomotor nerve activity (SNA) from skin conductance data recorded during stimulus presentation. A variational Bayes approximation was employed to invert a forward model that describes how hidden SNA translates into observable SCRs (see Materials and Methods). Previous work has shown that this approach is more sensitive than conventional SCR peak-to-peak analysis [18–20]. 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. To quantify the magnitude of physiological reversal learning, we calculated a reversal learning index for each participant (see Materials and Methods). The reversal learning index was positive (indicating successful reversal) and significantly greater than zero for both the CFS and no-CFS groups (Figure 2B).

Specifically, a linear mixed model (see Materials and Methods for details) revealed a significant interaction of stage and spider in both groups (CFS: $\beta = 0.27$, $t\ (2935) = 4.23$, $P < 0.001$; no-CFS: $\beta = 1.23$, $t\ (2935) = 7.29$, $P < 0.001$); the interaction was stronger in the no-CFS group than the CFS group (interaction of group, stage, and spider; $\beta = -0.96$, $t\ (2935) = -5.35$, $P < 0.001$). Note that a significant interaction is formally equivalent to a significant reversal learning index. These results indicate that reversal learning was evident in both groups, and more pronounced in the no-CFS group.

As previous work has found a negative association between anxiety and threat acquisition with and without awareness [10], we also calculated correlations between the CFS group's baseline anxiety measures (STAIT, STAIS, FSQ) and the reversal learning index. Overall, reversal learning decreased significantly with increasing levels of state and trait anxiety, and to a lesser but non-significant extent for spider phobia (Figure 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 non-consciously. The CFS group showed significantly lower accuracy in response to the "which seen?" question (M = 0.46, SD = 0.29) compared to 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 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, $P < 0.001$).

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. To control for possible artifacts of regression to the mean (see Supplementary Methods), we first calculated the correlation between two independent estimates of the awareness index [16], one calculated from even-numbered trials, the other from odd-numbered trials. These measures were strongly correlated (r (84) = 0.96, $P < 0.001$; Figure 3A); participants' awareness level in one set of trials was thus overwhelmingly predictive of their awareness in the other set, demonstrating the reliability of the awareness measure and indicating that an individual's overall awareness index was unlikely to have an extreme value that was due to measurement-level noise.

Next, we examined the association between the reversal learning index and the awareness index, using values of both indices obtained separately from even (Figure 3B) and odd (Figure 3C) trials. As the color-coding of Figure 3 shows, the relation between individual participants' reversal learning and their awareness 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 (Figure 3D). The parameter of interest was the intercept, which corresponds to the magnitude of reversal learning at zero perceptual awareness. The intercept was positive and significantly different from zero, and 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 Materials and Methods; Figure 3E and Table 1).

Comparing learning and expectation-based accounts

Well-controlled lab-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. Even without any awareness of the CSs, a participant should have been able to distinguish two types of trials: reinforced (with shock) and non-reinforced (no-shock). In a study with two CSs and a 100% reinforcement rate like ours, such expectations would correspond to an anticipated pattern of alternating 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.

To answer this question, we used a Bayesian approach to compare the probability of our findings being accounted for by a classic Rescorla-Wagner learning model [21] and a trial-sequence model. We hypothesized that successful threat reversal without perceptual awareness should be better explained by the Rescorla-Wagner learning model, whereas simple pattern-based expectation would be better explained by the trial-sequence learning model. 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. The Rescorla-Wagner model (BIC: 562.1) outperformed the pattern-based expectation model (BIC: 584.9), with the difference (Δ BIC: 22.9) greater than 10, suggesting that the evidence against the trial switch model is very strong [22]. Repeating this comparison for just the participants with zero mean awareness confirmed the lower BIC for the Rescorla-Wagner model (BIC: 114.3) compared to the pattern-based expectation model (BIC: 125.7), with the difference again greater than 10 (Δ BIC: 11.3; see also Figure S2). This model comparison provides convincing evidence that a classical Rescorla-Wagner learning model explains our findings better than an alternative expectation-based model.

These results indicate that participants were able to update their defensive physiological responses independently of their awareness of threat-related cues. Previous studies have shown that new threat associations can be formed without perceptual awareness of the conditioned stimuli [5, 9–11]. However, until now it was unknown whether the far more complex process of threat reversal - shifting reactions from a stimulus that no longer predicts danger to one that now does - can be accomplished without awareness. Our finding of reversal learning occurring independently of the level of perceptual awareness suggests that separate processes underlie affective flexibility and conscious processing [23]. Conversely, the negative correlation between reversal learning and anxiety suggests that the various impairments caused by anxiety are not limited to systems underlying conscious processes.

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 [24], Bayesian statistics [25], or parametric variation of the experimental manipulation [26]. 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.

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 [1] where threat associations are formed [27]. Consistent with this, it is known that patients with anxiety disorders often show rigid and inflexible threat responses in conjunction with prefrontal cortex dysfunction [28, 29]. 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 [30]. Compared to 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 non-adaptive). Threat learning without perceptual awareness is also negatively correlated with baseline state anxiety in healthy participants [10]. Our new finding that baseline anxiety is negatively correlated with affective flexibility suggests a potential use for reversal learning as a model paradigm for investigating how anxiety modulates various processes in a variety of disorders, including, for example, posttraumatic stress disorder, in which there is an impairment of threat inhibition [31].

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). 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 [32]; STAIT and STAIS, respectively) and spider phobia (Fear of Spider Questionnaire; FSQ [33]) were taken prior to participation and did not differ between the groups (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 procedure

Participants viewed the stimuli monocularly, through a mirror stereoscope (StereoAids, Australia) placed at a distance of 45 cm from a 17-inch 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, multi-colored, 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. 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 s 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 of the same trial type ever occurred consecutively.

Trial order and spider identity were counterbalanced across participants. To assess the effectiveness of the awareness manipulation [34], 1 s 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 s 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 s inter-trial interval.

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 (200ms; 50pulse/s) were delivered at a level determined individually by each participant as "uncomfortable but not painful" (maximum of 60V), 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

We estimated SNA from SCR data with a model-based variational Bayes approximation [18], inverting a forward model that describes how (hidden) SNA translates into (observable) SCR. A unit increase in SNA corresponds to an increase in SCR of 1 micro Siemens. The model assumes that the observed SCR can be decomposed into different components including anticipation, evocation, and spontaneous fluctuations, each of which are generated by bursts of SNA driven by changes in sympathetic arousal. The generative (forward) model thus describes how sympathetic arousal, the physiological measure that is taken as an index of the psychological process of threat, translates into sudomotor nerve bursts which then generate the observable SCR [18]. Using Bayesian inference, the forward model can then be reversed in order to estimate the most likely underlying SNA given the observed SCR:

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)}, \tag{1}$$

where the most likely parameter vector θ (corresponding to the SNA) given the observed outcome y (corresponding to the SCR) is given by the prior estimate of θ weighted by the likelihood of y given θ. Solving this equation involves integration over the model evidence p(y) which is analytically hard to compute (and possibly intractable). This can be resolved by replacing this integration problem by an optimization problem, which can be approximated with Variational Bayes procedures [35], where the log of the model evidence can be framed as the sum of the Kullback-Leibler divergence and the Free Energy. By maximizing the Free Energy the Kullback-Leibler divergence is minimized, and a lower bound to the log model evidence can be derived iteratively.

The SNA estimates were computed using previously developed software package PsPM [18] implemented in MATLAB R2016b (The Mathworks Inc, Natick, MA, USA). The statistical analyses were conducted with the R software [36] (R version 3.4.2 (2017-09-28)) and the libraries lme4 [37] and lsmeans [38]. Welch's t-tests were used instead of two sample t-tests when groups had unequal variances.

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):

$$
\begin{aligned}
\text{Reversal learning index} &= \Delta\text{Acquisition} - \Delta\text{Reversal} \\
\Delta\text{Acquisition} &= [\text{mean(Spider A)} - \text{mean(Spider B)}]_{\text{Acquisition}} \qquad (2) \\
\Delta\text{Reversal} &= [\text{mean(Spider A)} - \text{mean(Spider B)}]_{\text{Reversal}}
\end{aligned}
$$

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, spider B; Figure 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:

$$f_{p,k}(i, n) = p f_{p,k}(i + 1, n - 1) + (1 - p) f_{p,k}(0, n - 1), \tag{3}$$

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., $f_{0.5,8}(0, 32)$, 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 Figure S1B, the result for 8 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 (8 or more consecutive hits); notably, 3 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 8 or more consecutive hits.

Comparing learning and expectation-based models

The Rescorla-Wagner model [21] describes how the prediction for each trial is updated according to a prediction error and learning rate:

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

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 $V_0 = 0.5$, reflecting the assumption that getting a shock or not was equally likely for the first trial.

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 larger than 0.5, a trial switch is expected), α' is the learning rate, and δ'_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 τ_n, 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.

We performed a formal model comparison between the conventional Rescorla-Wagner model and the trial switch model for our data set (Figure S2), using maximum likelihood estimation and non-linear optimization (implemented with the fmincon function in MATLAB R2016b (The Mathworks Inc, Natick, MA, USA). Using the log likelihood, we calculated the Bayesian Information Criterion (BIC) to compare the two models as follows:

$$\text{BIC} = \log(n)\, k - 2 \cdot \log(\hat{L}), \tag{7}$$

where n is the number of data points, k is the number of regressors, and $\hat{L}$ is the maximized value of the likelihood function.

Acknowledgments

All authors declare no conflicts of interest with regard to the current study. 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 NIMH 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. The funding source had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

References

[1] Morris JS, Dolan RJ (2004) Dissociable amygdala and orbitofrontal responses during reversal fear conditioning. *Neuroimage* 22(1):372–80.

[2] Schiller D, et al. (2008) Evidence for recovery of fear following immediate extinction in rats and humans. *Learning and Memory* 15(6):394–402.

[3] Fleming SM, Dolan RJ, Frith CD (2012) Metacognition: computation, biology and function. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 367(1594):1280–6.

[4] Critchley HD, Mathias CJ, Dolan RJ (2002) Fear conditioning in humans: the influence of awareness and autonomic arousal on functional neuroanatomy. *Neuron* 33(4):653–63.

[5] Morris JS, Ohman A, Dolan RJ (1998) Conscious and unconscious emotional learning in the human amygdala. *Nature* 393(6684):467–70.

[6] Ohman A, Soares JJ (1994) "unconscious anxiety": phobic responses to masked stimuli. *Journal of Abnormal Psychology* 103(2):231–40.

[7] Whalen PJ, et al. (1998) Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. *Journal of Neuroscience* 18(1):411–8.

[8] Katkin ES, Wiens S, Ohman A (2001) Nonconscious fear conditioning, visceral perception, and the development of gut feelings. *Psychol Sci* 12(5):366–70.

[9] Manns JR, Clark RE, Squire LR (2002) Standard delay eyeblink classical conditioning is independent of awareness. *Journal of Experimental Psychology: Animal Behavior Processes* 28(1):32–7.

[10] Raio CM, Carmel D, Carrasco M, Phelps EA (2012) Nonconscious fear is quickly acquired but swiftly forgotten. *Current Biology* 22(12):R477–9.

[11] 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(6):989–97.

[12] Tsuchiya N, Koch C (2005) Continuous flash suppression reduces negative afterimages. *Nature Neuroscience* 8(8):1096–101.

[13] Stein T, Hebart MN, Sterzer P (2011) Breaking continuous flash suppression: A new measure of unconscious processing during interocular suppression? *Frontiers in Human Neuroscience* 5:167.

[14] Carmel D, Arcaro M, Kastner S, Hasson U (2010) How to create and use binocular rivalry. *J Vis Exp* (45).

[15] Gayet S, Stein T (2017) Between-subject variability in the breaking continuous flash suppression paradigm: Potential causes, consequences, and solutions. *Frontiers in Psychology* 8:437.

[16] Shanks DR (2016) Regressive research: The pitfalls of post hoc data selection in the study of unconscious mental processes. *Psychonomic Bulletin & Review*.

[17] Stein T, Seymour K, Hebart MN, Sterzer P (2014) Rapid fear detection relies on high spatial frequencies. *Psychol Sci* 25(2):566–74.

[18] Bach DR, Daunizeau J, Friston KJ, Dolan RJ (2010) Dynamic causal modelling of anticipatory skin conductance responses. *Biological Psychology* 85(1):163–70.

[19] Bach DR (2014) A head-to-head comparison of scralyze and ledalab, two model-based methods for skin conductance analysis. *Biological Psychology* 103:63–8.

[20] Staib M, Castegnetti G, Bach DR (2015) Optimising a model-based approach to inferring fear learning from skin conductance responses. *Journal of Neuroscience Methods* 255:131–8.

[21] Rescorla R, Wagner A (1972) *A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement*, eds. Black A, Prokasy W. (Appleton-Century-Crofts, New York), pp. 64–99.

[22] Raftery AE (1995) Bayesian model selection in social research. *Sociological methodology* pp. 111–163.

[23] Lau H, Rosenthal D (2011) Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences* 15(8):365 – 373.

[24] Fleming SM, Lau HC (2014) How to measure metacognition. *Frontiers in Human Neuroscience* 8:443.

[25] Dienes Z (2015) *How Bayesian statistics are needed to determine whether mental states are unconscious*, ed. Overgaard M. (Oxford University Press, Oxford), pp. 199–221.

[26] Schmidt T (2015) Invisible stimuli, implicit thresholds: Why invisibility judgments cannot be interpreted in isolation. *Adv Cogn Psychol* 11(2):31–41.

[27] Roy M, Shohamy D, Wager TD (2012) Ventromedial prefrontal-subcortical systems and the generation of affective meaning. *Trends in Cognitive Science* 16(3):147–56.

[28] Ressler KJ, Mayberg HS (2007) Targeting abnormal neural circuits in mood and anxiety disorders: from the laboratory to the clinic. *Nature Neuroscience* 10(9):1116–24.

[29] Rauch SL, Shin LM, Phelps EA (2006) Neurocircuitry models of posttraumatic stress disorder and extinction: human neuroimaging research–past, present, and future. *Biological Psychiatry* 60(4):376–82.

[30] Duits P, et al. (2015) Updated meta-analysis of classical fear conditioning in the anxiety disorders. *Depression and Anxiety* 32(4):239–253.

[31] Jovanovic T, Norrholm SD (2011) Neural mechanisms of impaired fear inhibition in posttraumatic stress disorder. *Frontiers in Behavioral Neuroscience* 5:44.

[32] Spielberger CD (1983) State-trait anxiety inventory for adults. *PsycTESTS Dataset*.

[33] Szymanski J, O'Donohue W (1995) Fear of spiders questionnaire. *PsycTESTS Dataset*.

[34] Merikle PM, Smilek D, Eastwood JD (2001) Perception without awareness: perspectives from cognitive psychology. *Cognition* 79(1-2):115–34.

[35] Friston K, Mattout J, Trujillo-Barreto N, Ashburner J, Penny W (2006) Variational free energy and the laplace approximation. *NeuroImage* 34(1):220–234.

[36] R Core Team (2016) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria).

[37] Bates DM (2005) Fitting linear mixed models in R. *R News* 5:27–30.

[38] Lenth RV (2016) Least-squares means: The R package lsmeans. *Journal of Statistical Software* 69(1):1–33.

[39] Li J, Schiller D, Schoenbaum G, Phelps EA, Daw ND (2011) Differential roles of human striatum and amygdala in associative learning. *Nature Neuroscience* 14(10):1250–1252.

Tables and Figures

Model	Predictor	Beta	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

Table 1: Regression coefficients for all awareness index models. 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; model 3 additionally included STAIT and tracking scores.

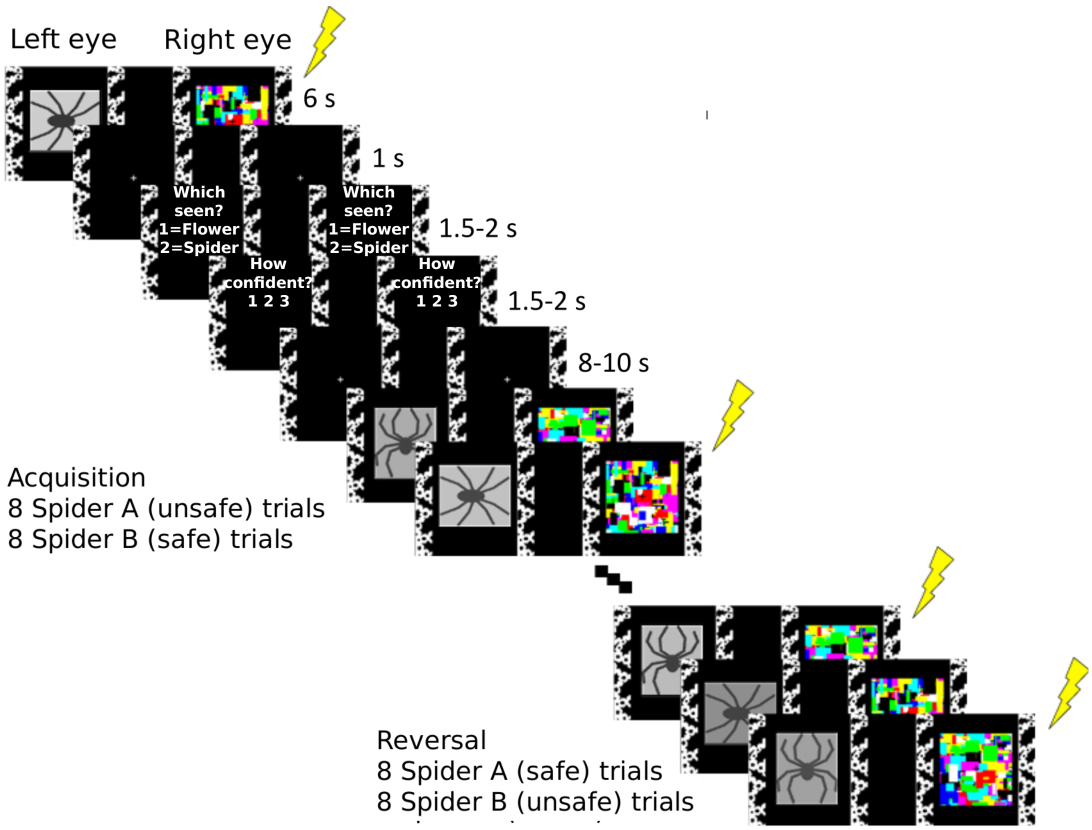

Figure 1: Experimental design and procedure. In each trial of the acquisition phase, participants were presented with one of two stimuli (schematic pictures of spiders, presented monocularly for 6 sec and suppressed from awareness by a CFS mask shown to the other eye). One image (spider A) always terminated 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 8 times in each phase. Trial order was pseudorandomized (see Materials and Methods) and spider identity (A and 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 s each), beginning 1 s after the offset of every CS, and followed by an 8-10 s inter-trial interval. Participants who underwent the same procedure without CFS were shown identical CSs, but the CFS mask was absent.

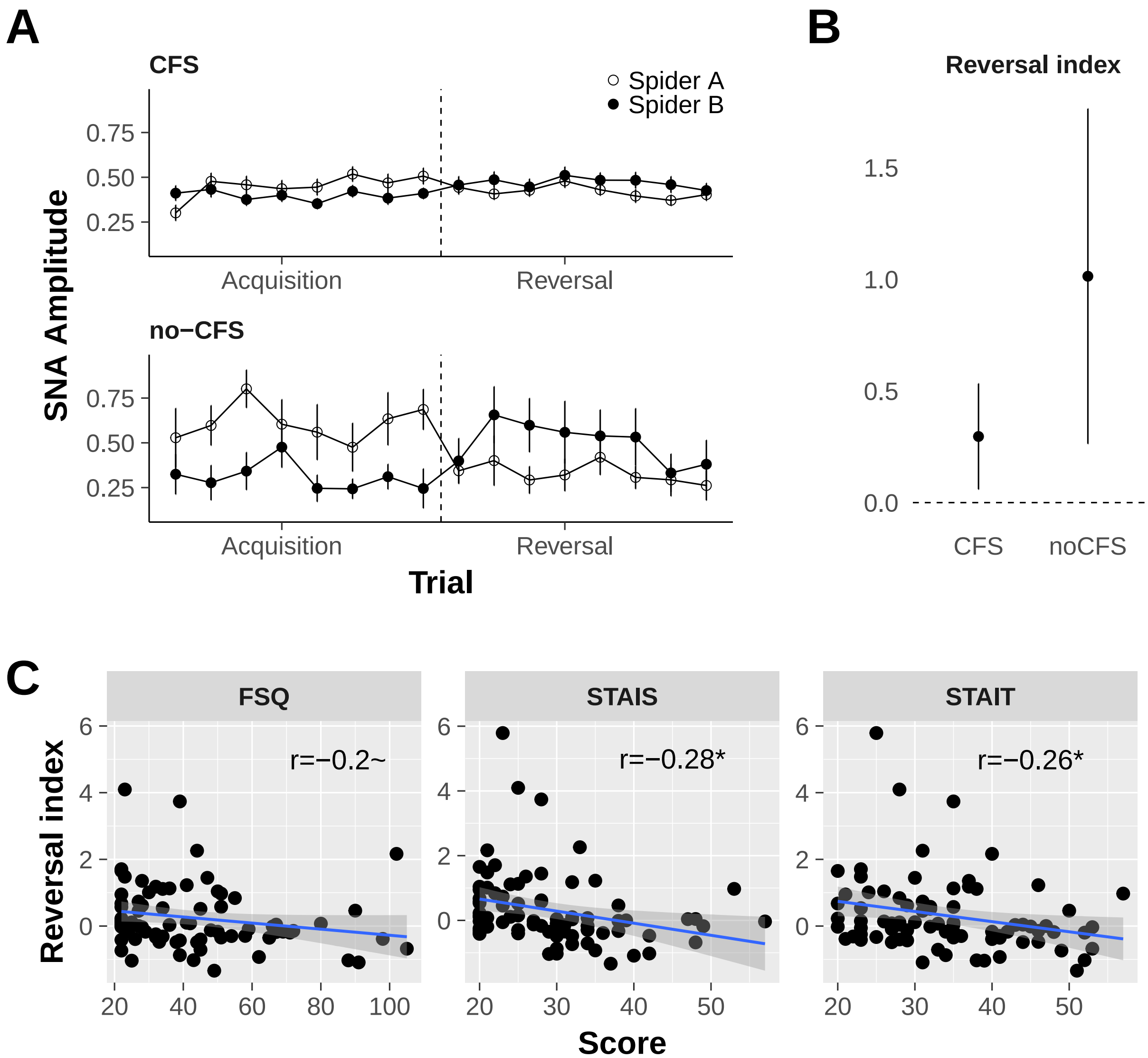

Figure 2: Physiological reversal learning. A. Time courses reveal reversal of threat responses with and without continuous flash suppression. Data points represent trial-wise mean responses to spider A (the CS+ during acquisition) and spider B (the CS- during acquisition). Both groups showed evidence for reversal learning as indicated by the interaction of stage (acquisition, reversal) and stimulus (spider A, spider B) and quantified by the reversal index. 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. Abbreviations: STAIS/STAIT, state/trait anxiety subscale of the Spielberger State-Trait Anxiety Inventory; FSQ, Fear of Spider Questionnaire, ~, $P < .1$; *, $P < .05$.

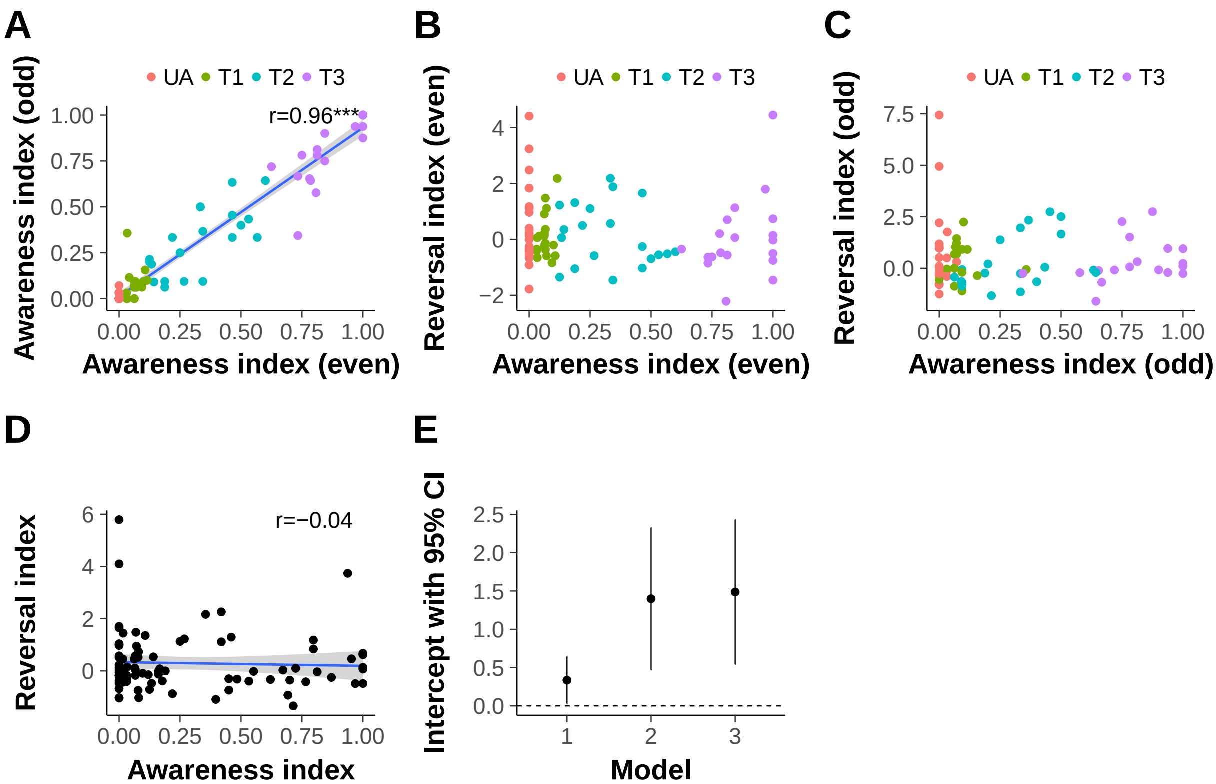

Figure 3: Characterizing the relation between perceptual awareness and reversal learning. A. Correlation between the awareness index of even and odd-numbered trials. Each data point represents an individual participant. The strong positive correlation between these independent measures of awareness demonstrates 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 sample in 3 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 Panel 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, 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. Intercepts and their 95% confidence intervals in a series of regression models.** Reversal learning is predicted by the perceptual awareness index (model 1), the perceptual awareness index and STAIT scores (model 2), and perceptual awareness, STAIT and tracking scores (model 3). (Excluding the potential outlier in the top left corner of panel D weakens significance of the intercept in model 1, P = 0.07; the intercepts of model 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.

Supplementary Information

Supplementary Methods

Investigating unconscious perception relies on the effectiveness of the technique used to suppress stimuli from awareness. Although CFS is highly effective, it is not foolproof - different observers are not equally susceptible to it, and stimuli often break through the suppression. The standard approach to dealing with breakthrough of the suppressed stimulus is to remove from analysis those trials and individuals in which it occurred. This approach is problematic, though, as it may lead to various artifacts. Below we detail the concerns that are addressed by the analysis approach we adopted in the present study.

Regression to the mean

Recent computational work suggests that findings in the trials that remain after removing suppression failures (unless the number of such trials is negligibly small) could be the result of regression to the mean [16]: If two noisy measures of the same underlying phenomenon are used (e.g., behavioral and skin conductance responses may both measure conscious processing, each with its own, independent measurement noise), then selecting only the extreme cases of one measure (e.g., only those cases where behavior indicates complete absence of awareness) is unlikely to yield similarly extreme results for the other measure (e.g., it is unlikely that the skin conductance responses will also be close to zero). Thus, a result that looks as if it indicates unconscious processing would in fact be an artifact that is entirely due to regression to the mean.

To address this issue, we refrain from removing trials or participants with breakthrough. Instead, we include all of them in our analyses and assess CS awareness in a continuous manner by assigning each participant an awareness rating; we base this rating on two independent measures of awareness (odd and even-numbered trials), as proposed by Shanks [16] to control for regression to the mean in awareness estimates. We then examine the association between the level of CS awareness and the amount of reversal learning indicated by physiological stimulus-evoked responses (Figure S1).

Sensitivity of awareness assessments

The tasks used to assess whether participants were aware of the suppressed stimulus (typically through examining, after each trial, whether they can confidently and accurately report which stimulus was presented) may not be sensitive enough to detect such awareness. This is because breakthrough, leading to at least some awareness, may not always reach a level that allows above-chance performance on the chosen discrimination task (e.g., one might be able to distinguish between two stimuli without being able to label either, a problem that falls within the general concern of "sensitivity dissociation" between direct and indirect measures; [17]).

To address this issue, we analyze individual participants' response patterns across trials, searching for any link between stimuli and behavioral responses - including consistently wrong ones - that may indicate potential tracking of the stimuli, suggesting some perceptual awareness (even in the absence of an explicit ability to identify the images) and thus possibly accounting for physiological findings. We incorporate a binary factor indicating whether or not tracking behavior was observed in each participant into the regression model accounting for physiological responses.

Participants expectations

Laboratory models of conditioning require a pre-defined number of trials for each CS category, to avoid confounding learning with exposure to different stimulus ratios. Furthermore, the order of trials is often not completely random but constrained to allow for roughly similar distributions of CSs across the experiment [2,39]. This means that participants would not be wrong to predict the probability of receiving aversive stimulation on a given trial using a heuristic that takes into account the overall rate of the aversive stimulus and the number of trials since its last occurrence (this is akin to the gambler's fallacy, but in such cases has a higher-than-chance success rate). The likelihood of participants basing their anticipation of a shock - and of the physiological arousal that comes with such anticipation - on such probabilistic estimates may be even higher when CSs are suppressed from awareness, leaving the overall rate and distribution of shocks as the only bases for such predictions. We address this issue by we using Bayesian model comparison, demonstrating that a model featuring updating of learning account for the findings better than a shock-expectation model based on the distribution of shocks across the experiment.

Characteristic	N (CFS)	Mean	SD	N (no-CFS)	Mean	SD
Male	38	nil	nil	7	nil	nil
Female	48	nil	nil	5	nil	nil
Age, y	86	29.7	8.3	12	31.7	10.6
STAIT	80	34.2	9.6	11	33.5	12.3
STAIS	80	29.4	8.6	11	28.4	9.2
FSQ	79	43.5	21.5	12	36.8	12.1
UR, SNA units	86	1.8	0.5	12	1.8	0.7

Table 1: Sample characteristics. Note: STAI data for 7 participants (6 in the CFS condition) and FSQ data for 7 participants (all in the CFS condition) were lost due to an archiving error. Exclusion of these participants from subsequent analyses does not alter the overall pattern of results. *Abbreviations*: CFS, Continuous Flash Suppression; SD, Standard deviation; STAIS/STAIT, state/trait anxiety subscale of the Spielberger State-Trait Anxiety Inventory; FSQ, Fear of Spider Questionnaire; UR, unconditioned response (shock response); SNA, sudomotor nerve activity.

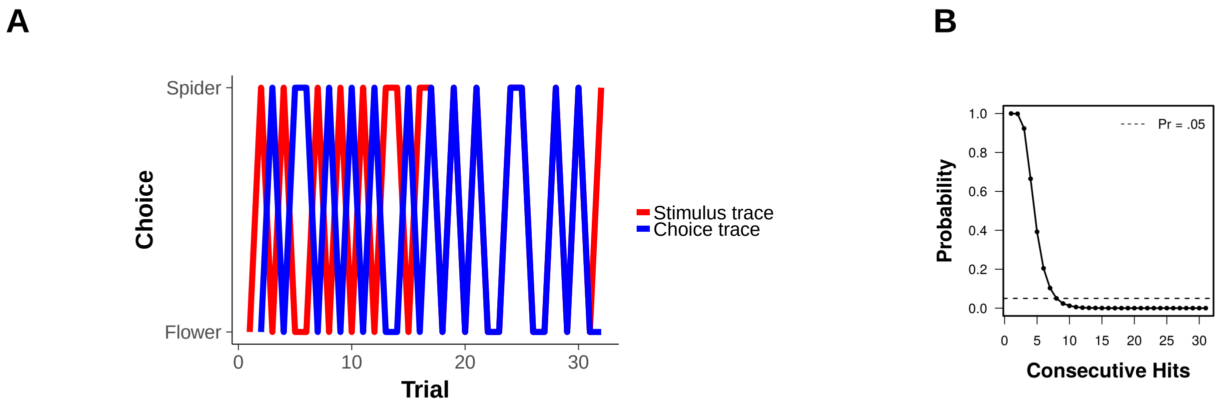

Figure 1: Examination of tracking behavior in the correspondence between time courses of forced-choice responses and stimulus presentation in the CFS group. A. Forced choice responses (blue) versus stimulus presentation (red) in an illustrative participant. This participant shows periods in which forced-choice responses track or are in direct opposition to both the presented stimuli and shock application, suggesting that this participant tracked the spider and/or the shock with his responses. **B. Probability of consecutive hits (10^5 simulations).** Hits were defined as the number of consecutive trials where the time courses of forced-choice responses and stimulus presentation were either identical or directly opposed to each other. To quantify the probability of consecutive hits, we simulated the probabilities for all possible levels of consecutive hits (e.g., from 1 to 31 for 32 trials). We simulated 10^5 draws of a binomial distribution, and calculated the average probability to see (at least) each possible number of consecutive hits. The probability exponentially declines as the number of consecutive hits increases, reaching a probability of about 5% at 8 consecutive hits. Thus, observing 8 or more consecutive hits in 32 trials in a perceptually unaware participant who is truly guessing is unlikely, which is why we considered this an indication for stimulus tracking behavior.

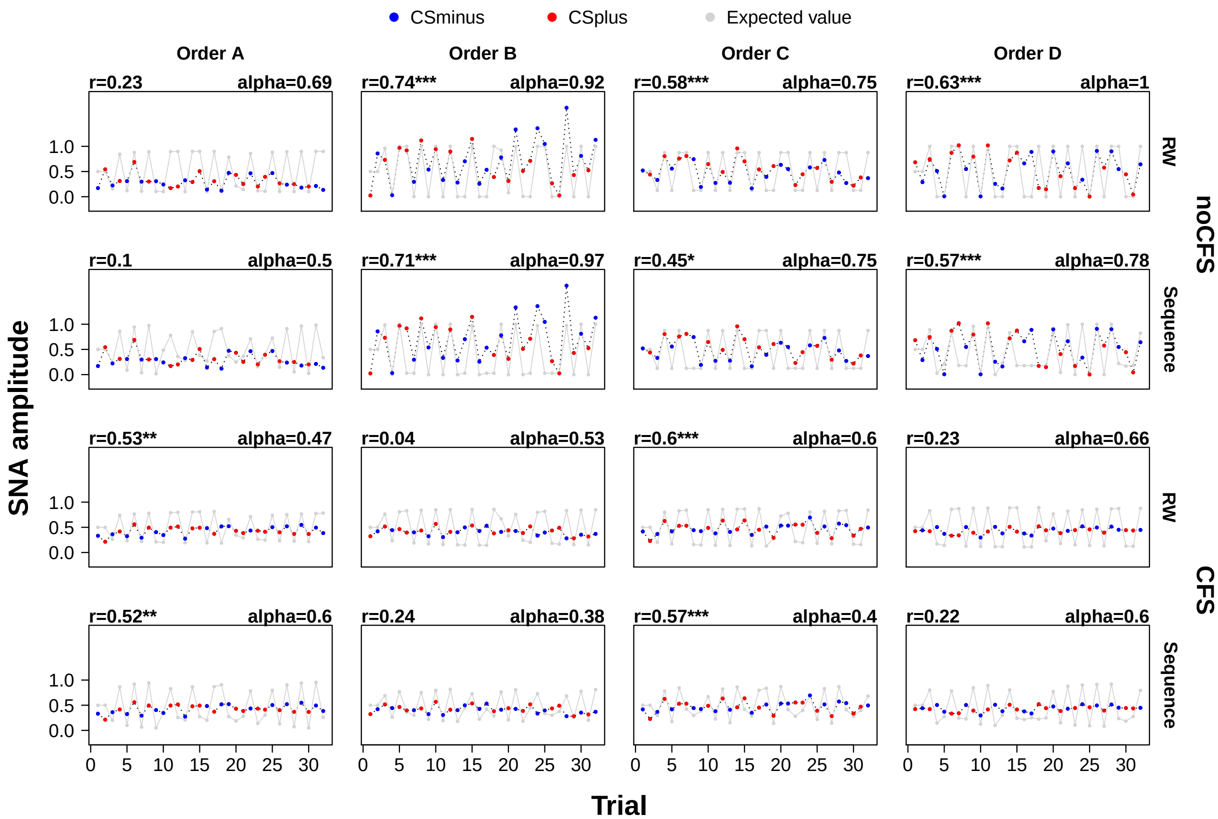

Figure 2: Time series of measured and predicted threat response indicate that the Rescorla-Wagner learning model explains the data better than an alternative trial-sequence learning model that simply assumed a sequence of alternating trial types (no-shock/shock or vice versa). Trial-wise mean responses to CS+ and CS- are shown by group, learning model and experimental order. Each panel also shows the Pearson correlation between predicted and measured time series and the learning rate (parameter alpha in each of the learning models). The generally higher Pearson correlation coefficients are consistent with the formal model comparison between the Rescorla-Wagner model and the trial-sequence model, which confirmed that the Rescorla-Wagner model provided stronger evidence than the alternative learning model. r, Pearson correlation coefficient; alpha, learning rate; RW, Rescorla-Wagner learning model; Sequence, trial-sequence learning model; SNA, sudomotor nerve activity; *, $P < 0.001$; **, $P < 0.01$; *, $P < 0.05$.