The impact of prior and ongoing threat on the false alarm threshold for facial discrimination

Santiago Papini, Joseph E. Dunsmoor, Jasper A. J. Smits

The University of Texas at Austin

Corresponding Author:

Santiago Papini, Institute for Mental Health Research, The University of Texas at Austin

305 E. 23rd St., Stop E9000, Austin, TX, 78712

512-471-7694 | spapini@utexas.edu

Abstract

Perceptual adaptations facilitate rapid responses to threats but can come with the cost of false alarms, or the failure to discriminate safe or novel stimuli from signals of true threat. For example, a fatigued colleague might be avoided when their tired expression is interpreted as a scowl, or a glimpse at a stranger might cause a rush of anxiety if they resemble a known adversary. We examined false alarms in the context of facial cues, which can become exaggerated signals of threat across anxiety disorders. In Experiment 1, ongoing threat lowered the false alarm threshold for discrimination based on anger intensity compared to prior and no threat. In Experiment 2, prior and ongoing threat each lowered the false alarm threshold for identity-based facial discrimination compared to no threat. These results could be relevant for anxiety disorders in which excessive false alarms may contribute to overgeneralized threat responses.

In the context of threat perception, a false alarm occurs when a safe or novel stimulus is mistaken for a signal of threat. An experimental analysis of the impact of prior and ongoing threats on false alarms may clarify their impact on the generalization of threat responses, which is a putative mechanism in the development and maintenance of anxiety disorders. Toward this aim, we conducted two facial discrimination experiments designed to answer two questions: 1) How do varying degrees of threat impact the false alarm threshold? 2) What are the relationships between varying levels of threat, perceptual adaptations, behavior, and subjective anxiety? In this introduction we provide the basis for our primary hypothesis by drawing from an evolutionary account of the adaptive value of false alarms in the context of facial cues related to social threat perception. Next, we review the role of false alarms in anxiety disorders and outline recent experimental evidence of their potential role in the overgeneralization of threat responses, which are the key areas in which our experiments aim to make a contribution.

Facial expressions can be powerful social signals of potential threats (Öhman, 1986; Olsson & Phelps, 2004). We get frightened when we observe intense anger or fear in others. We also remember the faces of people that have threatened us before, and are quick to recognize them in future encounters regardless of their emotional expression. The capacity for facial discriminations across the dimensions of emotional expression and identity facilitates rapid activation of diverse responses tuned to threats spanning physical harm, competition for resources, and social rejection (Neuberg, Kenrick, & Schaller, 2011; D. J. Stein & Nesse, 2011). Despite their adaptive value, threat responses can compete with other behaviors that are necessary for reproductive success (LeDoux & Daw, 2018). Therefore, an excessively low threshold for false alarms – one that reduces all risk at the expense of unnecessary protective responses – is not optimal (Nesse, 2001).

The cost of unnecessary protective responses is particularly salient for humans living in relatively safe environments far removed from the contexts in which these responses evolved. Cognitive, physiological, and behavioral responses to perceived threats pull time, attention, and energy away from activities that are crucial for an individual's wellbeing (Piccirillo, Taylor Dryman, & Heimberg, 2016). In addition to these costs, excessive use of otherwise adaptive protective actions, such as escape and avoidance, prevents the acquisition of evidence necessary to calibrate the false alarm threshold (Goetz, Davine, Siwiec, & Lee, 2016; Telch & Lancaster, 2012). Not surprisingly, inaccurate threat perceptions are a prominent transdiagnostic feature of anxiety disorders (Pittig, Treanor, LeBeau, & Craske, 2018; M. B. Stein & Paulus, 2009). Therefore, an increased understanding of the complex process of discrimination between true threats and false alarms may lead to novel therapeutics or increased precision in the application of existing treatments for anxiety disorders.

Despite the role of perceptual false alarms in anxiety disorders, experimental research has primarily focused on Pavlovian conditioning (Struyf, Zaman, Vervliet, & Van Diest, 2015), which pairs innocuous stimuli with aversive outcomes to examine the generalization of threat responses to cues resembling the conditioned stimulus (Pavlov (1927), 2010). Within the context of facial discrimination, conditioned physiological and neuronal responses to fearful faces generalized to faces with greater emotional intensity that were never paired with aversive outcomes such as shock (Dunsmoor, Mitroff, & LaBar, 2009; Dunsmoor, Prince, Murty, Kragel, & LaBar, 2011). This effect was larger in participants with posttraumatic stress disorder (Morey et al., 2015), which is characterized by excessive emotional and physiological reactivity to cues that resemble some aspect of the traumatic event. Experiments that blended the facial features of two distinct individuals to examine generalization along identity showed the activation of autonomic responses and explicit threat predictions were limited to faces that could not be perceptually discriminated from the threat conditioned target (Holt et al., 2014; Tuominen et al., 2019). However, retrospective identification of the target face elicited false alarms to similar but discriminable faces and neuronal responses also generalized beyond the threshold of discrimination (Tuominen et al., 2019), pointing to differences in the role of perception on the generalization of diverse responses. Across these experiments, aversive stimuli were presented during the generalization test, so it is plausible that ongoing threat impacted perceptual discrimination of threat-associated faces in a manner consistent with an evolutionary account of the adaptive value of rapid threat detection (Lynn & Barrett, 2014; D. J. Stein & Nesse, 2011).

Our goal was to examine the impact of prior and ongoing threat on false alarms, which may clarify the boundary between adaptive and pathological generalization of threat responses, and contribute to an improved understanding of their role in anxiety disorders. Two experiments were conducted that involved an initial learning phase, during which a target (one of two facial stimuli) was paired with an outcome, followed by a subsequent testing phase, during which participants had to discriminate between the target and facial morphs that varied in their degree of similarity to the target (Fig. 1). Ongoing threat was modeled by pairing the target face with aversive outcomes throughout both phases and prior threat was modeled by pairing the target with aversive outcomes during the learning phase only. The absence of threat was modeled by pairing the target with a neutral outcome during the learning phase only. These experiments were designed to test the primary hypothesis that ongoing threat would lower the threshold for false alarms, which would facilitate rapid reactions. To examine potentially causal relationships between varying degrees of threat and the perceptual, emotional, and behavioral responses we measured, we used exploratory graph analysis. To determine whether results varied as a function of stimulus dimension we conducted two experiments: Experiment 1 (*N* = 90) used facial stimuli that varied on the dimension of anger intensity and Experiment 2 (*N* = 90) used facial stimuli that varied on the dimension of identity.

Experiment 1: Facial Discrimination Along the Dimension of Anger Intensity

Method

Participants

Participants were 90 undergraduate students recruited from the University of Texas at Austin Psychology Subject Pool who reported no current or past psychiatric hospitalizations, diagnoses, or medication, no history of head injury that required hospitalization or resulted in a loss of consciousness, and had normal or corrected vision. We selected a sample size comparable or larger than experiments examining generalization in anxiety disorders (Kaczkurkin et al., 2016; Kaczkurkin & Lissek, 2013; Lissek et al., 2014, 2010) and within-subjects examinations of the effect of shock on perception of fear in faces (Lim & Pessoa, 2008). All participants viewed a video summary and read a detailed document of study procedures before providing written informed consent. The mean age of participants was 18.54 (*SD* = 0.75), 59 (65.56%) were females, and 33 (36.67%) self-identified as White, 29 (32.22%) as Asian or Pacific Islander, 19 (21.11%) as Hispanic or Latino/a, 8 (8.89%) as Black, and 1 (1.11%) as Other. All randomized participants received research credit. Procedures were approved by the Institutional Review Board at The University of Texas at Austin.

Overview of experiment

The experiment was programmed in SuperLab (Version 5.0.5) and run on a 21.5” iMac. All responses were made on a Dell optical two-button mouse. Participants first learned and practiced responding *yes* by clicking the left mouse button and *no* by clicking the right mouse button whenever the word *yes* or *no* was shown on the screen. Participants had to complete 20 trials in a row without error in order to begin the experiment.

The key phases of the experiment were a learning and a testing phase. In the learning phase identical instructions were displayed on the computer screen for all participants: *You are ready to begin the task. In this task the target image is always followed by a sound. The other images will never be followed by a sound. Do not respond in this phase. Just learn which is the target, which is followed by sound.* The learning phase used two images of a Caucasian male individual's face with either a neutral or an angry expression; the angry face was always the target. Each face was presented twice in random order for 5 s with a 2 s intertrial interval (ITI). Facial stimuli were selected from the NimStim standardized and validated stimuli set (Tottenham et al., 2009). Next, participants read the following instructions: *Your task is to identify the target by clicking yes or no. If the image is exactly the same as the target, click yes with your index finger. If the image is different in any way, click no with your middle finger. Please be as accurate as possible.* During the first 4 trials, participants received feedback about the accuracy of their response, and in the remaining 16 trials participants did not receive feedback (total of 20 trials, 10 of each stimulus in random order). During this part of the learning phase, stimuli remained on the screen until participants responded. Throughout the learning phase, a sound was played immediately after every presentation of the target face.

The testing phase began after a 10 m break. Participants read these instructions: *In this phase we are testing your memory of the target. If the image is exactly the same as the target from the previous phase, click yes with your index finger. If the image is different in any way, click no with your middle finger. The images may repeat themselves. You may or may not hear a sound. Your response should be based on what you learned in the previous phase. Please be as* *accurate as possible.* The testing face used nine additional images of the same individual with a facial expression that morphed from the neutral face presented in learning (i.e., 0% similarity to the target) to the angry (target) face (i.e., 100% similarity) in 10% increments, for a total of 11 different stimuli that were presented in six randomized blocks. As with the learning phase, stimuli were presented until a response was made, and there was a 2 s ITI between trials.

Threat manipulation

Participants were randomized to one of three groups: Ongoing-Threat, Prior-Threat, or No-Threat. Group names denote the presence or absence of aversive outcomes during the experimental phases. In the learning phase, the target was followed by a 500 ms startling sound (i.e., a scream) in the Ongoing-Threat and Prior-Threat groups delivered through Bose over-ear noise-cancelling headphones. In the No-Threat group, a neutral 500 ms tone was used instead of an aversive US. The target in the Ongoing-Threat and Prior-Threat groups was also paired with shock, which was delivered simultaneously with the startle sound. During the testing phase, the target face was no longer paired with an outcome in the Prior-Threat and No-Threat groups, but continued to be paired with scream+shock in the Ongoing-Threat group.

Participants randomized to a group that included shock watched a brief video explaining the calibration procedure. Two electrodes were attached to the ring and middle fingers of the dominant hand, and the shocks (0.2 - 4 milliamps) were delivered through a 9V battery powered Coulbourn E13-22 (Coulbourn Instruments). Research assistants manually delivered shocks beginning at the lowest level and gradually increasing one level at a time. Throughout calibration, participants were reminded that if a level was too uncomfortable to continue, the shock could be reduced to a previous level. During the experiment, the shock was activated with a digital pulse and lasted 0.5 s. Importantly, all participants received identical instructions (i.e., learn the target followed by the sound), but the aversive quality of sound and its combination with a shock in the Ongoing-Threat and Prior-Threat groups were designed to condition a threat association with the target.

In sum, Ongoing-Threat was modeled by pairing the target with an aversive outcome across both phases, Prior-Threat was modeled by pairing the target with an aversive outcome in the learning phase only, and No-Threat was modeled by pairing the target with a neutral sound. One participant randomized to Ongoing-Threat discontinued participation after the learning phase, and data from two participants in Prior-Threat were excluded due to random responding in the test phase. These participants were identified during ongoing data collection, therefore recruitment continued until the final sample size of 90 was reached (30 participants in each group).

Subjective anxiety measure

To examine the effect of aversive outcomes (i.e., shock and startle sound) on anxiety levels, participants completed a brief version of the State-Trait Anxiety Inventory (STAI-6) (Marteau & Bekker, 1992) three times during the experiment: prior to randomization to measure baseline state anxiety, immediately after completion of the learning phase, and immediately after completion of the testing phase.

Analyses

Reproducibility

All analyses were carried out in R version 3.6.1. Data and complete analysis outputs with session information output of the analysis environment (e.g., exact package versions used) are available on the Open Science Foundation (OSF) repository for this project: https://osf.io/m582k/.

False alarm threshold

Signal Detection Theory (SDT) was applied to model responses in the testing phase (Green & Swets, 1966). In a prototypical signal detection experiment, a stimulus (e.g., light or sound) is presented at various levels of intensity spanning from the range of imperceptible to always perceived. The 50% detection threshold, sometimes referred to as the *point of subjective equality*, represents the intensity at which the stimulus has a probability of being detected half of the time. In other words, stimuli with intensities greater than the 50% detection threshold are more likely to be detected than not. The analytic methods of SDT have been applied beyond the perceptual domain, including two-alternative forced choice tasks with facial stimuli where a sigmoid shaped function models the change in probability of a response as a function of a facial feature that is parametrically modulated across stimuli, such as emotional expression (Lynn & Barrett, 2014). The interpretation of the threshold depends on the task instructions: “Is the face angry?” would yield an *anger perception threshold*; “Do you expect a shock?” would yield a *threat expectancy threshold*; “Are you afraid?” would yield a *fear activation threshold*. Our experiment asked participants to identify whether stimuli were identical to the previously learned target, and incorrect *yes* responses were false alarms. As such, we defined the threshold in our model as the *false alarm threshold* or the level of similarity to the target above which participants were more likely to have false alarms (i.e., respond *yes*) than correct rejections (i.e., respond *no*).

To model response data we used the cumulative normal function, which estimates the probability of a *yes* response as a function of stimulus similarity to the target. Two parameters were extracted from the models: $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}$. The false alarm threshold corresponds to the target-similarity level, $\boldsymbol{\alpha}$, at which the probability of responding *yes* was 50%, and beyond which false alarms become more likely than correct rejections. For example, if $\boldsymbol{\alpha} = 70\%$, then faces with more than 70% anger intensity would be more than 50% likely to activate a false alarm. We tested the hypothesis that the Ongoing-Threat group would exhibit a lower false alarm threshold by comparing **α** between groups. We also reported the values of **α** at which the maximum between-groups differences in false alarm probability were observed. These models excluded trials with the target image, for which a *yes* response would not be considered a false alarm. Instead, the proportions of correctly identified targets were reported separately.

We also examined the **σ** parameter, which is the standard deviation of **α** and provides a measure of precision. Larger **σ** indicate greater uncertainty around the false alarm threshold (i.e., lower precision), whereas smaller **σ** indicate more precise discrimination between stimuli. We did not have *a priori* hypotheses about **σ**, but compared groups on this metric to examine whether ongoing threat had an effect on precision. The mean and 95% confidence intervals (CI) of **α** and **σ** were estimated in 1,000 non-parametric bootstraps applying a maximum-likelihood criterion (Linares & López i Moliner, 2016). Goodness of fit was assessed with the deviance statistic (Wichmann & Hill, 2001), which tests the hypothesis that the model fit is significantly different than the observed data. Therefore, a nonsignificant result suggests a good model fit. Group differences in parameter estimates and false alarm probabilities were considered significant when the 95% CI of the difference did not overlap with zero, or $p < .050$. We also examined whether results varied across stages of the testing phase, where Early was defined as the first 3 blocks and Late was defined as the last 3 blocks.

Response time

Analyses were conducted to test whether ongoing threat was associated with more rapid responses. Response times, recorded in ms, were log transformed to reduce skew, and participant-level averages were calculated separately for *yes* and *no* responses to assess the specificity of the effect of ongoing threat on response time. The effects of Group (Ongoing-

Threat, Prior-Threat, No-Threat), Response (*yes* or *no*), and Group $\times$ Response on response time were examined in a mixed model with random subject effects to account for the within-subject correlation between responses. When necessary, models were rerun excluding outliers with a standardized residual exceeding 2.5 standard deviations. Significant interactions were followed up with pairwise contrasts. We also examined whether results depended on phase, where Early was defined as the first 3 blocks and Late was defined as the last 3 blocks.

Subjective anxiety

Analyses were conducted to test whether the threat manipulation was associated with increases in subjective anxiety. The effects of Group (Ongoing-Threat, Prior-Threat, No-Threat), Phase (post-learning and post-testing), and Group $\times$ Phase on the STAI-6 total scores were examined in a mixed model with random subject effects to account for correlated repeated measures and controlling for pre-randomization STAI-6 scores. When necessary, models were rerun excluding outliers with a standardized residual exceeding 2.5 standard deviations. Significant interactions were examined with pairwise contrasts of between- and within-group effects.

Exploratory graph analysis

To complement the primary analyses and examine potentially causal relations between threat and the perceptual, behavioral, and emotional measures acquired throughout the experiment, we generated directed acyclic graphs (DAG), which can uncover nonlinear, directed relationships among variables (Rohrer, 2018). Bayesian structure learning algorithms were applied in a bootstrapping framework that tests the directionality of effects between variables using only basic information about known relationships (e.g., temporal sequence of variables). To disaggregate the effects of prior and ongoing threat we used two binary variables (threat during learning and during testing). All data from the previous analyses were included: anxiety scores from the learning and testing phases, reaction times for *yes* and *no* responses (raw data was used since this analysis is nonparametric), and false alarm threshold and precision parameters, which were estimated fitting models on the subject-level data using the same approach as the group-level analysis.

Models were estimated with the following prior information based on the temporal sequence of measurement: causal pathways to threat or anxiety in the learning phase were prohibited, given that the remaining variables were measured after the learning phase. Highly correlated variables that do not have a putative causal relationship can bias DAG models; therefore, we excluded pathways between reaction times for *yes* and *no* responses. However, we did not make any assumptions about potential relationships between the false alarm threshold and precision, as it has been argued that these may affect one another (Lynn & Barrett, 2014). The graph structure was estimated by averaging results from 1,000 bootstrap replications. Pathways that appeared in the majority of replications were preserved and represented by unidirectional arrows on a graph, where arrow thickness corresponds to path replicability.

Results

False alarm threshold

Figure 2A shows the model estimated probabilities of false alarms as a function of stimulus similarity to the target. Deviance statistics suggested good model fit across groups (all ps > .998). Relative to Prior-Threat, Ongoing-Threat resulted in a significantly lower false alarm threshold, mean difference -7.61%, 95% CI:[-9.44, -5.87], and a peak difference of 17.99% higher probability of false alarms, 95% CI: [13.60, 22.22], which occurred at 73.75% morph similarity. Relative to No-Threat, Ongoing-Threat resulted in a significantly lower false alarm threshold, mean difference -8.19%, 95% CI:[-10.12, -6.40], and a peak difference of 19.22%

higher probability of false alarms, 95% CI: [14.96, 23.40], which occurred at 75.85% morph similarity. Consistent with the degree of overlap in Figure 2A, there were no differences between Prior-Threat and No-Threat across these analyses. Moreover, the proportion of correctly identified targets was not significantly different between groups, Ongoing-Threat *M* = 96.11%, 95% CI: [92.97, 99.25], Prior-Threat *M* = 96.11%, 95% CI: [92.97, 99.25] (identical to Ongoing-Threat), and No-Threat *M* = 91.67%, 95% CI: [85.61, 97.73].

To examine whether effects on the false alarm threshold were similar across the experiment, we reran analyses with a Stage term, which consisted of Early (i.e., first 3 blocks) and Late (i.e., last 3 blocks). Deviance statistics suggested good model fit across groups and stages (all *p*s > .955). A consistent pattern of results was observed across stages such that the false alarm threshold was significantly lower for Ongoing-Threat relative to No-Threat, Early mean difference -8.61%, 95% CI: [-10.89, -6.50], Late mean difference -8.22%, 95% CI: [-11.14, -5.72], and relative to Prior-Threat, Early mean difference -6.95%, 95% CI: [-9.12, -4.91], Late mean difference -7.32%, 95% CI: [-9.17, -5.66].

Precision parameters were not significantly different between groups, suggesting that ongoing threat did not significantly affect the capacity to discriminate between stimuli near the false alarm threshold. However, when examined across the Early and Late stages of the experiment, Prior-Threat was associated with significantly better precision than No-Threat in the Late stage, mean difference -5.68%, 95% CI: [-10.19, -0.39]. All other differences were nonsignificant.

Response time

Figure 2B shows mean response times in ms with 95% CI for *yes* and *no* responses. Analyses showed a significant Group × Response interaction, $F(2, 86.14) = 4.87$, $p = .010$. There were no significant differences between groups in response times when responding *no*, all ps > .192. However, for *yes* responses, Ongoing-Threat resulted in significantly faster RTs relative to No-Threat, $t(127) = 3.48$, $p < .001$, $d = 0.62$, and Prior-Threat, $t(124) = 3.05$, $p = .003$, $d = 0.54$, whereas the difference between Prior-Threat and No-Threat was nonsignificant, $t(127) = 0.47$, $p = .64$.

To examine whether effects on response times were similar across the experiment, we reran analyses with a Stage term, which consisted of Early (i.e., first 3 blocks) and Late (i.e., last 3 blocks). The Group × Response × Stage interaction was nonsignificant, $p = .344$, suggesting that the relationship between Group and Response in reaction time did not differ across stages. However, the Group × Response interaction was significant $F(2, 253.61) = 11.79$, $p < .001$, suggesting that on average and after taking Stage into account, reaction time patterns differed as a function of Group and Response type. Consistent with the prior model that excluded Stage, pairwise comparisons showed significantly faster reaction times for *yes* responses for Ongoing-Threat compared to both No-Threat, $t(114) = 3.48$, $p < .001$, $d = 0.62$, and Prior-Threat, $t(127) = 3.48$, $p < .001$, $d = 0.62$, but no significant differences to *no* responses, all ps > .207. Together, these results suggest that Group effects were consistent across the Early and Late stages of the task.

Subjective anxiety

Figure 2C shows mean STAI-6 anxiety scores with 95% CI across phases of the experiment. Analyses showed a significant Group × Phase interaction, $F(2, 84.81) = 16.25$, $p < .001$. Relative to No-Threat, significantly higher levels of anxiety were reported for Ongoing-Threat, $t(117) = 7.09$, $p < .001$, $d = 1.31$, and for Prior-Threat, $t(118) = 6.43$, $p < .001$, $d = 1.18$, immediately after the learning phase, which demonstrates that the use of aversive outcomes during the learning phase resulted in increased anxiety. Immediately post-testing, anxiety levels were significantly higher for Ongoing-Threat relative to both Prior-Threat, $t(118) = 4.49$, $p < .001$, $d = 0.83$, and relative to No-Threat, $t(117) = 6.62$, $p < .001$, $d = 1.22$. Only the Prior-Threat group showed a significant change from post-train to post-test, $t(85.7) = 5.64$, $p < .001$, $d = 1.22$, which demonstrates that the discontinuation of aversive outcomes resulted in a decrease in anxiety. However, the difference in post-test anxiety levels between Prior-Threat and No-Threat remained significant, $t(118) = 2.12$, $p = .039$, $d = 0.39$.

Exploratory graph analysis

Figure 3A shows the directed acyclic graph (DAG) for Experiment 1. Results from the DAG are consistent with the other analyses, and suggest independent causal pathways from threat to anxiety and to the perceptual features of facial discrimination. Specifically, ongoing threat had a direct effect on anxiety in the testing phase that was independent of its direct effect on the false alarm threshold. Reaction time for *yes* responses was affected directly by threat in the learning phase, and indirectly by the effect of ongoing threat on the false alarm threshold. Thus, the faster *yes* responses during Ongoing-Threat may be due to the cumulative effects of previously acquired threat associations and a decreased false alarm threshold. Although differences in precision were not observed in the main SDT analysis, the DAG model, which adjusts for relations among all variables, identified a path from threat during the learning phase to precision in the testing phase.

Experiment 2: Facial Discrimination Along the Dimension of Identity

Method

Participants

Participants were undergraduate students recruited from The University of Texas at Austin Psychology Subject Pool who reported no current or past psychiatric hospitalizations, diagnoses, or medication, no history of head injury that required hospitalization or resulted in a loss of consciousness, and had normal or corrected vision. All participants viewed a video summary and read a detailed document of study procedures before providing written informed consent. Two participants randomized to the Prior-Threat group discontinued participation, and data from two participants (one in the Prior-Threat group and one in the No-Threat group) were excluded due to random responding in the test phase, leaving a final sample size of 90 (30 participants in each group). The mean age of participants was 18.91 (SD = 1.07), 53 (58.89%) were females, and 32 (35.56%) self-identified as White, 29 (32.22%) as Asian or Pacific Islander, 22 (24.44%) as Hispanic or Latino/a, 4 (4.44%) as Black, and 3 (3.33%) as Middle Eastern. All randomized participants received research credit. Procedures were approved by the Institutional Review Board at The University of Texas at Austin.

Overview of experiment

Experimental procedures for Experiment 2 were identical to Experiment 1. However, facial stimuli depicted two distinct Caucasian male individuals with a neutral facial expression, selected from the same standardized and validated stimuli set as Experiment 1. The learning phase used two images (i.e., the two different individuals). The testing face used nine additional images that were parametrically morphed from a distinct identity (i.e., 0% similarity to the target) to the target identity (i.e., 100% similarity) in 10% increments. Thus, whereas Experiment 1 stimuli used the same identity and morphed along the dimension of emotional expression from neutral to angry, Experiment 2 stimuli used two identities with neutral facial expressions and morphed along the dimension of identity.

Analyses

Analyses followed the same procedures as Experiment 1. In the DAG analysis, data from Experiment 2 were fit to the model from Experiment 1 to facilitate comparison.

Results

False alarm threshold

Figure 2D shows the model estimated probabilities of false alarms as a function of stimulus similarity to the target. Deviance statistics suggested good model fit across groups (all *p*s > .999). Relative to No-Threat, Ongoing-Threat had a significantly lower false alarm threshold, mean difference -4.38%, 95% CI:[-6.38, -2.41], and a peak difference of 17.19% higher probability of false alarms, 95% CI: [13.52, 20.88], which occurred at 55.08% morph similarity. The false alarm threshold was not significantly different between Ongoing-Threat and Prior-Threat. However, Prior-Threat resulted in a significantly lower false alarm threshold than No-Threat, mean difference -4.97%, 95% CI: [-6.61, -3.18]. The proportion of correctly identified targets was significantly lower for Ongoing-Threat, *M* = 88.33%, 95% CI: [82.87, 93.79], compared to Prior-Threat, *M* = 97.78%, 95% CI: [95.63, 99.93] and compared to No-Threat, *M* = 96.67%, 95% CI: [94.13, 99.20].

To examine whether effects on the false alarm threshold were consistent, we examined results across the Early and Late stages. Deviance statistics suggested good model fit across groups and stages (all *p*s > .991). In the Early stage, Ongoing-Threat resulted in a lower false alarm threshold compared to No-Threat, mean difference -6.88%, 95% CI: [-9.10, -4.59], and compared to Prior-Threat, mean difference -2.54%, 95% CI: [-5.06, -0.003]. However, by the Late stage, Ongoing-Threat resulted in a higher false alarm threshold compared to Prior-Threat, mean difference 3.90%, 95% CI: [1.65, 6.39]. Across both stages, Prior-Threat was associated with a significantly lower false alarm threshold relative to No-Threat, mean difference Early = -4.34%, 95% CI: [-6.44, -2.00], mean difference Late -5.70%, 95% CI: [-7.45, -3.83].

Precision was significantly lower (as indicated by a larger **σ** parameter) for Ongoing-Threat relative to No-Threat, mean difference 9.30%, 95% CI: [6.69, 11.83], and relative to Prior-Threat, mean difference 4.42%, 95% CI: [1.47, 7.40]. Precision was also significantly lower for Prior-Threat relative to No-Threat, mean difference 4.89%, 95% CI: [2.42, 7.44]. Together, these results suggest that threat associations formed in the learning phase and ongoing threat during the testing phase both impacted the capacity to discriminate between stimuli, such that as the threat level increased, precision decreased. When precision was examined across stages, Ongoing-Threat was not significantly different than Prior-Threat in the Early stage, but by the Late stage, Ongoing-Threat resulted in significantly lower precision compared to Prior-Threat, mean difference 7.35%, 95% CI: [3.79, 11.38]. Consistent with the overall analysis, Ongoing-Threat resulted in significantly lower precision than No-Threat across both stages, Early mean difference 8.11%, 95% CI: [5.14, 10.97], Late mean difference 10.35%, 95% CI: [7.24, 13.64].

Response time

Figure 2E shows mean response times in ms with 95% CI for *yes* and *no* responses. The Group × Response interaction was nonsignificant, $p = .334$. However, there was a main effect of Group, $F(2, 86.19) = 3.49$, $p = .035$. Post hoc pairwise comparisons showed a significantly faster overall response time for Ongoing-Threat relative to No-Threat, $t(87.2) = 2.61$, $p = .011$, $d = 0.56$. Response times between the other groups were not significantly different (all *p*s > .099).

There were no significant interactions with Group in the model that included Stage, all *p*s > .111. However, the main effect of Group was significant, $F(2, 85.32) = 3.64$, $p = .003$, suggesting that on average and after taking Stage into account, reaction time differed by Group. Pairwise comparisons showed significantly faster reaction time for Ongoing-Threat relative to No-Threat, $t(87.1) = 2.68$, $p = .009$, $d = 0.57$. Together, these results suggest that Group effects were consistent across the Early and Late stages of the task.

Subjective anxiety

Figure 2F shows mean STAI-6 anxiety scores with 95% CI across phases of the experiment. Analyses showed a significant Group × Phase interaction, $F(2, 85) = 14.30$, $p < .001$. Relative to No-Threat, there were significantly higher levels of anxiety reported in the Ongoing-Threat group, $t(120) = 5.85$, $p < .001$, $d = 1.07$, and the Prior-Threat group, $t(121) = 5.16$, $p < .001$, $d = 0.94$, immediately after the learning phase. Immediately post-testing, anxiety levels were significantly higher for Ongoing-Threat relative to Prior-Threat, $t(121) = 5.32$, $p < .001$, $d = 0.97$, and relative to No-Threat, $t(120) = 7.19$, $p < .001$, $d = 1.31$. Only the Prior-Threat group showed a significant change from post-train to post-test, $t(85) = 5.91$, $p < .001$, $d = 1.28$. The difference in anxiety levels between Prior-Threat and No-Threat at post-test was nonsignificant, $t(121) = 1.95$, $p = .053$, $d = 0.35$.

Exploratory graph analysis

Figure 3B shows the DAG results for Experiment 2. Pathways that replicated across experiments are represented by solid arrows, and pathways that were not significant in Experiment 2 are represented with dashed arrows. Within the context of Experiment 2, results from the DAG are consistent with the previous analyses, and suggest independent causal pathways from threat to anxiety and to precision in the testing phase. Specifically, threat during the learning phase impacted anxiety and had an independent, direct effect on precision, which is consistent with the SDT analyses. There was also a direct effect of the false alarm threshold on reaction time for *yes* responses. However, there were no direct or indirect effects of threat (learning or testing phase) on the false alarm threshold, or reaction time.

Discussion

Our primary aim was to test the hypothesis that ongoing threat would lower the false alarm threshold for facial discrimination. We found that both prior and ongoing threat can cause perceptual adaptations that lower the false alarm threshold, but that these effects may depend on the facial features involved in the discrimination. One key difference that may explain the divergence of results is that Experiment 1 involved a threat-relevant signal (i.e., anger intensity), whereas Experiment 2 involved a threat-irrelevant signal (i.e., facial identity). A large body of empirical data suggests stronger conditioning of physiological and threat expectancy responses for threatening (vs. neutral) faces through an enhanced resistance to extinction (for review see Dimberg & Öhman, 1996). If conditioned responses are activated by inaccurate threat perceptions, one may infer that the false alarm threshold is more likely to be lower for angry (vs. neutral) faces previously paired with aversive outcomes, yet across our two experiments we found the opposite: Prior-Threat decreased the false alarm threshold in the discrimination of neutral faces that morphed along the dimension of identity, whereas the false alarm threshold for angry morphs was not affected by Prior- relative to No-Threat. Future experiments can be designed to explicitly test whether the threat-relevance of stimuli differentially impacts false alarm thresholds for prior threats.

Another potential explanation for the divergence of Prior-Threat results is that as stimuli become less discriminable, uncertainty increases and false alarms become more widely distributed across the threshold (Lynn & Barrett, 2014), which is quantified by the precision parameter. Prior experiments focused on precision in the discrimination of threat conditioned stimuli showed enhanced precision when stimuli were olfactory (Li, Howard, Parrish, & Gottfried, 2008), and impaired precision when stimuli were auditory (Resnik, Sobel, & Paz, 2011). Yet it is unclear whether the differences across these separate experiments, which applied different procedures and sensory modalities, can be explained by differences in stimulus discriminability. In contrast, our experiments used identical procedures with facial stimuli that may have varied in their discriminability. The emotional expressivity of stimuli in Experiment 1 reflects a dynamic feature of faces for which a simple rule can be derived (e.g., the faces vary in the intensity of anger, and the target is the angriest face), whereas the invariant features that distinguished identities in our second experiment were not as amenable to simple rules. If discrimination between the identity morphs in Experiment 2, which showed reduced precision for Ongoing- and Prior-Threat, was more difficult than discrimination between the angry morphs in Experiment 1, which did not find significant differences in precision across groups, our results would be consistent with the hypothesis that the impact of prior threat on precision is stronger for stimuli that are more difficult to discriminate. Additionally, the exploratory graph analyses, which disaggregated the partial effects of prior and ongoing threat, found evidence of an effect of threat learning on precision across both experiments. An explanation that would be consistent with results across all analyses is that the effect of prior threat on precision was large enough in Experiment 2 to subsume additional perceptual effects of ongoing threat on the false alarm threshold. Future experiments can be designed to explicitly test this account and determine whether the differential impact of prior threat on precision is a function of stimulus discriminability.

As a secondary aim, we examined associations among threat, perceptual adaptations, behavioral responses, and emotional reactions. Across both experiments, the threat manipulations resulted in similar patterns of reaction time and subjective anxiety across groups and phases. While a lower false alarm threshold was associated with decreased response time as predicted, our experiments did not find a consistent link between subjective anxiety and perceptual adaptations. Recent experiments found that poor perceptual discrimination between circles that varied in size was associated with greater generalization of threat expectancy (Struyf, Zaman, Hermans, & Vervliet, 2017; Zaman, Ceulemans, Hermans, & Beckers, 2019; Zaman, Struyf, Ceulemans, Beckers, & Vervliet, 2019). One of these experiments found mixed evidence of increased perceptual errors when a target was learned through pairings with an aversive (as opposed to neutral) outcome, leading the authors to state that "one could argue that fear affects the perceptual system such that the perception of fear-evoking stimuli is favored" (Zaman, Struyf, et al., 2019). Alternatively, one could argue that threat affects the perceptual system by favoring false alarms and that subjective fear may be an epiphenomenon, which would be consistent with exploratory graph analyses showing that threat impacted subjective anxiety independently of its effects on perceptual features of discrimination

Divergence in the relationships between perception and indices of fear have also been observed in experiments that train perceptual discrimination across a variety of stimuli. For example, a threat conditioning and discrimination task that used bells that morphed across color found lower threat expectancy (but not physiological) responses among adults (Ginat-Frolich, Klein, Katz, & Shechner, 2017), whereas the same task found lower physiological (but not subjective) responses in children (Ginat-Frolich, Gendler, Marzan, Tsuk, & Shechner, 2019). Perceptual discrimination training resulted in decreased avoidance (but not threat expectancy) in a task that used geometric morphs (Lommen et al., 2017), and decreased avoidance and physiological responses to spider stimuli (Ginat-Frolich, Klein, Aderka, & Shechner, 2019).

Together, these results indicate that the impact of improved perceptual discrimination on threat responses depends on properties of the stimulus, type of threat response, and individual differences in participants such as age and level of anxiety.

In sum, our experiments captured the impact of threat on perceptual adaptations for the discrimination of facial stimuli. These findings provide avenues for further research on the role of perceptual false alarms in the overgeneralization of diverse threat responses, which has been reported in patients with panic disorder (Lissek et al., 2010), generalized anxiety disorder (Lissek et al., 2014), posttraumatic stress disorder (Kaczkurkin et al., 2016; Morey et al., 2015), and obsessive compulsive traits (Kaczkurkin & Lissek, 2013). Continued in-depth examination of the interplay between threat persistence, stimuli used, and the type of responses measured is necessary to facilitate identification of boundary conditions between adaptive and impaired discrimination of true threats from false alarms, which may have implications for anxiety disorders and other aspects of human functioning.

Author Contributions

S. Papini developed the study concept, supervised the collection of data by research assistants, and performed data analyses. All authors contributed to the interpretation of results. S. Papini drafted the manuscript, and J. Smits and J. Dunsmoor provided critical revisions. All authors approved the final version of the manuscript for submission.

Acknowledgments

We would like to acknowledge the research assistants and staff at the Anxiety and Health Behaviors Laboratory for their work in the data collection process, and the undergraduate students that volunteered to participate in the experiments.

Declaration of Conflicting Interests

S. Papini and J. E. Dunsmoor declared that there were no conflicts of interest with respect to the authorship or the publication of this article. J. A. J. Smits receives compensation for his role as a consultant to Big Health, Ltd.

Funding

S. Papini's work on this project was supported by the Donald D. Harrington Fellowship.

Open Practices

Data and analysis outputs can be downloaded at the Open Science Foundation repository for this project: https://osf.io/m582k/

References

Dimberg, U., & Öhman, A. (1996). Behold the wrath: Psychophysiological responses to facial stimuli. *Motivation and Emotion*, *20*(2), 149–182.

Dunsmoor, J. E., Mitroff, S. R., & LaBar, K. S. (2009). Generalization of conditioned fear along a dimension of increasing fear intensity. *Learning & Memory*, *16*(7), 460–469. https://doi.org/10.1101/lm.1431609

Dunsmoor, J. E., Prince, S. E., Murty, V. P., Kragel, P. A., & LaBar, K. S. (2011). Neurobehavioral mechanisms of human fear generalization. *NeuroImage*, *55*(4), 1878–1888. https://doi.org/10.1016/j.neuroimage.2011.01.041

Ginat-Frolich, R., Gendler, T., Marzan, D., Tsuk, Y., & Shechner, T. (2019). Reducing fear overgeneralization in children using a novel perceptual discrimination task. *Behaviour Research and Therapy*, *116*, 131–139. https://doi.org/10.1016/j.brat.2019.03.008

Ginat-Frolich, R., Klein, Z., Aderka, I. M., & Shechner, T. (2019). Reducing avoidance in adults with high spider fear using perceptual discrimination training. *Depression and Anxiety*, *36*(9), 859–865. https://doi.org/10.1002/da.22930

Ginat-Frolich, R., Klein, Z., Katz, O., & Shechner, T. (2017). A novel perceptual discrimination training task: Reducing fear overgeneralization in the context of fear learning. *Behaviour Research and Therapy*, *93*, 29–37.

Goetz, A. R., Davine, T. P., Siwiec, S. G., & Lee, H.-J. (2016). The functional value of preventive and restorative safety behaviors: A systematic review of the literature. *Clinical Psychology Review*, *44*, 112–124. https://doi.org/10.1016/j.cpr.2015.12.005

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1). Wiley New York.

Holt, D. J., Boeke, E. A., Wolthusen, R. P. F., Nasr, S., Milad, M. R., & Tootell, R. B. H. (2014). A parametric study of fear generalization to faces and non-face objects: Relationship to discrimination thresholds. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00624

Kaczkurkin, A. N., Burton, P. C., Chazin, S. M., Manbeck, A. B., Espensen-Sturges, T., Cooper, S. E., … Lissek, S. (2016). Neural Substrates of Overgeneralized Conditioned Fear in PTSD. *American Journal of Psychiatry*. https://doi.org/10.1176/appi.ajp.2016.15121549

Kaczkurkin, A. N., & Lissek, S. (2013). Generalization of Conditioned Fear and Obsessive-Compulsive Traits. *Journal of Psychology & Psychotherapy*, *7*, 3.

LeDoux, J., & Daw, N. D. (2018). Surviving threats: Neural circuit and computational implications of a new taxonomy of defensive behaviour. *Nature Reviews Neuroscience*, *19*(5), 269–282. https://doi.org/10.1038/nrn.2018.22

Li, W., Howard, J. D., Parrish, T. B., & Gottfried, J. A. (2008). Aversive Learning Enhances Perceptual and Cortical Discrimination of Indiscriminable Odor Cues. *Science (New York, N.Y.)*, *319*(5871), 1842–1845. https://doi.org/10.1126/science.1152837

Lim, S.-L., & Pessoa, L. (2008). Affective learning increases sensitivity to graded emotional faces. *Emotion*, *8*(1), 96.

Linares, D., & López i Moliner, J. (2016). *quickpsy: An R Package to Fit Psychometric Functions for Multiple Groups*. Retrieved from http://diposit.ub.edu/dspace/handle/2445/116040

Lissek, S., Kaczkurkin, A. N., Rabin, S., Geraci, M., Pine, D. S., & Grillon, C. (2014). Generalized Anxiety Disorder Is Associated With Overgeneralization of Classically Conditioned Fear. *Biological Psychiatry*, *75*(11), 909–915.

https://doi.org/10.1016/j.biopsych.2013.07.025

Lissek, S., Rabin, S., Heller, R. E., Lukenbaugh, D., Geraci, M., Pine, D. S., & Grillon, C. (2010). Overgeneralization of Conditioned Fear as a Pathogenic Marker of Panic Disorder. *American Journal of Psychiatry*, *167*(1), 47–55. https://doi.org/10.1176/appi.ajp.2009.09030410

Lommen, M. J. J., Duta, M., Vanbrabant, K., Jong, R. de, Juechems, K., & Ehlers, A. (2017). Training discrimination diminishes maladaptive avoidance of innocuous stimuli in a fear conditioning paradigm. *PLOS ONE*, *12*(10), e0184485. https://doi.org/10.1371/journal.pone.0184485

Lynn, S. K., & Barrett, L. F. (2014). “Utilizing” Signal Detection Theory. *Psychological Science*, *25*(9), 1663–1673. https://doi.org/10.1177/0956797614541991

Marteau, T. M., & Bekker, H. (1992). The development of a six-item short-form of the state scale of the Spielberger State—Trait Anxiety Inventory (STAI). *British Journal of Clinical Psychology*, *31*(3), 301–306. https://doi.org/10.1111/j.2044-8260.1992.tb00997.x

Morey, R. A., Dunsmoor, J. E., Haswell, C. C., Brown, V. M., Vora, A., Weiner, J., … LaBar, K. S. (2015). Fear learning circuitry is biased toward generalization of fear associations in posttraumatic stress disorder. *Translational Psychiatry*, *5*(12), e700–e700. https://doi.org/10.1038/tp.2015.196

Nesse, R. M. (2001). The Smoke Detector Principle. *Annals of the New York Academy of Sciences*, *935*(1), 75–85. https://doi.org/10.1111/j.1749-6632.2001.tb03472.x

Neuberg, S. L., Kenrick, D. T., & Schaller, M. (2011). Human threat management systems: Self-protection and disease avoidance. *Neuroscience & Biobehavioral Reviews*, *35*(4), 1042–

1051. https://doi.org/10.1016/j.neubiorev.2010.08.011

Öhman, A. (1986). Face the Beast and Fear the Face: Animal and Social Fears as Prototypes for Evolutionary Analyses of Emotion. *Psychophysiology*, *23*(2), 123–145. https://doi.org/10.1111/j.1469-8986.1986.tb00608.x

Olsson, A., & Phelps, E. A. (2004). Learned Fear of "Unseen" Faces after Pavlovian, Observational, and Instructed Fear. *Psychological Science*, *15*(12), 822–828. https://doi.org/10.1111/j.0956-7976.2004.00762.x

Pavlov (1927), P. I. (2010). Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex. *Annals of Neurosciences*, *17*(3), 136–141. https://doi.org/10.5214/ans.0972-7531.1017309

Piccirillo, M. L., Taylor Dryman, M., & Heimberg, R. G. (2016). Safety Behaviors in Adults With Social Anxiety: Review and Future Directions. *Behavior Therapy*, *47*(5), 675–687. https://doi.org/10.1016/j.beth.2015.11.005

Pittig, A., Treanor, M., LeBeau, R. T., & Craske, M. G. (2018). The role of associative fear and avoidance learning in anxiety disorders: Gaps and directions for future research. *Neuroscience & Biobehavioral Reviews*, *88*, 117–140. https://doi.org/10.1016/j.neubiorev.2018.03.015

Resnik, J., Sobel, N., & Paz, R. (2011). Auditory aversive learning increases discrimination thresholds. *Nature Neuroscience*, *14*(6), 791–796. https://doi.org/10.1038/nn.2802

Rohrer, J. M. (2018). Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27–42. https://doi.org/10.1177/2515245917745629

Stein, D. J., & Nesse, R. M. (2011). Threat detection, precautionary responses, and anxiety disorders. *Neuroscience & Biobehavioral Reviews*, *35*(4), 1075–1079. https://doi.org/10.1016/j.neubiorev.2010.11.012

Stein, M. B., & Paulus, M. P. (2009). Imbalance of Approach and Avoidance: The Yin and Yang of Anxiety Disorders. *Biological Psychiatry*, *66*(12), 1072–1074. https://doi.org/10.1016/j.biopsych.2009.09.023

Struyf, D., Zaman, J., Hermans, D., & Vervliet, B. (2017). Gradients of fear: How perception influences fear generalization. *Behaviour Research and Therapy*, *93*, 116–122. https://doi.org/10.1016/j.brat.2017.04.001

Struyf, D., Zaman, J., Vervliet, B., & Van Diest, I. (2015). Perceptual discrimination in fear generalization: Mechanistic and clinical implications. *Neuroscience & Biobehavioral Reviews*, *59*, 201–207. https://doi.org/10.1016/j.neubiorev.2015.11.004

Telch, M. J., & Lancaster, C. L. (2012). Is There Room for Safety Behaviors in Exposure Therapy for Anxiety Disorders? In P. Neudeck & H.-U. Wittchen (Eds.), *Exposure Therapy: Rethinking the Model—Refining the Method* (pp. 313–334). https://doi.org/10.1007/978-1-4614-3342-2_18

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., … Nelson, C. (2009). The NimStim set of facial expressions: Judgments from untrained research participants. *Psychiatry Research*, *168*(3), 242–249. https://doi.org/10.1016/j.psychres.2008.05.006

Tuominen, L., Boeke, E., DeCross, S., Wolthusen, R. PF., Nasr, S., Milad, M., … Holt, D. (2019). The relationship of perceptual discrimination to neural mechanisms of fear generalization. *NeuroImage*, *188*, 445–455. https://doi.org/10.1016/j.neuroimage.2018.12.034

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*(8), 1293–1313. https://doi.org/10.3758/BF03194544

Zaman, J., Ceulemans, E., Hermans, D., & Beckers, T. (2019). Direct and indirect effects of perception on generalization gradients. *Behaviour Research and Therapy*, *114*, 44–50. https://doi.org/10.1016/j.brat.2019.01.006

Zaman, J., Struyf, D., Ceulemans, E., Beckers, T., & Vervliet, B. (2019). Probing the role of perception in fear generalization. *Scientific Reports*, *9*(1), 1–8. https://doi.org/10.1038/s41598-019-46176-x

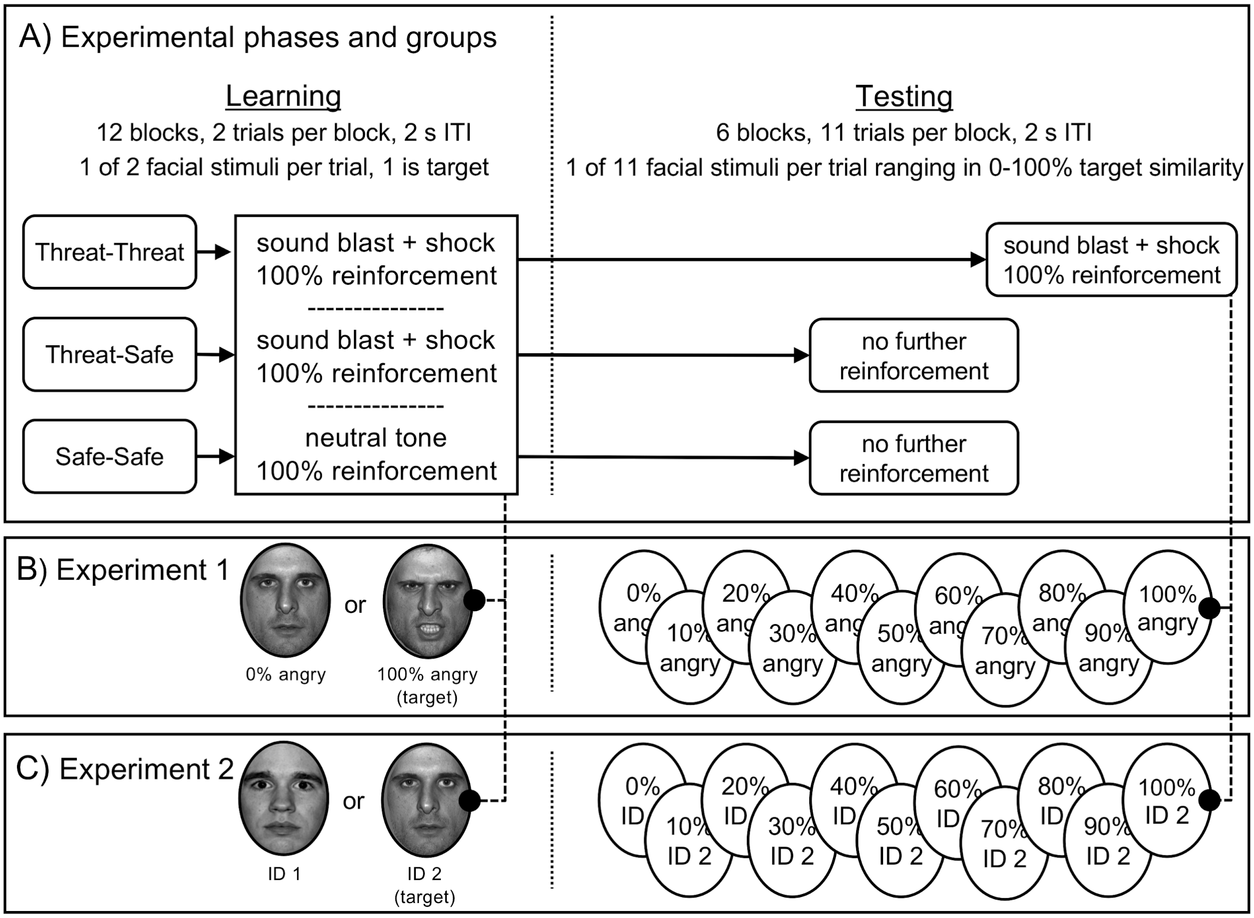

Fig. 1. Schematic of experimental procedure. Both experiments included three groups that varied in the type of outcome paired with the target face across learning and testing phases (A). In the first two blocks of the learning phase, stimuli appeared for 5 s and no response was required. In the remaining 10 blocks of learning, stimuli appeared until participants responded. Experiment 1 used facial stimuli of the same identity that varied in similarity to the target along the dimension of anger intensity (B) and Experiment 2 used facial stimuli with neutral expression that varied in similarity to the target along the dimension of identity (C). This figure shows example images from the NimStim database that are permitted to be published (Tottenham et al., 2009).

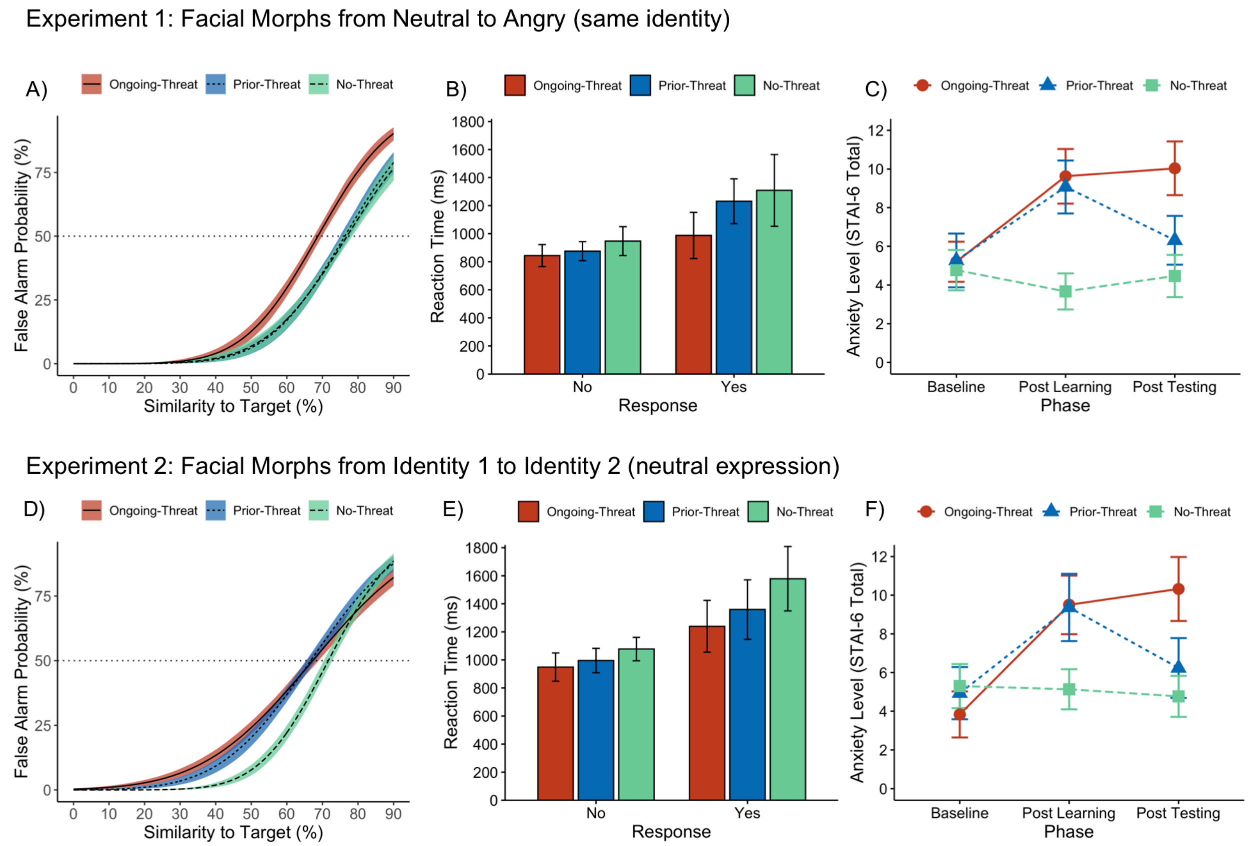

Fig. 2. Subjective, perceptual, and behavioral responses measured across experiments. Mean model estimated probabilities of false alarms with 95% CI as a function of each morph's similarity to the target for Experiment 1 (A) and Experiment 2 (D). Mean response times with 95% CI for *yes* and *no* responses reflecting whether the participant thought the stimulus was identical to the target for Experiment 1 (B) and Experiment 2 (E). Mean anxiety level with 95% CI measured by the STAI-6 across stages of Experiment 1 (C) and 2 (F).

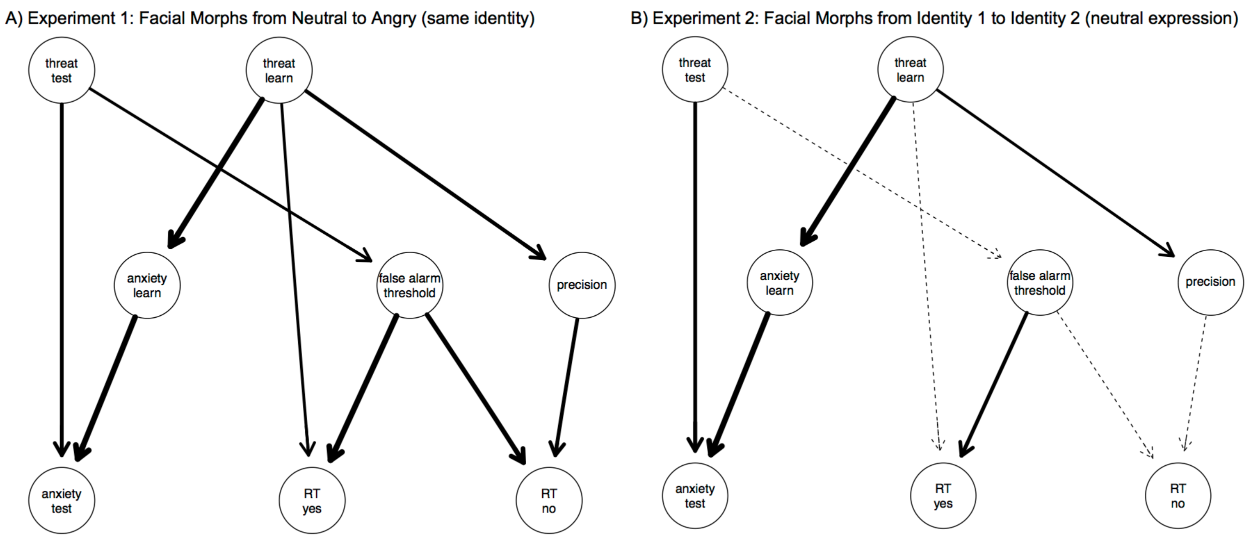

Fig. 3. Directed acyclic graphs depicting potentially causal relations among experimental variables. Line thickness represent the strength of the relationship between variables. Results from Experiment 2 (B) were mapped on to the graph structure from Experiment 1 (A) to facilitate comparison. Dashed lines in (B) denote paths that were not present in Experiment 2.