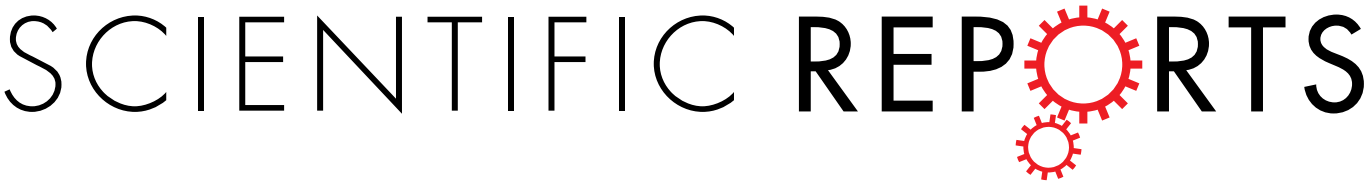

OPEN

Verbal instructions override the meaning of facial expressions

Florian Bublatzky[1,2], Pedro Guerra[3] & Georg W. Alpers[2]

Received: 28 February 2018
Accepted: 17 September 2018
Published online: 09 October 2018

Psychological research has long acknowledged that facial expressions can implicitly trigger affective psychophysiological responses. However, whether verbal information can alter the meaning of facial emotions and corresponding response patterns has not been tested. This study examined emotional facial expressions as cues for instructed threat-of-shock or safety, with a focus on defensive responding. In addition, reversal instructions were introduced to test the impact of explicit safety instructions on fear extinction. Forty participants were instructed that they would receive unpleasant electric shocks, for instance, when viewing happy but not angry faces. In a second block, instructions were reversed (e.g., now angry faces cued shock). Happy, neutral, and angry faces were repeatedly presented, and auditory startle probes were delivered in half of the trials. The defensive startle reflex was potentiated for threat compared to safety cues. Importantly, this effect occurred regardless of whether threat was cued by happy or angry expressions. Although the typical pattern of response habituation was observed, defense activation to newly instructed threat cues remained significantly enhanced in the second part of the experiment, and it was more pronounced in more socially anxious participants. Thus, anxious individuals did not exhibit more pronounced defense activation compared to less anxious participants, but their defense activation was more persistent.

The ability to communicate about future events and their potential consequences is highly advantageous for gaining benefits and avoiding danger. Such vital information can be transmitted using non-verbal communication (e.g., facial expressions, body posture)[1,2], but also via verbal or written instructions (e.g., 'beware of …'). Both sources of information – visual facial expressions and language – have been shown to effectively modulate the activity of motivational systems in the brain, to prepare for adequate responding in a given situation[3–7]. However, to what degree facial expressions and verbal instructions interact in guiding person perception and social behavior is not well-understood.

There is a strong body of research examining the role of facial expressions and their capability to mediate perceptual processing and behavioral responding in social situations. Viewing threat-related emotional expressions – such as fear or anger – has been shown to be associated with enhanced activation of the autonomous nervous system and speeded behavioral responding[8–10]. Similarly, happy facial expressions have been suggested to receive preferential access to attentional processing resources compared to neutral faces. For instance, happy faces have been linked to better detection rates[11,12] and facilitated electrocortical processing (e.g., LPP component)[13]. However, observing unknown people who smile might also be more ambivalent as their actual intention remains uncertain[5,14]. Together, these psychophysiological response patterns have been suggested to reflect the workings of basic motivational systems that organize behavioral approach or withdrawal (e.g., defense behavior)[15,16]. Accordingly, facial expressions of emotion are presumed to be evolutionarily prepared to receive more attentional resources and prime emotion-specific motor-behavioral responding[10,17].

Language is another evolutionary prepared communication system. Affective language, such as insults or compliments, is especially effective at catching the listeners' attention. This is particularly evident when information directly refers to the listener or reader[18–20]. Accordingly, verbal instructions about imminent aversive events (threat-of-shock) effectively enhance perceptual processing[21–23], defensive activation[4,7,24,25], and modulate overt behavioral responding (e.g., in decision-making tasks)[26,27]. Importantly, this verbal information does not need to be substantiated by first-hand experiences of the anticipated aversive events. For instance, despite the lack of aversive reinforcement, instructed threat contingencies are very resistant to extinction even across several

[1]Department of Psychosomatic Medicine and Psychotherapy, Central Institute of Mental Health Mannheim, Medical Faculty Mannheim/Heidelberg University, Mannheim, Germany. [2]Clinical Psychology and Biological Psychology and Psychotherapy, Department of Psychology, School of Social Sciences, University of Mannheim, Mannheim, Germany. [3]University of Granada, Department of Personality, Granada, Spain. Correspondence and requests for materials should be addressed to F.B. (email: florian.bublatzky@zi-mannheim.de)

days[28]. However, verbal instructions can reverse threat expectancies, for instance, when an instructed threat cue is newly learned as a safety cue[29–31]. Given the centrality of threat perception for interpersonal relations and social behavior, associations between threatening events and facial information might be malleable and flexibly change according to social settings.

The present study examined the joint impact of visual and verbal affective information on defensive responding. To this end, pictures of happy and angry facial expressions were verbally instructed as cues for the threat of electric shocks or safety. As dependent variables, we chose both somatic (startle reflex) and autonomic indices (skin conductance response [SCR] and heart rate [HR]), which have been shown to be sensitive to facial expressions and verbal instructions[5,9,24]. In addition to physiological measures, we also obtained subjective ratings about the perceived threat, affective valence, and emotional arousal. Following two previous studies that used pictures of affective scenes as shock cues (i.e., pleasant and unpleasant IAPS pictures)[24,32], threat instructions were predicted to change the inherent valence of facial expressions. This should be evinced by threat-potentiated startle reflex, enhanced SCRs, and potentiated initial HR deceleration, as well as higher threat ratings for threat-of-shock relative to safety cues[4,7,24,25].

Focusing on the interaction of facial emotions and verbal threat/safety instructions, the congruency of affective information (e.g., an angry face instructed to signal shock threat compared to safety) was of particular interest. According to the motivational priming theory[15,16], an interaction of threat/safety instruction by facial expression was expected: When serving as a threat cue, inherently unpleasant stimuli (i.e., angry faces) will elicit more pronounced defensive responding than inherently pleasant stimuli (i.e., happy faces). Alternatively, incongruent information might be particularly effective in guiding defense activation to unknown people. For instance, a smiling person instructed to signal shocks may be considered as particularly dangerous[14], with implications for social interactions and behavior towards this person (e.g., impression formation, social bonding)[33,34].

Also, we expected to gain insight into the malleability of instruction effects by examining reversal learning. Reversal learning reflects the shift of threat associations from one stimulus to another, with the concurrent inhibition (previous threat cue becomes safe) and acquisition of threat contingencies (a previous safety cue becomes threatening)[35]. To this end, a second experimental block was preceded by additional instructions, which aimed at reversing previously learned threat/safety contingencies (e.g., from threat to safety or vice versa)[36,37]. Here, it is of interest whether the impact of reversal instructions on fear extinction learning depends on prepared learning mechanisms in person perception[36]. Specifically, we examined whether threat effects were more stable when angry (relative to happy) faces served as reversed safety cues (i.e., previously cueing threat). Moreover, we predicted that pleasant facial expressions might be less effective as a threat cue[37], or that they are more quickly associated with safety in the reversal test.

Methods

Participants. Sample size was determined using G*Power[38], which indicated that N = 40 was required to detect all relevant physiological effects at a medium effect size (f = 0.25, α error = 0.05, power = 0.8, and assumed correlation of repeated measures = 0.4). This stop rule for data collection was also in line with previous startle studies using emotional facial expression and threat-of-shock instructions[3,24,30]. Forty healthy volunteers (10 males) were recruited from the students of the psychology department at the University of Mannheim. Participants' age ranged between 17 and 52 ($M = 22.7$, $SD = 7.1$) and the sample was within the normal range of state and trait anxiety (STAI, $M = 35.3$ and 35.1, $SD = 5.8$ and 8.8), social anxiety (SPIN, $M = 10.9$, $SD = 8.2$), and depression (BDI, $M = 5.9$, $SD = 6.5$). All participants were informed about the general study procedure before informed consent was obtained. The ethical review committee of the University of Mannheim approved all utilized procedures and methods. Participants received course credits for their participation.

Stimulus materials and presentation. Face pictures were selected from the Karolinska Directed Emotional Faces (KDEF[39]), a well-established stimulus set providing pictures of human facial expressions of emotion. Sixteen actors (eight females) displaying happy, neutral, and angry facial expressions, were selected based on visual inspection (i.e., seven raters agreed upon the clarity and recognizability of facial expressions). The KDEF face identifiers were af01, af07, af09, af11, af19, af20, af22, af29, am02, am03, am07, am08, am10, am13, am14, and am25.

All 48 pictures (1024 × 768 pixels) were presented for 6 s separated by variable inter-trial intervals (ITI) ranging from 10 to 15 s to allow response recovery (see Fig. 1). To provoke the defensive startle reflex, auditory startle probes (white noise, 105 dB, 50 ms) were presented during half of the picture trials. The 48 pictures (including the 24 picture-startle trials) were evenly distributed across two experimental blocks (instantiation, reversal) and three facial expressions (happy, neutral, angry), resulting in four picture-startle trials for each experimental condition per participant. To prevent the predictability of auditory stimulation, startle probes were presented at either 4, 4.5, 5 or 5.5 s after picture onset (i.e., while the picture was still visible), and six additional startle probes (three per block) were presented during the ITI. Startle probes were presented binaurally using headphones (AKG K44 Perception) and the average lag between probes was 28.8 s.

Presentation software (Neurobehavioral Systems, Inc., Albany, CA, USA) served to control the stimulus presentation, which was pseudorandom regarding picture sequence (no immediate repetition of the same face actor, no more than three pictures of the same facial expression in a row) and regarding startle presentation (no more than two picture-startle trials in a row). Electric stimuli for the shock work-up procedure were presented using a Digitimer Stimulator DS-5 (up to 10 shocks, with maximal 10 mA, 100 ms).

Experimental task and instructions. Participants' task was to look at all pictures, which were presented during the two experimental blocks (instantiation and reversal; see Fig. 1). Immediately before the first block started (instantiation), participants were verbally instructed that they might receive up to three electric shocks

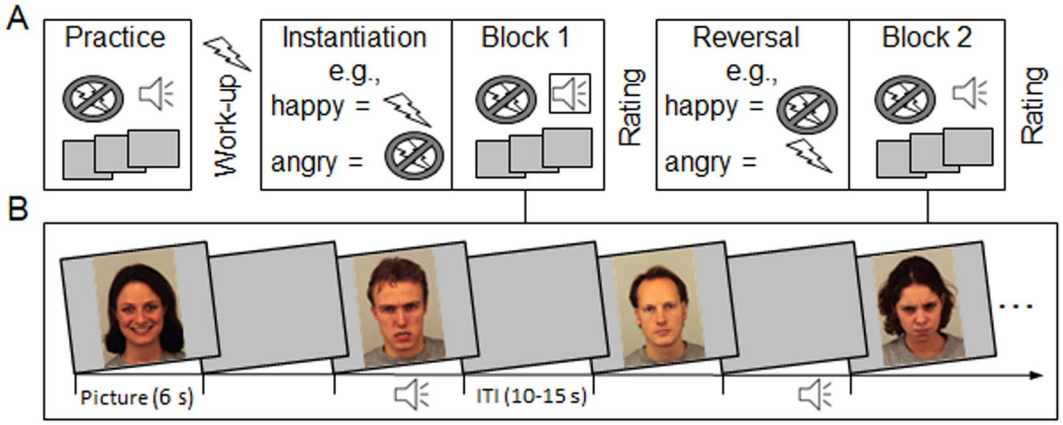

Figure 1. Schematic illustration of the experimental procedures and stimulus presentation. (**A**) After a brief practice run and shock work-up procedure, participants were verbally instructed that one particular emotional facial expression serves as a cue for threat-of-shock (e.g., happy) or safety (e.g., angry) and the first experimental block started (instantiation). Preceding the second experimental block (reversal), a verbal reversal instruction stated that now threat and safety contingencies are reversed (e.g., now angry faces cue threat and happy cue safety). The order in which facial expressions cued threat or safety was tested in two groups (each N = 20 completed the happy-angry or angry-happy threat order). Please note, neutral faces always cued safety. Following each block, threat and safety cues were rated regarding valence, arousal, and perceived threat. (**B**) Within each block, face pictures displaying happy, neutral, and angry facial expressions were presented (each 6 s) with a variable intertrial interval (ITI, 10 to 15 s). Auditory startle probes were presented occasionally during pictures and ITIs, no shocks were presented during the experiment. Example pictures are taken from the KDEF (identifiers: af01has, am08ans, am10nes, and af20ans).

under specific conditions. One group of participants (N = 20) was told that electric shocks might be administered whenever an angry face is presented (angry = threat) but not when they see a happy face (happy = safety). The other group (N = 20) received the opposite instruction, stating that happy facial expressions cued threat-of-shock (happy = threat), and safety condition being signaled by angry faces (angry = safety). For the second experimental block (reversal), all participants were verbally instructed that now threat and safety contingencies were reversed. Specifically, the previous threat cue becomes safe, and the previous safety cue becomes threatening. Thus, across both experimental groups, happy and angry facial expressions served equally often as instructed and reversed threat and safety cue; neutral faces always signaled safety.

Procedure. Participants completed questionnaires on general and social anxiety and depression (State-Trait Anxiety Inventory [STAI-state/trait], Social Phobia Inventory [SPIN], Social Interaction Anxiety Scale [SIAS], Beck Depression Inventory [BDI]). Sensors for physiological recordings were attached, and an electric stimulation electrode was placed at the right upper arm. Next, a brief shock work-up procedure (without picture presentation) was carried out to ensure the credibility of the threat instruction[22,40]. To set the shock intensity individually at a level rated as "maximally unpleasant but not yet painful", participants received up to 10 shocks with increasing intensity. Participants were then instructed that the intensity of the electric shocks given during the experiment would be equal to the most unpleasant test stimulus.

Practice trials served to familiarize participants with the picture and startle presentation procedure and to allow for initial habituation of the startle reflex. Afterward, verbal instructions regarding threat and safety contingencies were given (i.e., which facial expression signals threat-of-shock and which signals safety) and the first experimental block started (instantiation). Following this block, participants rated the hedonic valence and arousal using the Self-Assessment Manikin (SAM)[41], and perceived threat of the facial expressions using a visual analog scale ranging from *not at all* to *highly threatening* (1 to 10). Then all participants received the instruction that threat/safety contingencies were now reversed (e.g., the threat cue becomes safe, and safety cue becomes threatening), and the second block started (reversal). Facial expressions were rated again after the reversal block. Finally, participants were debriefed. No shocks were presented during the experiment. Thus, results reflect physiological responding during the anticipation (but not experience) of electric shocks.

Data recording and reduction. Psychophysiological measures were recorded continuously with a vAmp amplifier (BrainProducts, Munich, Germany). Startle amplitudes were derived from the electromyogram of the orbicularis muscle using two miniature Ag/AgCl electrodes. The raw signal was recorded at a 1000 Hz sampling rate and frequencies below 28 Hz and above 500 Hz were filtered out with a band-pass filter (24 dB/octave roll-off). Raw electromyogram (EMG) data were rectified and smoothed with a moving average procedure (50 ms) in VisionAnalyzer 2.0 (BrainProducts). Startle responses were scored with an automated procedure and defined as the maximum peak in the 21–150 ms time window following each startle probe. Peak amplitudes were calculated relative to a mean baseline period (50 ms preceding startle response time window)[28,42].

As an index of phasic autonomic activation, skin conductance responses (SCRs) were recorded with Ag/AgCl electrodes (constant voltage of 0.5 V; 20 Hz sampling rate) placed at the hypothenar eminence of the non-dominant hand. SCRs to picture onset were calculated as the maximum increase in skin conductance in the interval of 1 to 6 s (relative to a 1 s pre-stimulus period). A minimum threshold of 0.02 μS was used for zero-response detection, and range and distribution corrections were applied. Phasic heart rate changes to picture onset was derived from the electrocardiogram recorded at lead II. The electrocardiogram signal was recorded at 1000 Hz, and frequencies below 0.1 and above 13 Hz were filtered. The weighted HR averages every half second were expressed in terms of differential scores with respect to a 2 s baseline period[24].

Data analysis and statistical design. Self-report (valence, arousal, and threat ratings) and physiological data (startle-EMG and SCR) were submitted to (2 × 2) × 2 repeated measures ANOVA, including the within-subject factors Instruction (threat vs. safety) and Block (instantiation Block 1 vs. reversal Block 2), as well as the between-group factor Order (happy-angry vs. angry-happy). The Order referred to the sequence in which emotional facial expression cued threat or safety in which experimental block. Specifically, for the happy-angry order, happy faces served as threat cue during instantiation block (Block 1), and angry faces cued threat during the following reversal block (Block 2). This was reversed for the angry-happy order, in which angry faces during Block 1 and happy faces in Block 2 cued threat-of-shock. Regarding phasic changes in heart rate, an additional factor, Time (12), was implemented to compare half-second changes after picture onset.

To examine the impact of emotional facial expression on the instantiation and reversal of threat instructions, planned comparisons focused separately on each Order (happy-angry vs. angry-happy). Please note that for reasons of brevity and to reduce the complexity of the overall design, neutral faces cued safety in both blocks and were thus excluded from the analyses of instructed and reversed threat. However, supplementary analyses were conducted to compare Facial expression (happy vs. neutral vs. angry) when serving as a safety cue (see Supplemental Material). Covariation analyses were conducted to test the impact of inter-individual differences in reported social- and trait-anxiety on defense activation.

Greenhouse-Geisser corrections were used where relevant, and the partial η^2 is reported as a measure of effect size. To control for Type 1 error, Bonferroni correction was applied for post-hoc *t*-tests.

Results

Self-report data. *Threat ratings.* Overall, instructed threat cues were rated as more threatening relative to safety cues, $F(1,39) = 17.37$, $p < 0.001$, $\eta_p^2 = 0.31$, and perceived threat decreased from the instantiation block to the reversal block, $F(1,39) = 5.98$, $p < 0.05$, $\eta_p^2 = 0.13$. The interaction Instruction × Block did not reach significance, $F(1,39) = 1.47$, $p = 0.23$, $\eta_p^2 = 0.04$, however, a significant three-way interaction Instruction × Block × Order emerged, $F(1,38) = 56.72$, $p < 0.001$, $\eta_p^2 = 0.60$.

Follow-up analyses run separately for the Happy-Angry order (see Fig. 2A; Table 1) showed that instructed threat effects varied across blocks, $F(1,19) = 21.26$, $p < 0.001$, $\eta_p^2 = 0.53$. Specifically, pronounced threat ratings were observed for happy facial expressions cueing threat-of-shock during instantiation, $F(1,19) = 5.23$, $p < 0.05$, $\eta_p^2 = 0.22$, and for angry faces cueing threat in the subsequent reversal block, $F(1,19) = 31.30$, $p < 0.001$, $\eta_p^2 = 0.62$. Similarly, for the Angry-Happy order, an interaction Instruction × Block emerged $F(1,19) = 35.46$, $p < 0.001$, $\eta_p^2 = 0.65$. Separate comparisons of threat and safety cues indicate more pronounced threat ratings for angry faces during the instantiation block, $F(1,19) = 35.04$, $p < 0.001$, $\eta_p^2 = 0.65$, and happy expressions in the reversal block, $F(1,19) = 16.50$, $p < 0.01$, $\eta_p^2 = 0.47$.

Valence ratings. Overall, threat cues were rated as more unpleasant than safety cues, $F(1,39) = 11.01$, $p < 0.01$, $\eta_p^2 = 0.22$, and unpleasantness decreased across blocks, $F(1,39) = 4.62$, $p < 0.05$, $\eta_p^2 = 0.11$. Whereas no interaction of Instruction × Block was observed, $F(1,39) = 1.99$, $p = 0.17$, $\eta_p^2 = 0.05$, a significant three-way interaction was found, Instruction × Block × Order $F(1,38) = 60.87$, $p < 0.001$, $\eta_p^2 = 0.62$.

Separate analysis for the Happy-Angry order (Fig. 2B) showed a significant interaction Instruction × Block, $F(1,19) = 13.53$, $p < 0.01$, $\eta_p^2 = 0.42$. Interestingly, happy expressions were rated as more pleasant than angry faces even when happy faces served as instructed threat cue in the instantiation block, $F(1,19) = 11.01$, $p < 0.01$, $\eta_p^2 = 0.37$. In the reversal block, angry faces cueing threat were rated as more unpleasant than happy faces cueing safety, $F(1,19) = 11.56$, $p < 0.01$, $\eta_p^2 = 0.38$. For the Angry-Happy order, a significant interaction Instruction × Block was evident, $F(1,19) = 58.77$, $p < 0.001$, $\eta_p^2 = 0.76$. Angry faces cueing threat were more unpleasant during the instantiation block, $F(1,19) = 72.96$, $p < 0.001$, $\eta_p^2 = 0.79$, and this threat effect was less pronounced when happy facial expressions served as new threat cues in the reversal block, $F(1,19) = 11.28$, $p < 0.01$, $\eta_p^2 = 0.37$.

Arousal ratings. Overall, instructed threat cues were rated as more arousing than safety cues, $F(1,39) = 10.41$, $p < 0.01$, $\eta_p^2 = 0.21$, and arousal decreased from the instantiation to the reversal block, $F(1,39) = 21.90$, $p < 0.001$, $\eta_p^2 = 0.36$. Moreover, the interaction of Instruction × Block was significant, $F(1,39) = 4.83$, $p < 0.05$, $\eta_p^2 = 0.11$, showing pronounced threat effects in the instantiation block, $F(1,39) = 13.52$, $p < 0.01$, $\eta_p^2 = 0.26$, but not in the reversal block, $F(1,39) = 1.05$, $p = 0.31$, $\eta_p^2 = 0.03$. This pattern did not significantly differ between experimental orders (Angry-Happy or Happy-Angry; Fig. 2C), Instruction × Block × Order $F(1,38) = 2.91$, $p = 0.10$, $\eta_p^2 = 0.07$.

Startle reflex. The defensive startle reflex was more pronounced when viewing threat as compared to safety cues (Fig. 3A), $F(1,39) = 56.87$, $p < 0.001$, $\eta_p^2 = 0.59$, and a pronounced pattern of response habituation was observed across experimental blocks (Fig. 4A), $F(1,39) = 58.19$, $p < 0.001$, $\eta_p^2 = 0.60$. Moreover, reflex amplitudes varied as a function of Instruction × Block, $F(1,39) = 4.26$, $p < 0.05$, $\eta_p^2 = 0.10$, indicating that threat-potentiation decreased from the instantiation to the reversal block. Post-hoc tests revealed pronounced differences between threat and safety cues within the instantiation block, and less markedly but still highly significant in the following

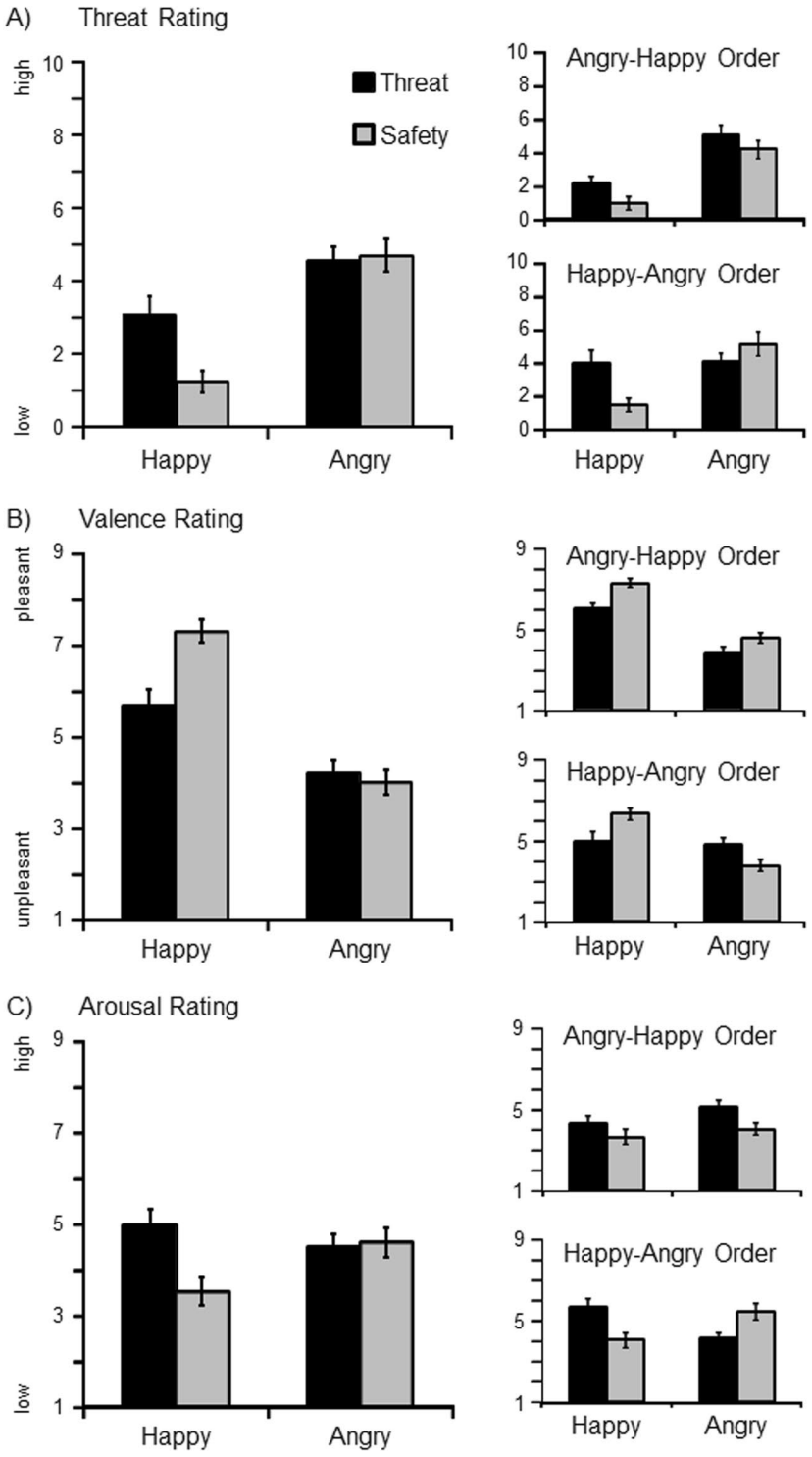

Figure 2. Self-reported threat (**A**), valence (**B**), and arousal (**C**) ratings as a function of facial expression (happy, angry) and instructions (threat, safety). The left side illustrates overall means (with SEM) averaged across experimental blocks and orders. On the right side, separate means are plotted for each order. The angry-happy order started with angry facial expression cueing threat during the instantiation block and happy faces cueing threat during the following reversal block. For the happy-angry order, instructed threat/safety contingencies were vice versa.

reversal block, $Fs(1,39) = 48.78$ and 16.52, $ps < 0.001$, $\eta_p^2 = 0.56$ and 0.30. Importantly, the inherent valence of emotional facial expressions did not modulate the instantiation and reversal of threat as observed for the startle reflex, Order × Instruction × Block, $F(1,38) = 0.73$, $p = 0.40$, $\eta_p^2 = 0.02$. Planned follow-up tests focused on each experimental order separately.

Block	Instruction	Order	Startle		SCR		HR		Valence		Arousal		Threat	
			M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Instantiation (Block 1)	Threat	Angry-Happy	57.85	5.14	0.115	0.11	−2.75	1.66	3.55	1.73	5.15	1.66	5.05	2.61
		Happy-Angry	57.48	5.31	0.097	0.17	−2.75	2.71	5.05	2.48	5.80	2.14	4.05	3.61
	Safe	Angry-Happy	49.54	3.31	0.030	0.04	−0.78	1.85	7.90	1.25	3.30	2.13	0.95	1.76
		Happy-Angry	49.77	3.60	0.032	0.07	−1.30	4.36	3.50	1.67	5.45	2.19	5.20	3.08
Reversal (Block 2)	Threat	Angry-Happy	50.26	5.99	0.048	0.10	−1.64	3.28	6.30	1.75	4.15	2.08	2.15	1.95
		Happy-Angry	51.21	5.44	0.079	0.24	−2.58	2.65	4.85	1.69	3.90	1.71	4.10	2.38
	Safe	Angry-Happy	46.58	4.22	0.014	0.02	−1.33	2.62	4.50	1.54	3.75	1.62	4.23	2.46
		Happy-Angry	45.41	4.32	0.034	0.05	−1.29	2.31	6.70	1.75	3.75	1.89	1.55	1.79

Table 1. Mean amplitudes and standard deviations (*M*, *SD*) as a function of Block (instantiation vs. reversal), Instruction (threat vs. safety), and Order of threat instruction (Angry-Happy vs. Happy-Angry). Means are provided for the startle reflex, skin conductance responses (SCR), heart rate (HR), and ratings of the self-reported valence, arousal, and perceived threat. Heart rate changes refer to averages across the significant time intervals from 3 to 5 s after picture onset.

When angry faces signaled threat during the instantiation block and served as safety cue in the subsequent reversal block (Angry-Happy order), main effects of Instruction and Block were significant, $Fs(1,19) = 23.67$ and 49.31, $ps < 0.001$, $\eta_p^2 = 0.56$ and 0.72. Moreover, startle amplitudes tended to vary as a function of Instruction × Block, $F(1,19) = 3.38$, $p = 0.08$, $\eta_p^2 = 0.15$. Separate analyses for each block revealed that angry faces as threat cue triggered highly significant threat effects during instantiation, $F(1,19) = 24.66$, $p < 0.001$, $\eta_p^2 = 0.57$, but only marginal threat effects were observed when angry faces served as safety cue in the reversal block, $F(1,19) = 4.02$, $p = 0.06$, $\eta_p^2 = 0.17$. In contrast, for the Happy-Angry order, when happy faces served as threat cue during instantiation and as safety cue in the reversal block, main effects of Instruction and Block were found, $Fs(1,19) = 19.97$ and 32.59, $ps < 0.001$, $\eta_p^2 = 0.51$ and 0.63. However, threat effects were not reduced across blocks when angry faces cued shock threat in the reversal block, Instruction × Block $F(1,19) = 0.96$, $p = 0.34$, $\eta_p^2 = 0.05$. Follow-up tests revealed pronounced threat-potentiated startle for happy faces cueing threat in the instantiation block, $F(1,19) = 22.95$, $p < 0.001$, $\eta_p^2 = 0.55$, which was similarly pronounced in the subsequent reversal block when angry faces cued threat, $F(1,19) = 16.1$, $p = 0.001$, $\eta_p^2 = 0.46$. Thus, during the reversal block, instruction effects were more resistant to extinction when angry rather than happy faces cued threat.

Exploratory analyses revealed that the overall interaction Instruction × Block varied as a function of inter-individual differences in reported social- and trait-anxiety (Fig. 4B). Specifically, significant covariation effects were observed with SPIN scores, $F(1,38) = 8.15$, $p < 0.01$, $\eta_p^2 = 0.18$, STAI-trait, $F(1,38) = 7.81$, $p < 0.01$, $\eta_p^2 = 0.17$, and marginally with SIAS, $F(1,38) = 3.71$, $p = 0.06$, $\eta_p^2 = 0.09$. To follow up on these interactions, correlational analyses were conducted between anxiety scores and startle amplitudes (i.e., the difference scores between threat minus safety) separately for each block. For the instantiation block, threat effects did not vary with anxiety level ($r_{\text{trait-anxiety}} = -0.20$, $p = 0.21$; $r_{\text{SPIN}} = -0.25$, $p = 0.12$; $r_{\text{SIAS}} = -0.08$, $p = 0.61$). In the subsequent reversal block, however, threat-potentiated startle was more pronounced in participants who scored higher on anxiety ($r_{\text{trait-anxiety}} = 0.36$, $p < 0.05$; $r_{\text{SPIN}} = 0.32$, $p < 0.05$; $r_{\text{SIAS}} = 0.32$, $p < 0.05$). Thus, anxious participants did not exhibit more pronounced, but more persistent defense activation compared to less socially anxious participants.

Skin conductance responses. Enhanced skin conductance responses (SCR) were observed for threat relative to safety cues, $F(1,39) = 9.29$, $p < 0.01$, $\eta_p^2 = 0.19$ (see Fig. 3B). SCRs diminished over time, they were more pronounced in the instantiation block than in the subsequent reversal block, $F(1,39) = 7.21$, $p < 0.05$, $\eta_p^2 = 0.16$. The interaction Instruction × Block didn't reach significance, $F(1,39) = 2.95$, $p = 0.09$, $\eta_p^2 = 0.07$. Exploratory follow-up analyses revealed significant threat effects during instantiation, $F(1,39) = 20.93$, $p < 0.001$, $\eta_p^2 = 0.35$, but not in the reversal block, $F(1,39) = 2.31$, $p = 0.14$, $\eta_p^2 = 0.06$. Importantly, the inherent facial valence did not modulate SCRs for instantiation and reversal of threat, Order × Instruction × Block, $F(1,38) = 0.53$, $p = 0.47$, $\eta_p^2 = 0.01$. Planned comparisons focused separately on each experimental order.

When angry faces served initially as threat cues, and later as safety cues (Angry-Happy order), there were significant main effects of Instruction, $F(1,19) = 13.34$, $p < 0.01$, $\eta_p^2 = 0.41$, and Block, $F(1,19) = 8.43$, $p < 0.01$, $\eta_p^2 = 0.31$, as well as a trend to an interaction Instruction × Block, $F(1,19) = 3.55$, $p = 0.08$, $\eta_p^2 = 0.16$. Follow-up analyses revealed threat-enhanced SCRs when angry faces cued threat during the instantiation block, $F(1,19) = 15.15$, $p = 0.001$, $\eta_p^2 = 0.44$, but not when happy faces cued threat in the reversal block, $F(1,19) = 2.63$, $p = 0.12$, $\eta_p^2 = 0.12$. For the Happy-Angry order, in contrast, SCRs did not differ for Instruction or Block, $Fs(1,19) = 2.55$ and 0.53, $ps = 0.13$ and 0.48, $\eta_p^2 = 0.21$ and 0.03. Whereas no interaction of Instruction × Block was found, $F(1,19) = 0.41$, $p = 0.53$, $\eta_p^2 = 0.02$, exploratory analyses indicated threat-enhanced SCRs to happy faces cueing threat during the instantiation, $F(1,19) = 6.85$, $p < 0.05$, $\eta_p^2 = 0.27$, but not when threat was cued by angry faces in the subsequent reversal block, $F(1,19) = 0.79$, $p = 0.38$, $\eta_p^2 = 0.04$. No covariation effects were observed between SCRs and anxiety scores.

Phasic heart rate changes. Overall, heart rate revealed a significant deceleration when viewing threat relative to safety cues, $F(1,39) = 4.03$, $p = 0.05$, $\eta_p^2 = 0.09$ (see Fig. 3C). Furthermore, there was an interaction of Time × Instruction, $F(11,429) = 5.48$, $p < 0.01$, $\eta_p^2 = 0.12$. Follow-up analyses were calculated separately for each time interval and indicated significant heart rate deceleration for threat relative to safety cues between 3 and 5 s

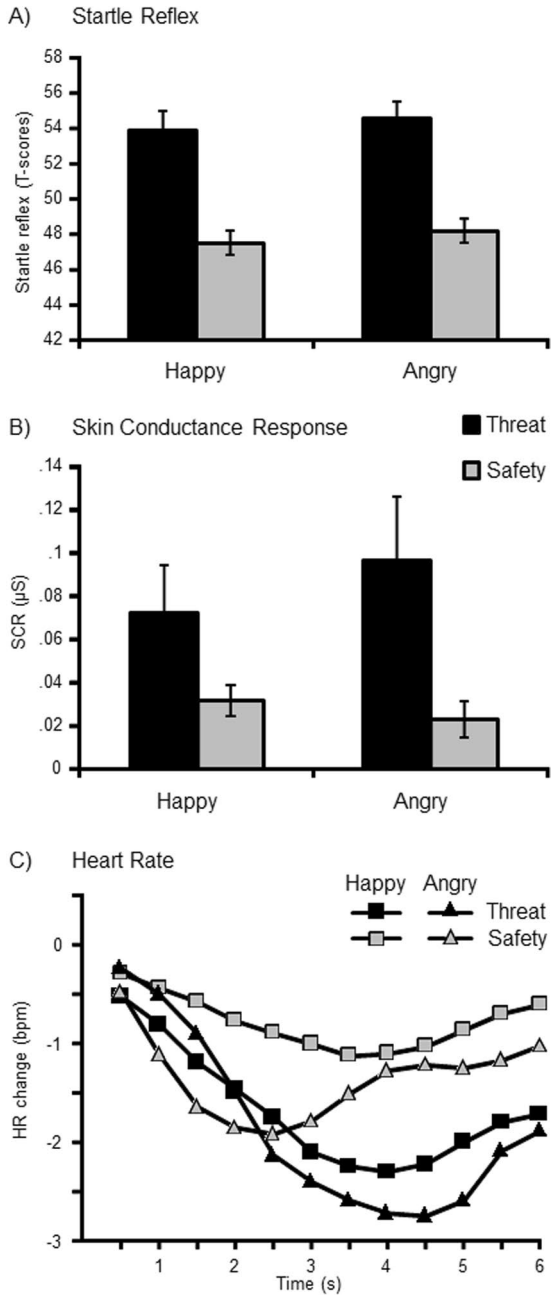

Figure 3. Mean responses of the startle reflex (**A**) and skin conductance (**B**) for happy and angry facial expressions serving as threat or safety cue (with SEM). Heart rate changes (**C**) are averaged every half a second after stimulus onset. As no interaction effects occurred with the sequence of instructions, averages across experimental orders (happy-angry and angry-happy) are illustrated.

after picture onset (all $ps < 0.05$). Neither the main effect Block, $F(1,39) < 0.01$, $p = 0.98$, $\eta_p^2 < 0.01$, nor the interactions Instruction × Block, $F(1,39) = 0.78$, $p = 0.38$, $\eta_p^2 = 0.02$, Time × Instruction × Block, $F(11,429) = 1.04$, $p = 0.38$, $\eta_p^2 = 0.03$, nor the four-way interaction by Order reached significance, $F(11,418) = 0.65$, $p = 0.58$, $\eta_p^2 = 0.02$.

Exploratory analyses focused separately on the different experimental orders. For the Angry-Happy order, when angry faces served as a threat cue in the instantiation block, a substantial threat deceleration was evident, $F(1,19) = 8.98$, $p < 0.01$, $\eta_p^2 = 0.32$, which developed over time following the threat cue onset, $F(11,209) = 5.22$, $p < 0.01$, $\eta_p^2 = 0.22$. However, no threat effects emerged in the subsequent reversal block when happy faces served as new threat cue, $F(1,19) = 0.01$, $p = 0.93$, $\eta_p^2 < 0.01$. For the Happy-Angry order, in contrast, no threat effect occurred, $F(1,19) = 2.62$, $p = 0.12$, $\eta_p^2 = 0.12$, nor did any interaction including Instruction reach significance, $Fs < 1.68$, $ps > 0.20$, $\eta_p^2 < 0.08$. No covariation effects were found between phasic heart rate changes and anxiety scores.

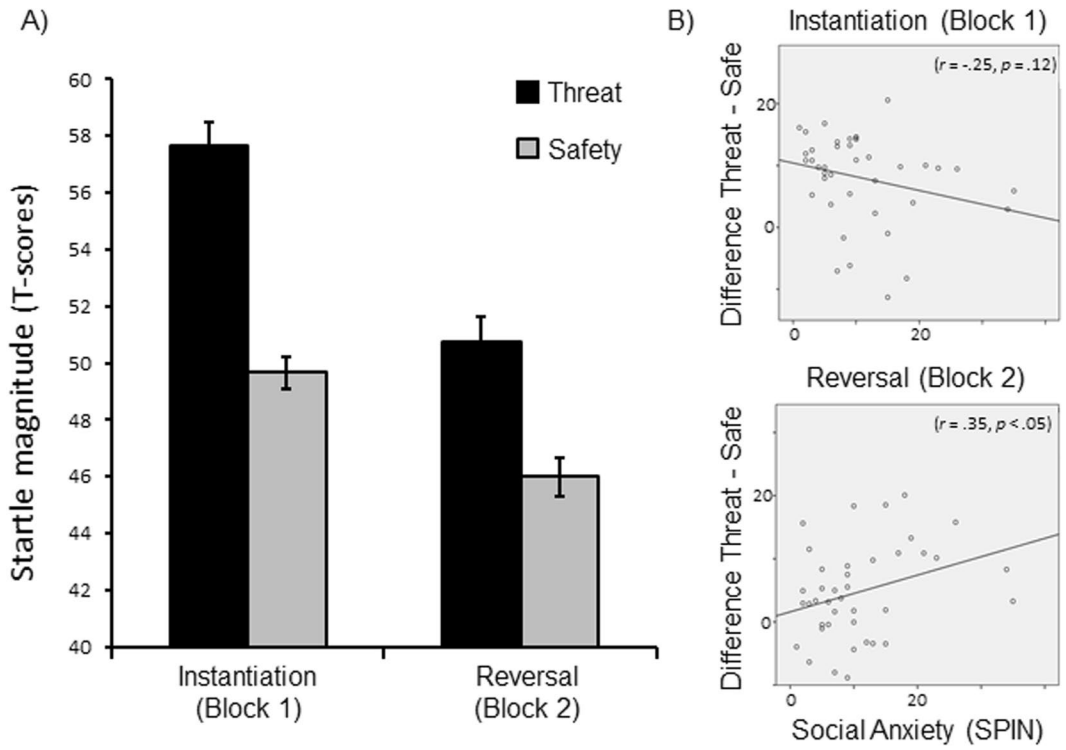

Figure 4. (**A**) Mean startle responses as a function of threat/safety instructions averaged across experimental blocks and orders (with SEM). Threat and safety contingencies were instantiated in Block 1 (e.g., angry faces cue threat) and reversed in Block 2 (e.g., happy faces cue threat; or vice versa). (**B**) Scatterplots illustrate the covariation between individuals' social anxiety and threat-potentiated startle effects (differences between threat and safety) separately for the instantiation and reversal block. Overall, threat effects are malleable and stable; anxious participants reveal a more persistent pattern of defense activation after reversal instructions.

Discussion

The present study examined the capability of emotional facial expressions as cues for verbally instructed threat-of-shock or safety. Also, we tested the flexibility of threat and safety associations using reversal instructions. Verbal communication about threat contingencies triggered, as expected, a pronounced pattern of psychophysiological defense reactions. This was evident in potentiated eye-blink startle reflex, enhanced skin conductance responses, and heart rate deceleration. For self-report data, interaction effects of facial expressions and verbal instructions emerged. Specifically, when smiling faces cued threat, they were rated as aversive as angry faces. In contrast, physiological responding was independent of whether the threat was cued by a happy or an angry facial expression. Moreover, reversal instructions flexibly changed defense activation, leading to relatively stable threat effects despite substantial response habituation across the experimental blocks. Interestingly, after reversal instructions, threat-potentiated startle was more pronounced in more socially anxious participants. Thus, anxious individuals did not exhibit more pronounced defense activation compared to less anxious participants, but their defense activation was more persistent.

When confronting other people's facial expressions, which were previously learned as signals for shock threat, pronounced activation of the somatic and autonomic nervous system was observed (i.e., potentiated eye-blink startle and enhanced skin conductance responses). These findings replicate previous studies showing defense activation to visual stimuli cueing instructed threat-of-shock[24–29]. Defensive responding, however, occurred regardless of whether the threat was cued by a smiling or an angry face. Thus, the inherent valence of the threat cue (happy or angry expression) was not relevant for the acquisition of threat contingencies. This finding adds to previous research using the threat-of-shock paradigm with complex natural affective scenes[4,7,24]. For example, Bradley and colleagues[24] observed comparable threat-potentiated startle reflex to pleasant and unpleasant pictures when these served as instructed threat cues. Moreover, when pictures were not predictive for threat-of-shock (i.e., presented within a threatening context), threat effects were found similarly pronounced for pleasant, neutral, and unpleasant pictures[4,43]. The present study extends this view to face and person perception and shows that the emotional salience of happy and angry facial expressions can be easily overridden by verbal instructions about threat contingencies. This finding contributes to the rather mixed evidence on whether human faces serve as an evolutionary prepared conditional stimulus[36,37,44,45]. Compared to pictures of snakes or spiders, the human face may be a less reliable source of threat or safety information, probably because facial expressions can be manipulated consciously and are subject to social display rules[46].

The inherent valence of an emotional face did not interact with the verbally transmitted acquisition of threat or safety contingencies. This finding is supported by previous neuroimaging research, for instance, showing that threat instructions led to a more general sensitization of stimulus processing[32,47], regardless of the a priori

meaning of a shock cue (e.g., unpleasant or neutral social scenes). Moreover, in the present study, neither the somatic (eye-blink startle) nor the autonomic nervous system (SCR and phasic HR responses) showed an interaction between visual and instructed information. Supplementary analyses using Bayesian statistics supported these findings. Likelihood estimates of the null hypothesis (i.e., no Order × Instruction × Block interaction) indicated that the null relative to the alternative hypotheses were around 19-, 37-, and 142-times more likely for the startle reflex, SCR, and HR measures respectively. Thus, the present data lend support for the notion that the processing of visual and verbal threat information is organized in (partially) distinct neural circuitry. For example, affective modulation of the startle reflex triggered by emotional pictures is impaired in patients with right rather than left temporal lobectomy, whereas the opposite pattern can be observed when instructed threat cues are presented[48]. Interestingly, our self-report data revealed result patterns that were partly in contrast to physiological measures. Valence, arousal, and threat ratings confirmed that verbal communication about potential threats clearly induced aversive anticipations. Moreover, these threat/safety contingencies were highly malleable and reversible using subsequent instructions. In contrast to physiology, however, rating data showed that the impact of threat and safety instructions varied with the inherent valence of the facial threat/safety cue. When cueing threat, a smiling face becomes as aversive as an angry face, and both cues were highly effective in triggering defensive responding to cope with the anticipated aversive event.

Overall, reversal instructions flexibly changed threat/safety associations and the corresponding physiological response patterns. In line with previous studies, verbal instructions were highly effective at reducing defensive responding using reversing affective contingencies from threat to safety[29–31]. Similarly effective was the reversal of contingencies from safety to threat. Newly learned threat cues (previously safe), compared to the newly learned safety cues (previously threatening), were associated with lower valence and higher threat ratings. Moreover, potentiation of the startle reflex was observed for the new threat cues despite pronounced response habituation across experimental blocks. This result adds to the findings of previous research, which show that instructed threat effects may be highly persistent within and across repeated sessions, even without any aversive reinforcement[4,28].

Interestingly, after reversal instructions, threat effects varied as a function of social and trait anxiety. Specifically, anxious participants did not exhibit more pronounced defense activation compared to less socially anxious participants but did exhibit a more persistent defense activation. From a clinical perspective, this is an important finding showing that inter-individual differences in anxiety might account for the capability to learn new safety contingencies and to reduce psychophysiological defense activation. As many fears and anxieties rely on aversive anticipations rather than experiences, the mere absence of aversive events or omission of reinforcement is not sufficient for successful fear extinction learning (e.g., in generalized anxiety disorder or social phobia)[49–51]. To optimize social communication about threats and safety in a therapeutic context, different means of social learning need to be accounted for (i.e., learning by instructions and observations)[25,52]. Building upon the present inter-individual differences in reversal learning, testing (sub-)clinical samples high in social anxiety or interpersonal disturbances might be particularly informative[53,54]. Here, the implementation of a full reversal design[29,35] might focus on safety learning and elucidate the impact of reversed compared to maintained social safety cues.

Several noteworthy aspects of the present design and findings need to be acknowledged and should be addressed in future research. Exploratory analyses provided some indication for the hypothesis that facial emotions might differentially modulate reversal learning. Specifically, for the startle reflex during the reversal block, instruction effects were more resistant to extinction when angry rather than happy faces cued threat. This finding might result from anger-superiority in threat learning (i.e., angry faces more readily associated with threat)[9,10] or happy-superiority[11,12] in safety learning. For directly comparing these opposing hypotheses, the use of a non-affective threat cue condition would have been useful (i.e., neutral faces cueing threat during reversal block) and cannot be resolved with the data at hand. Future research could examine the capability of distinct non-affective social stimuli as reversed threat/safety cue. For instance, invariant facial features – such as person identity and the color of the skin – have been shown to be powerful factors that bias threat learning and can be pitted against each other (e.g., viewing other-race, but same team faces)[34,36,45]. Here, social approaches to initiate persistent reversal learning (i.e., shifting aversive contingencies to other non-social cues) may help to counteract stereotypes, social avoidance, and ostracism[35,55]. From an evolutionary perspective, it appears likely that combined variant and invariant facial information (e.g., facial expression and person identity cues)[56] critically guide behavioral responding. For instance, an angry looking out-group member or a smiling mother might be more readily learned as a signal for threat or safety; such congruency effects in prepared learning can be tested with personalized stimulus materials (e.g., pictures of attachment figures)[57,58]. Finally, the transfer to behavioral output measures appears pertinent to test the implications and consequences of threat and safety learning in social interaction situations, for instance, regarding interpersonal trust[59], stereotyping and social group biases[33,34], or choice behavior in clinical settings (e.g., decision to undergo treatment)[54,60].

In summary, verbal communication about threats might easily prime defensive response programs regardless of the inherent valence of the threat or safety cue (i.e., happy or angry facial expression). Moreover, threat effects were malleable by additional verbal instructions, and the persistence of threat effects varied with inter-individual differences in social and trait anxiety. Anxious participants did not exhibit more pronounced defense activation compared to less anxious participants but did exhibit more persistent defense activation. As threat instructions were not substantiated by the individual's own experiences (i.e., no shocks during the experiment), these findings demonstrate the effects of mere anticipatory processes in person perception relevant to maladaptive extinction learning in anxiety disorders.

Data Availability

The datasets generated and analyzed during the current study are available from F.B on request.

References

1. De Gelder, B. Towards the neurobiology of emotional body language. *Nature Reviews. Neuroscience*, **7**(3), 242, doi:1038/nrn1872 (2006).
2. Ekman, P. Facial expression and emotion. *American Psychologist* **48**(4), 384, https://doi.org/10.1037/0003-066X.48.4.384 (1993).
3. Alpers, G. W., Adolph, D. & Pauli, P. Emotional scenes and facial expressions elicit different psychophysiological responses. *International Journal of Psychophysiology* **80**(3), 173–181, https://doi.org/10.1016/j.ijpsycho.2011.01.010 (2011).
4. Bublatzky, F., Guerra, P. M., Pastor, M. C., Schupp, H. T., & Vila, J. Additive effects of threat-of-shock and picture valence on startle reflex modulation. https://doi.org/10.1371/journal.pone.0054003 (2013).
5. Bublatzky, F. & Alpers, G. W. Facing two faces: Defense activation varies as a function of personal relevance. *Biological Psychology* **125**, 64–69, https://doi.org/10.1016/j.biopsycho.2017.03.001 (2017).
6. Bublatzky, F., Pittig, A., Schupp, H. T., & Alpers, G. W. Face-to-face: Visual attention to emotional facial expressions depend on face orientation. *Social Cognitive and Affective Neuroscience*, https://doi.org/10.1093/scan/nsx001 (2017).
7. Grillon, C., Ameli, R., Woods, S. W., Merikangas, K. & Davis, M. Fear-potentiated startle in humans: Effects of anticipatory anxiety on the acoustic blink reflex. *Psychophysiology* **28**(5), 588–595, https://doi.org/10.1111/j.1469-8986.1991.tb01999.x (1991).
8. Adolphs, R. Fear, faces, and the human amygdala. *Current Opinion in Neurobiology* **18**(2), 166–172, https://doi.org/10.1016/j.conb.2008.06.006 (2008).
9. Grillon, C. & Charney, D. R. In the face of fear: anxiety sensitizes defensive responses to fearful faces. *Psychophysiology* **48**(12), 1745–1752, https://doi.org/10.1111/j.1469-8986.2011.01268.x (2011).
10. Öhman, A., Lundqvist, D., & Esteves, F. The face in the crowd revisited: a threat advantage with schematic stimuli. *Journal of Personality and Social Psychology*, **80**(3), https://doi.org/10.1037/0022-3514.80.3.381 (2001).
11. Becker, D. V., Anderson, U. S., Mortensen, C. R., Neufeld, S. L. & Neel, R. The face in the crowd effect unconfounded: Happy faces, not angry faces, are more efficiently detected in single- and multiple-target visual search tasks. *Journal of Experimental Psychology: General* **140**, 637–659, https://doi.org/10.1037/a0024060 (2011).
12. Craig, B. M., Becker, S. I. & Lipp, O. V. Different faces in the crowd: A happiness superiority effect for schematic faces in heterogeneous backgrounds. *Emotion* **14**(4), 794–803, https://doi.org/10.1037/a0036043 (2014).
13. Bublatzky, F., Gerdes, A. B. M., White, A. J., Riemer, M. & Alpers, G. W. Social and emotional relevance in face processing: Happy faces of future interaction partners enhance the late positive potential. *Frontiers in Human Neuroscience* **8**, 493, https://doi.org/10.3389/fnhum.2014.00493 (2014).
14. Gerdes, A. B. M., Wieser, M. J., Alpers, G. W., Strack, F. & Pauli, P. Why do you smile at me while I'm in pain? Pain selectively modulates voluntary facial muscle responses to happy faces. *International Journal of Psychophysiology* **85**, 161–167, https://doi.org/10.1016/j.ijpsycho.2012.06.002 (2012).
15. Lang, P. J. & Bradley, M. M. Emotion and the motivational brain. *Biological Psychology* **84**(3), 437–450, https://doi.org/10.1016/j.biopsycho.2009.10.007 (2010).
16. Lang, P. J., Bradley, M. M., & Cuthbert, B. N. Motivated Attention: Affect, activation, and action. Attention and orienting: Sensory and motivational processes (pp. 97–135). (1997).
17. Adams, R. B., Ambady, N., Macrae, C. N. & Kleck, R. E. Emotional expressions forecast approach-avoidance behavior. *Motivation and Emotion* **30**(2), 177–186, https://doi.org/10.1007/s11031-006-9020-2 (2006).
18. Herbert, C., Pauli, P. & Herbert, B. M. Self-reference modulates the processing of emotional stimuli in the absence of explicit self-referential appraisal instructions. *Social Cognitive and Affective Neuroscience* **6**, 653–61, https://doi.org/10.1093/scan/nsq082 (2011).
19. Schindler, S. & Kissler, J. People matter: Perceived sender identity modulates cerebral processing of socio-emotional language feedback. *NeuroImage* **134**, 160–169, https://doi.org/10.1016/j.neuroimage.2016.03.052 (2016).
20. Wieser, M. J. *et al*. Not so harmless anymore: how context impacts the perception and electrocortical processing of neutral faces. *NeuroImage* **92**, 74–82, https://doi.org/10.1016/j.neuroimage.2014.01.022 (2014).
21. Baas, J. M., Milstein, J., Donlevy, M. & Grillon, C. Brainstem correlates of defensive states in humans. *Biological Psychiatry* **59**(7), 588–593, https://doi.org/10.1016/j.biopsych.2005.09.009 (2006).
22. Bublatzky, F., Flaisch, T., Stockburger, J., Schmälzle, R. & Schupp, H. T. The interaction of anticipatory anxiety and emotional picture processing: An event-related brain potential study. *Psychophysiology* **47**(4), 687–696, https://doi.org/10.1111/j.1469-8986.2010.00966.x (2010).
23. Mechias, M. L., Etkin, A. & Kalisch, R. A meta-analysis of instructed fear studies: implications for conscious appraisal of threat. *NeuroImage* **49**(2), 1760–1768, https://doi.org/10.1016/j.neuroimage.2009.09.040 (2010).
24. Bradley, M. M., Moulder, B. & Lang, P. J. When good things go bad: The reflex physiology of defense. *Psychological Science* **16**(6), 468–473, https://doi.org/10.1111/j.0956-7976.2005.01558.x (2005).
25. Olsson, A. & Phelps, E. A. 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 (2004).
26. Bublatzky, F., Alpers, G. W. & Pittig, A. From avoidance to approach: The influence of threat-of-shock on reward-based decision making. *Behaviour Research and Therapy* **96**, 47–56, https://doi.org/10.1016/j.brat.2017.01.003 (2017).
27. Schlund, M. W. *et al*. "Watch out!": Effects of instructed threat and avoidance on human free-operant approach–avoidance behavior. *Journal of the Experimental Analysis of Behavior* **107**(1), 101–122, https://doi.org/10.1002/jeab.238 (2017).
28. Bublatzky, F., Gerdes, A. & Alpers, G. W. The persistence of socially instructed threat: Two threat-of-shock studies. *Psychophysiology* **51**(10), 1005–1014, https://doi.org/10.1111/psyp.12251 (2014).
29. Costa, V. D., Bradley, M. M. & Lang, P. J. From threat to safety: Instructed reversal of defensive reactions. *Psychophysiology* **52**(3), 325–332, https://doi.org/10.1111/psyp.12359 (2015).
30. Mertens, G. & De Houwer, J. Potentiation of the startle reflex is in line with contingency reversal instructions rather than the conditioning history. *Biological Psychology* **113**, 91–99, https://doi.org/10.1016/j.biopsycho.2015.11.014 (2016).
31. Rowles, M. E., Lipp, O. V. & Mallan, K. M. On the resistance to extinction of fear conditioned to angry faces. *Psychophysiology* **49**(3), 375–380, https://doi.org/10.1111/j.1469-8986.2011.01308.x (2012).
32. Bublatzky, F. & Schupp, H. T. Pictures cueing threat: brain dynamics in viewing explicitly instructed danger cues. *Social Cognitive and Affective neuroscience* **7**, 611–622, https://doi.org/10.1093/scan/nsr032 (2012).
33. Fiske, S. T., & Neuberg, S. L. A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In Advances in experimental social psychology (Vol. 23, pp. 1–74). Academic Press. https://doi.org/10.1016/S0065-2601(08)60317-2 (1990).
34. Golkar, A. & Olsson, A. The interplay of social group biases in social threat learning. *Scientific Reports* **7**(1), 7685, https://doi.org/10.1038/s41598-017-07522 (2017).
35. Schiller, D., Levy, I., Niv, Y., LeDoux, J. E. & Phelps, E. A. From fear to safety and back: reversal of fear in the human brain. *Journal of Neuroscience* **28**(45), 11517–11525, https://doi.org/10.1523/JNEUROSCI.2265-08.2008 (2008).
36. Mallan, K. M., Sax, J. & Lipp, O. V. Verbal instruction abolishes fear conditioned to racial out-group faces. *Journal of Experimental Social Psychology* **45**(6), 1303–1307, https://doi.org/10.1016/j.jesp.2009.08.001 (2009).
37. Hugdahl, K. Electrodermal conditioning to potentially phobic stimuli: Effects of instructed extinction. *Behaviour Research and Therapy* **16**(5), 315–321, https://doi.org/10.1016/0005-7967(78)90001-3 (1978).
38. Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* **39**, 175–191, https://doi.org/10.3758/BF03193146 (2007).

39. Lundqvist, D., Flykt, A., & Öhman, A. The Karolinska directed emotional faces (KDEF). CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet (1998).
40. Riemer, M., Bublatzky, F., Trojan, J. & Alpers, G. W. Defensive activation during the rubber hand illusion: Ownership versus proprioceptive drift. *Biological psychology* **109**, 86–92, https://doi.org/10.1016/j.biopsycho.2015.04.011 (2015).
41. Bradley, M. M. & Lang, P. J. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry* **25**(1), 49–59, https://doi.org/10.1016/0005-7916(94)90063-9 (1994).
42. Blumenthal, T. D. *et al*. Committee report: Guidelines for human startle eyeblink electromyographic studies. *Psychophysiology* **42**(1), 1–15, https://doi.org/10.1111/j.1469-8986.2005.00271.x (2005).
43. Dunning, J. P., Del Donno, S. & Hajcak, G. The effects of contextual threat and anxiety on affective startle modulation. *Biological Psychology* **94**(1), 130–135, https://doi.org/10.1016/j.biopsycho.2013.05.013 (2013).
44. Lipp, O. V. & Edwards, M. S. Effect of instructed extinction on verbal and autonomic indices of Pavlovian learning with fear-relevant and fear-irrelevant conditional stimuli. *Journal of Psychophysiology* **16**(3), 176, https://doi.org/10.1027//0269-8803.16.3.176 (2002).
45. Navarette, C. D. *et al*. The roles of race and gender in the persistence of learned fear. *Psychological Science* **20**, 155–198 (2009).
46. Zaalberg, R., Manstead, A. & Fischer, A. Relations between emotions, display rules, social motives, and facial behaviour. *Cognition and Emotion* **18**(2), 183–207, https://doi.org/10.1080/02699930341000040 (2004).
47. Klinkenberg, I. A. *et al*. Healthy individuals maintain adaptive stimulus evaluation under predictable and unpredictable threat. *NeuroImage* **136**, 174–185, https://doi.org/10.1016/j.neuroimage.2016.05.041 (2016).
48. Funayama, E. S., Grillon, C., Davis, M. & Phelps, E. A. A double dissociation in the affective modulation of startle in humans: effects of unilateral temporal lobectomy. *Journal of Cognitive Neuroscience* **13**(6), 721–729, https://doi.org/10.1162/08989290152541395 (2001).
49. Lovibond, P. F. & Shanks, D. R. The role of awareness in Pavlovian conditioning: empirical evidence and theoretical implications. *Journal of Experimental Psychology: Animal Behavior Processes* **28**(1), 3 (2002).
50. Mineka, S. & Zinbarg, R. A contemporary learning theory perspective on the etiology of anxiety disorders: it's not what you thought it was. *American Psychologist* **61**(1), 10, https://doi.org/10.1037/0003-066X.61.1.10 (2006).
51. Rachman, S. The conditioning theory of fear acquisition: A critical examination. *Behaviour Research and Therapy*, **15**(5), 375–387, doi:10.1.1.527.294 (1977).
52. Askew, C., Reynolds, G., Fielding-Smith, S. & Field, A. P. Inhibition of vicariously learned fear in children using positive modeling and prior exposure. *Journal of Abnormal Psychology* **125**, 279–291, https://doi.org/10.1037/abn0000131 (2016).
53. Keltner, D. & Kring, A. M. Emotion, social function, and psychopathology. *Review of General Psychology* **2**(3), 320 (1998).
54. Paret, C., Jennen-Steinmetz, C. & Schmahl, C. Disadvantageous decision-making in borderline personality disorder: Partial support from a meta-analytic review. *Neuroscience & Biobehavioral Reviews* **72**, 301–309, https://doi.org/10.1016/j.neubiorev.2016.11.019 (2017).
55. Lindström, B. & Tobler, P. N. Incidental ostracism emerges from simple learning mechanisms. *Nature Human Behaviour*, **1**, https://doi.org/10.1038/s41562-018-0355-y (2018).
56. Kaufmann, J. M. & Schweinberger, S. R. Expression influences the recognition of familiar faces. *Perception* **33**(4), 399–408, https://doi.org/10.1068/p5083 (2004).
57. Guerra, P. *et al*. Filial versus romantic love: contributions from peripheral and central electrophysiology. *Biological Psychology* **88**(2), 196–203, https://doi.org/10.1016/j.biopsycho.2011.08.002 (2011).
58. Langeslag, S. J., Jansma, B. M., Franken, I. H. & Van Strien, J. W. Event-related potential responses to love-related facial stimuli. *Biological Psychology* **76**(1-2), 109–115, https://doi.org/10.1016/j.biopsycho.2007.06.007 (2007).
59. Tang, H. *et al*. Interpersonal brain synchronization in the right temporo-parietal junction during face-to-face economic exchange. *Social Cognitive and Affective Neuroscience* **11**(1), 23–32, https://doi.org/10.1093/scan/nsv092 (2015).
60. Pittig, A., Alpers, G. W., Niles, A. N. & Craske, M. G. Avoidant decision-making in social anxiety disorder: a laboratory task linked to *in vivo* anxiety and treatment outcome. *Behaviour Research and Therapy* **73**, 96–103, https://doi.org/10.1016/j.brat.2015.08.003 (2015).

Acknowledgements

We are grateful to O. Denzler, O. Grüner, S. Reuter, and C. Zala for their help with data collection. We received helpful comments from Fatih Kavcioglu. This research was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft; BU 3255/1-1, granted to F. Bublatzky).

Author Contributions

F.B., P.G. and G.W.A conceived the study and were involved in the generation and analyses of the data. F.B. wrote the manuscript and all authors revised the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-33269-2.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.