The impact of trust violations on emotional interference control

Shuge Yuan¹, Mengsi Xu², and Yue Zhu³

¹Chinese Academy of Sciences
²Shaanxi Normal University
³Southwest University

March 15, 2024

Abstract

Processing negative stimuli holds profound significance for survival, yet excessive focus on this information can lead to poor social adaptation. Trust violations, common occurrences in negative interpersonal interactions, result in individuals allocating greater cognitive resources towards processing negative stimuli. However, it is still unclear how trust violations influence the ability to control interference from negative stimuli. Therefore, this study employed the Emotional Stroop paradigm (Experiment 1) and the cued Ultimatum game (Experiment 2), using ERP technology to explore how trust violations influence emotional interference control abilities. In Experiment 1, the results showed that the violation group exhibited smaller SP amplitudes in the negative interference condition compared to the no-interference condition during the emotional Stroop paradigm. In Experiment 2, the violation group exhibited smaller SP amplitudes when benign cues were contaminated by threatening information relative to when they were not. Moreover, the violation group displayed higher P3 amplitudes for untrustworthy cues and angry faces than for trustworthy cues and happy faces, respectively. These findings were not observed in the control group. Collectively, these results suggest that trust violations not only draw more cognitive resources towards processing negative stimuli but also undermine interference control ability towards these stimuli.
SP amplitudes in the negative interference condition compared to the no-interference condition during the emotional Stroop paradigm. In Experiment 2, the violation group exhibited smaller SP amplitudes when benign cues were contaminated by threatening information relative to when they were not. Moreover, the violation group displayed higher P3 amplitudes for untrustworthy cues and angry faces than for trustworthy cues and happy faces, respectively. These findings were not observed in the control group. Collectively, these results suggest that trust violations not only draw more cognitive resources towards processing negative stimuli but also undermine interference control ability towards these stimuli.

Keywords: trust violations, interference control, interpersonal conflict, negative bias

Introduction

Emotional cues can help people infer others’ mental states, however, they sometimes play a misleading role (Frith & Frith, 2006). For example, people are often fooled by the hypocritical displays of goodwill shown by liars. Controlling emotional interference is crucial for goal accomplishment and successful interpersonal interactions (Hung et al., 2018). Previous research has shown that trust violations, a negative social interaction, can strengthen negative cognitive biases (Yuan et al., 2021). However, the impact of trust violations on the ability to control emotional interference, particularly when the interference is negative, remains unclear.

Emotional interference control refers to the ability to monitor and resolve the interference from task-irrelevant emotional stimuli (Cheng & Huang, 2013; Wu et al., 2013). This ability not only manifests itself in the cognitive task but also interpersonal interactions (Koban et al., 2013; Zaki et al., 2010). On the cognitive level, the emotional face-word Stroop paradigm is often used to measure this ability. This task creates emotional interference by presenting opposite emotional valences between task-relevant faces and task-irrelevant words. On the interpersonal interaction level, Ruz and Tudela (2013) created emotional interference by presenting opposite valences between a partner’s expression and actual meaning during the Ultimatum game. For example, when a partner is untrustworthy, a happy face may signal that they may allocate less money to the participants. These emotional interferences are common in daily life, and effectively overcoming them would be beneficial for individuals’ social functioning.

Trust violations refer to situations where the trustor’s expectations of reciprocity are not met by the trustee’s behaviors (Fry, 2019). Such violations can lead to an increase in the trustor’s negative cognitive bias (Bies & Tripp, 1996; Reimann et al., 2018; Yuan et al., 2021). For instance, Yuan et al. (2021) employed the trust game to induce trust violations and the dot-probe paradigm to measure attentional bias. Their findings revealed that trust violations enhanced attentional engagement towards angry faces and hindered attentional disengagement from these faces. Similarly, Reimann et al. (2018) observed a significant positive relationship between the level of trust violations and the state of rumination, indicating that people who experience more trust violations may have a higher likelihood of repeatedly focusing on negative events. This heightened negative bias may render individuals more susceptible to interference from negative task-irrelevant stimuli, potentially impeding their ability to control emotional interference effectively.

Furthermore, trust violations also impede the capacity for cognitive control (Deng et al., 2018; Yuan et al., 2022). In a study by Deng et al (2018), participants were asked to recall events related to trust violation, trust fulfillment, or unrelated content, and complete a survey designed to measure resource depletion. The results revealed that the violation group reported greater ego depletion compared to other groups. Yuan et al. (2021), using the n-back task, found that trust violations impaired the ability to update working memory under medium and high cognitive load levels. These findings suggest that trust violations deplete cognitive resources, which may further undermine subsequent cognitive control when greater resources are required. By combining the heightened negative bias with impaired cognitive control abilities under challenging conditions, trust violations may hinder negative interference control. Nevertheless, this hypothesis requires direct evidence for verification.

Electroencephalography (EEG) has been employed to examine variations in mental operations at different points in the information processing stream (Cohen, 2017). The neural mechanisms of interference control are mostly reflected in the N450 and slow potential (SP) components. The N450 component, a negative-peaking...
fronto-central waveform that occurs 400 and 500 ms after stimulus onset, serves as an index for conflict
detection and monitoring processes (Ma et al., 2016; Zhu et al., 2018). The SP component, a late slow wave
characterized by a centro-parietal waveform that appears between 700 and 800 ms after stimuli presentation,
is often associated with conflict/interference resolution, response selection, or response monitoring (Xue et
al., 2016, 2017). There are greater N450 (more negative) and SP amplitudes in the interference condition
than no-interference condition (Shen et al., 2013). In the interference condition, larger amplitudes on the
N450 and SP components reflect a higher ability to monitor and resolve interference, respectively (Xue et
al., 2016). Moreover, the P3 component, a centro-parietal waveform with a typical peak latency between
300-600 ms after the stimulus onset, signifies the amount of attentional allocation (Polich, 2007).

The present study used two experiments to explore how trust violations influence the ability of emotional
interference control. The trust game was used to induce trust violations, and the emotional face-word
Stroop (in experiment 1) and adapted Ultimatum game (in experiment 2) were used to measure the ability of
emotional interference control while recording event-related potentials. We hypothesized that trust violations
would impair negative interference control. In both two experiments, compared to the control group, the
violation group would have worse performances (i.e., longer RTs and lower ACCs) and smaller N450 and
SP amplitudes in the negative interference condition. Additionally, in the adapted Ultimatum game, the
violation group would display higher P3 amplitudes towards threatening stimuli, indicating an enhanced
negative bias.

Experiment 1

Method

Participants

A total of 60 volunteers participated in this study and they were randomly assigned to either the violation
group (n = 29) or the control group (n = 31). Two female participants from the violation group were
excluded because of low accuracy (below 60%) in the emotional Stroop task. One male from the control
group and one female from the violation group were excluded because they had too many EEG artifacts
(more than 50% of the trials were excluded).

Ultimately, there were 56 participants in the experiment (17-24 years, M = 19.66 years, SD = 1.49): 26
participants (16 females, 10 males) in the violation group and 30 participants (20 females, 10 males) in the
control group. All participants were right-handed, had a normal or corrected-to-normal vision, and had no
history of physical or mental illness. This research was approved by the Local Ethics Committee.

The sample size was estimated using G*power software (Version 3.1) (Faul et al., 2007). To identify a medium
effect (f = 0.25) for interaction between groups (violation, control) × congruency (congruent, incongruent)
and to provide 80% power, at least 34 participants would have been needed. Therefore, our final sample (n
= 56) provided enough power to detect the medium effect size.

Procedure and materials

Procedure

First, the participants practiced the trust game and emotional Stroop task. Then, they were randomly
assigned to either the violation or the control group, and they successively performed the trust game and
Stroop task. The trust evaluation and emotional state were assessed both before and after the trust game.
At the end of the experiment, the participants were asked explain the purpose of the study. The trust game
and emotional Stroop task were programmed using E-Prime version 2.0 and all stimuli were presented on a
19-in. CRT monitor viewed at a distance of 60 cm.

Trust game

This trust game used in this experiment has been employed in previous study to induce trust violations (Yuan
et al., 2022). At the beginning of the game, the participant selects a partner using a computer interface, and
both parties contribute their own 10 Chinese yuan as the initial investment amount. In each round of the game, the participant decides how much money to invest in their partner's account. This amount of money is multiplied by three and then transferred to the partner's account. The partner then decides how much money to return to the participant. If the participant decides not to invest any money, the money of both parties remains unchanged in that round (Wang et al., 2021). The total amount of money from each round serves as the principal investment for the next round. At the end of the game, the amount of money received by the participants constitutes their additional compensation (See Figure 1).

![Fig. 1. An example trial of the trust game](image)

The trust violation occurred when participants received less money than they invested in the game. The game ended either when the remaining amount of money was zero or when the game reached the fifth round. In the violation group, the proportion of returns from the partner was 50%, 10%, zero, 1/3, and zero, respectively, from the first to the fifth round (Kuwabara et al., 2014; Lount Jr et al., 2008). In the control group, after making a decision, the participants received the following feedback: ‘the decision of your partner will be presented later, please start the next round of investment’. This type of feedback was intended to maintain trust relatively unchanged.

Questionnaires

Emotion evaluation

The Positive and Negative Affect Schedule (PANAS) is a questionnaire used to measure emotional state. It includes 10 items assessing the positive emotion and 10 items assessing negative emotion. Participants were asked to self-report their current emotional state on a five-point scale (1 = very slightly or not at all to 5 = extremely). The Cronbach’s alpha for the first measure was 0.816, and for the second measure it was 0.794 (Watson et al., 1988).

Trust evaluation

One item was used to measure the participant’s evaluation of the partner before the first round of the game and at the end of the game. This item was “I think this person is trustworthy” (De Cremer, 2004).

Trust behavior

Before the end of the game, participants were asked to decide how much money they wanted to invest in the same partner if they had 10 CNY.

Emotional Stroop Task

Stimuli

A total of 40 emotional faces were selected from the Chinese Affective Faces System as the stimuli (Wang & Luo, 2005). There were 10 happy faces and 10 angry faces per gender. They were assessed for their valence and arousal on a nine-point scale in a previous survey. The two categories of faces differed significantly in valence ($t_{(38)} = -25.94$, $p < 0.001$, power $(1-\beta) = 1.00$; $M_{negative} = 2.74$, $SD = 0.32$; $M_{positive} = 6.66$, $SD = 0.59$). There was no significant difference between the arousal of negative and positive faces ($t_{(38)}$)
= 1.11, \( p = 0.275; M_{\text{negative}} = 6.08, SD = 1.16; M_{\text{positive}} = 5.71, SD = 0.92 \). The Chinese words “高兴” (means “happy”) or “愤怒” (means “angry”) superimposed in prominent red colour on emotional faces. The angry incongruent condition consisted of the “happy” word and the angry face, while the angry congruent condition referred to the “angry” word printed on angry faces. The same applied to the happy faces. Finally, there were 20 faces in each of the four conditions: angry incongruent, angry congruent, happy incongruent, and happy congruent. The size of these faces was 255*283 pixels.

Task procedure

Each trial began with a 500 ms fixation duration, followed by a 300-500ms blank screen. Then an emotional face was presented, remaining on the screen until the participant made a response or 1000ms elapsed. The participants were asked to judge the emotion of the face and ignore the word. They were required to respond by pressing the “F” key for angry faces, and the “J” key for happy faces. After the emotional image disappeared, a blank screen appeared for 1000-1200ms (See Figure 2). This task was composed of four blocks, with 80 trials per block. There were 16 practice trials, and the practice phase was prolonged until 70% of correct responses had been achieved.

Fig. 2. An example trial of the emotional stroop

Electroencephalography recording and processing

Continuous electrophysiological signals were recorded using an electrode cap (Neuroscan, Herndon, VA, USA), with 64 Ag/AgCl scalp sites according to the International 10/20 system. The ground electrode was placed between FPz and Fz, while the online reference electrode was located between Cz and CPz. Vertical electrooculograms (EOGs) were recorded supra-orbitally and infra-orbitally relative to the left eye, while the horizontal EOG was recorded as the difference in activity of the right versus the left orbital rim. The impedance of all electrodes was kept below 5 kΩ. The EEGs and EOGs were amplified using a 0.05-100HZ band-pass and continuously digitized at 500HZ/channel.

After recording, the data were analyzed using custom-made MATLAB (R2013a, The MathWorks, Inc., Natick, MA) scripts supported by EEGLAB (Delorme & Makeig, 2004). The EEG data were first referenced to the average of the left and right mastoids and digitally filtered with band-pass between 0.1 and 30 Hz. Before averaging, independent component analysis (ICA) was performed to detect and remove artifacts associated with muscle, eye movements, and eye-blinks. Trials contaminated with artifacts, due to amplifier clipping and peak-to-peak deflection exceeding ± 100 μV, were also excluded from the average. Trials with error responses were excluded from analyses.

The EEG epochs were created for face-locked ERP averages (i.e., extending from 200ms prior to the face until 1000 ms after face onset), aligned to a baseline of -200 ms until face presentation. The average number of trials per subject for ERP analysis was 67 (SD = 6.97) under the negative incongruent condition, 72 (SD = 6.65) under the negative congruent condition, 67 (SD = 6.54) under the positive incongruent condition, and 72 (SD = 5.67) under the positive congruent condition.

Analysis
Behavior analyses
Firstly, to verify the successful induction of trust violations, scores of emotional states, trust evaluation, and trust behavior were used. For emotional states (positive and negative) and the trust evaluation, mixed-design ANOVAs with group (violation, control) × measurement (first, second) were conducted. For trust behavior, an independent t-test was conducted.

Secondly, in the emotional Stroop task, mean accuracies (ACCs) and response times (RTs) were analyzed using mixed-design ANOVAs with group (violation, control) × emotional face (negative, positive) × congruency (congruent, incongruent). RTs were considered outliers if they were less than 200 ms or deviated more than three standard deviations from the individual mean RT. Outliers and errors were excluded from RT analyses.

ERP analyses
Based on previous studies and topographical maps in this study, the N450 and SP components representing different conflict processing were analyzed (Ma et al., 2016; Shen et al., 2013; Xue et al., 2016). Specifically, N450 was measured as the mean amplitudes between 400 and 500 ms at electrodes AF3, AF4, F1, FZ, F2, and FCZ, while SP was measured as the mean amplitudes between 600 and 800 ms at electrodes CZ and CPZ.

Mixed-design ANOVAs was performed on N450 and SP mean amplitudes with group (violation, control) × emotional face (negative, positive) × congruency (congruent, incongruent). Greenhouse-Geisser adjustments to the degrees of freedom were used for all statistical analyses where appropriate.

Results
Behavior
First, on the trust evaluation, there was a significant interaction between measurement and group, $F(1, 54) = 23.99, p < 0.001, \eta^2_\pi = .308$, power (1-β) = 1.00. Simple effect analysis revealed that the violation group had lower trust evaluation ($M_{\text{violation}} = 2.54, SD = 1.66$) than the control group ($M_{\text{control}} = 4.77, SD = 1.17, p < 0.001$) in the second measurement, and there was no significant difference between two groups in the first measurement ($M_{\text{violation}} = 4.85, SD = 1.05; M_{\text{control}} = 4.93, SD = 0.87, p = 0.735$). Second, on the trust behavior, the independent t-test showed that the violation group had lower investment ($M_{\text{violation}} = 2.73, SD = 2.91$) than the control group ($M_{\text{control}} = 5.83, SD = 1.78, t(54) = -4.886, p < 0.001, \text{power (1-\beta)} = 1.00$).

Third, on the positive emotional state evaluation, there was a significant interaction between measurement and group, $F(1, 54) = 5.97, p = 0.018, \eta^2_\pi = .099$, power (1-β) = 1.00. The simple effect analysis revealed that there was no significant difference between the two groups in the second measurement ($M_{\text{violation}} = 2.58, M_{\text{control}} = 2.88, p = 0.077$), but not on the first measurement ($M_{\text{violation}} = 3.00, M_{\text{control}} = 2.96, p = 0.780$). On the negative emotional state evaluation, there was no significant main effects or interaction.

These results suggested that the violation group had lower scores on the trust evaluation and behavior than the control group; emotions were worse in the violation group than the control group. These results suggest that trust violations are successfully induced.

Emotional stroop task
Table 1 presents the mean ACCs and RTs on each condition. For ACCs, the main effect of congruency was significant, $F(1, 54) = 86.31, p < 0.001, \eta^2_\pi = 0.62$. There was lower ACCs in the incongruent condition ($M_{\text{incongruent}} = 0.88$) than the congruent condition ($M_{\text{congruent}} = 0.95$). The main effect of group was also significant, $F(1, 54) = 4.44, p = 0.040, \eta^2_\pi = 0.08$, and the violation group ($M_{\text{violation}} = 0.91$) had lower ACCs than the control group ($M_{\text{control}} = 0.93$). There was no significant main or interaction effects ($ps > 0.097$).

Table 1 Mean and standard deviations for Emotional Stroop task by group
For RTs, the main effect of congruency was significant, $F(1, 54) = 290.55$, $p < 0.001, \eta^2_p = 0.84$. There was longer RTs in the incongruent condition ($M_{\text{incongruent}} = 570.40$ ms) than the congruent condition ($M_{\text{congruent}} = 529.17$ ms). The main effect of emotional face was significant, $F(1, 54) = 92.67$, $p < 0.001, \eta^2_p = 0.63$. There was longer RTs for anger faces ($M = 565.08$ ms) than happy faces ($M = 534.49$ ms). The interaction among group, emotion and congruency was significant, $F(1, 54) = 6.57$, $p = 0.013, \eta^2_p = 0.11$. Other effects were nonsignificant ($ps > 0.055$).

For further analysis this three interaction, group (violation, control) $\times$ congruency (consistent, inconsistent) was further analyzed for each emotional face. On the angry face condition, the interaction was significant, $F(1, 54) = 13.21$, $p = 0.001, \eta^2_p = 0.20$, power (1-$\beta$) = 1.00. The simple effect analysis revealed that the violation group had longer RTs than the control group in the inconsistent ($p = 0.034$) not the consistent condition ($p = 0.415$). On the happy face condition, only the main effect of congruency was significant, $F(1, 54) = 159.81$, $p < 0.001, \eta^2_p = 0.75$. Other effects was nonsignificant ($ps > 0.320$).

### ERP Results

**N450** On the N450 mean amplitudes, the main effect of congruency was significant, $F(1, 54) = 18.26$, $p < 0.001, \eta^2_p = 0.25$. There was smaller amplitudes in the incongruent condition ($M_{\text{incongruent}} = 5.07 \mu V$) than the congruent condition ($M_{\text{congruent}} = 5.86 \mu V$). There was no other significant main or interaction effects, $nps > 0.058$.

**SP** On the SP mean amplitudes, the main effect of congruency was significant, $F(1, 54) = 40.90$, $p < 0.001, \eta^2_p = 0.43$. There was larger amplitudes in the incongruent condition ($M_{\text{incongruent}} = 9.31 \mu V$) than the congruent condition ($M_{\text{congruent}} = 9.31 \mu V$). The main effect of group was also significant, $F(1, 54) = 31.22$, $p < 0.001, \eta^2_p = 0.37$, the mean amplitudes for anger faces was larger ($M_{\text{anger}} = 9.39 \mu V$) than happy faces ($M_{\text{happy}} = 8.03 \mu V$). The interaction among group, congruency and emotion was significant, $F(1, 54) = 4.38, p = 0.04, \eta^2_p = 0.08$.

For further analysis this three interaction, group (violation, control) $\times$ congruency (congruent, incongruent) was further analyzed for each emotional face. On the angry face condition, only the main effects of congruency ($F(1, 54) = 5.39, p = 0.024, \eta^2_p = 0.09$) and group ($F(1, 54) = 4.24, p = 0.044, \eta^2_p = 0.07$) were significant. The violation group had smaller amplitudes ($M_{\text{violation}} = 8.23 \mu V$) than the control group ($M_{\text{control}} = 10.56 \mu V$). The interaction between them was nonsignificant, $F(1, 54) = 0.09, p = 0.769, \eta^2_p = 0.002$.

On the happy face condition, the interaction between group and congruency was significant, $F(1, 54) = 7.05$, $p = 0.010, \eta^2_p = 0.12$, power (1-$\beta$) = 1.00. The simple effect analysis revealed that the violation group had smaller amplitudes than the control group in the incongruent ($M_{\text{control}} = 10.23 \mu V, SE = 0.79$) not the congruent condition ($M_{\text{control}} = 7.65 \mu V, SE = 0.74$; $M_{\text{violation}} = 6.63 \mu V, SE = 0.80, p = 0.356$) (See Figure 3). There was a negative correlation trend between SP amplitudes and RTs ($r = -0.12, p = 0.375$).
Fig. 3. Face-locked SP results. (A) Grand-average ERP waveforms (left) and mean SP amplitudes between 600 and 800 ms in the incongruent condition (right) at CZ and CPZ following happy faces for each group; (B) Grand-average ERP waveforms (left) and mean SP amplitudes between 600 and 800 ms in the incongruent condition (right) at CZ and CPZ following angry faces for each group; (C) topographical maps assessed between 600 and 800 ms following emotional face for each group. Error bars represent standard errors of the means. An asterisk represents $p < 0.05$; n.s. represents nonsignificant.

Experiment 2

Method

Participants

A total of 50 volunteers participated in this experiment and were randomly assigned to either the violation group ($n = 25$) or the control group ($n = 25$). Five participants were excluded from the analysis because
of low accuracy in the adapted Ultimatum game. Specifically, one male and three female from the violation group, and one male from the control group, had accuracy rates below 70%. Additionally, one male from the violation group was excluded because of excessive EEG artifacts, which resulted in more than 50% of his trials being discarded.

Ultimately, there were 44 participants in the experiment (18-25 years, $M = 21.23$ years, $SD = 1.84$): 20 participants (15 females, 5 males) in the violation group and 24 participants (18 females, 6 males) in the control group. All participants were right-handed, had normal or corrected-to-normal vision, and had no history of physical or mental illness. This research was approved by the Local Ethics Committee.

The sample size was estimated using G*power software (Version 3.1) (Faul et al., 2007). To identify a medium effect \( f = 0.25 \) for interaction between groups (violation, control) \( \times \) trustworthiness cue (trustworthy, untrustworthy) and to provide 80% power, at least 34 participants would have been needed. Therefore, our final sample \((n = 44)\) provided enough power to detect the medium effect size.

Procedure and materials

Procedure

The procedure was the same with Experiment 1 except that the emotional Stroop paradigm was instead with the adapted Ultimatum game.

The adapted Ultimatum game

In the Ultimatum game, partners propose how to divide the money, and participants decide whether to accept or reject the offer. Acceptance leads to both parties receiving the proposed allocation, while rejection results in neither party receiving any money. In this modified Ultimatum game, decisions must be made prior to revealing the proposed allocation. Participants are aided in their decision-making by trust cues and emotional facial expression. Their goal is to secure a larger portion of the funds than their partners, with a successful outcome resulting in an additional Y=5 reward.

Trustworthiness cues are divided into two categories: trustworthy and untrustworthy. Emotional facial expressions include both angry and happy faces. When a partner is deemed trustworthy, emotions tend to predict “natural” consequences: a happy face indicates a higher likelihood of a more favorable offer, while an angry face suggests the opposite. On the other hand, an untrustworthy cue indicates an unreliable partner. In this high-conflict situations, emotional expressions predict outcomes contrary to the “natural” consequences: a happy face indicates a smaller offer for the participant, while an angry face indicates a larger offer. Participants must rely on these cues to determine whether to accept or reject their partner’s proposal. The proposed allocation is then revealed after a decision has been made.

Stimuli

First, squared and circular black shapes were used as trustworthiness cues, indicating whether a partner was trustworthy or untrustworthy in a given trial. These cues were counterbalanced across participants. Second, this experiment employed 10 happy faces and 10 angry faces for each gender, selected from the Chinese Affective Faces System. These faces were rated for valence and arousal on a nine-point scale prior to the experiment. The two categories of faces differed significantly in valence \((t_{(42)} = -30.78, p < 0.001, M_{\text{angry}} = 2.71, SD = 0.54; M_{\text{happy}} = 7.23, SD = 0.64)\). There was no significant difference in the arousal levels of negative and positive faces \((t_{(42)} = 0.68, p = 0.500, M_{\text{angry}} = 6.58, SD = 0.99; M_{\text{happy}} = 6.47, SD = 1.24)\). Third, this experiment employed 16 offers displayed as green and blue numbers (from 1 to 9). The color represented the participant and the partner, respectively, and the difference between the two numbers was always 1. The color and left-right location of offers were counterbalanced across subjects. Half of the offers were beneficial to participants, while the other half were detrimental to them. The validity of the trustworthiness cue and emotional faces were 80% in each trial. For example, in 80% of the trials with a trustworthy cue combined with a happy face or an untrustworthy cue combined with an angry face, the offer was beneficial to the participant.
Task procedure

Each trial began with a 600 ms fixation duration. Then a trustworthiness cue was presented for 500ms, followed by a blank screen for 1000ms. An emotional face replaced the blank screen for 1500ms or until the participant made a response. After this, another 1000ms blank screen was presented. Participants were asked to make decision whether to accept the offer by pressing the "F" key or the "J" key for reject. Finally, the offer was presented for 1000ms (See Figure 4). This task was composed of eight blocks, with 50 trials per block. There were also 16 practice trials, and the practice phase was prolonged until 70% of correct responses had been achieved.

Fig. 4. An example trial of the adapted Ultimatum game

Electroencephalography recording and processing

The EEG recording and processing are the same as in Experiment1, except for the epoch phase. EEG epochs were created for trustworthiness cue-locked and face-locked ERP averages (i.e., extending from 200ms prior to the face until 1000 ms after face onset). The average number of trials per subject for ERP analysis was 170 \((SD = 18)\) under trustworthy condition, 159 \((SD = 20)\) under untrustworthy condition, 90 \((SD = 9)\) under happy trustworthy condition, and 86 \((SD = 8)\) under angry trustworthy condition, 84 \((SD = 8.66)\) under happy untrustworthy condition, and 76 \((SD = 9.78)\) under angry trustworthy condition.

Analysis

Behavior analyses

First, the emotional state, trust evaluation, and trust behavior were used to verify the successful induction of the trust violations, which is the same as in Experiment1.

Second, in the adapted Ultimatum game, group (violation, control) \(\times\) trustworthiness cue (trustworthy, untrustworthy) \(\times\) emotional face (angry, happy) mixed-design ANOVAs were preformed on the mean ACCs and RTs. RTs less than 200 ms or longer than 1500ms, and values deviating more than three standard deviations from the individual mean RT, were defined as outliers. Errors and outliers were excluded from RT analyses.

ERP analyses

Based on previous studies and topographical maps in this study, the trustworthiness cue-P3 and emotional face-P3, N450 and SP components were analyzed (Liotti et al., 2000; McNeely et al., 2003; West & Alain, 2000). Specifically, cue-P3 was assessed as mean amplitude between 420 and 470 ms at electrodes CP1, CPZ, CP2, CP4, P1, PZ, P2, P4, and P6; target-P3 was assessed as mean amplitude between 350 and 650 ms at electrodes CP1, CPZ, CP2, CP4, P1, PZ, and P2; N450 was assessed as mean amplitude between 432 and 452 ms at electrodes AF3, AF4, F1, FZ, and F2; SP was assessed as mean amplitude between 760 and 960 ms at electrodes FC1, FC, FC2, C1, CZ, and C2.

Group (violation, control) \(\times\) trustworthiness cue (trustworthy, untrustworthy) mixed-design ANOVAs were performed on the cue-P3. Groups (violation, control) \(\times\) trustworthiness cue (trustworthy, untrustworthy)
emotional face (angry, happy) mixed ANOVAs were analyzed on the face-P3, N450 and SP. Greenhouse-Geisser adjustments to the degrees of freedom were used for all statistical analyses where appropriate.

**Results**

**Behavior**

First, on the trust evaluation, there was a significant interaction between measurements and groups, \( F(1, 42) = 78.50, p < 0.001, \eta^2_p = 0.65 \). Simple effect analysis revealed that the violation group had lower trust evaluation (\( M_{\text{violation}} = 1.75, SD = 0.91 \)) than the control group (\( M_{\text{control}} = 4.96, SD = 1.04, p < .001 \)) in the second measurement and there was no significant difference between the two groups in the first measurement (\( M_{\text{violation}} = 4.90, SD = 1.21; M_{\text{control}} = 5.17, SD = 0.87, p = 0.400 \)). Second, on trust behavior, the independent \( t \)-test showed that the violation group had lower investment (\( M_{\text{violation}} = 2.15, SD = 2.60 \)) than the control group (\( M_{\text{control}} = 6.42, SD = 1.79, t(42) = -6.42, p < 0.001 \)).

Third, on the positive emotional state evaluation, there was a significant interaction between measurements and groups, \( F(1, 42) = 10.30, p = 0.003, \eta^2_p = 0.20 \). The simple effect analysis revealed that there was marginally difference between the two groups on the second measurement (\( M_{\text{violation}} = 2.57, M_{\text{control}} = 2.85, p = 0.098 \)), but not on the first measurement (\( M_{\text{violation}} = 3.00, M_{\text{control}} = 2.85, p = 0.392 \)). On the negative emotional state evaluation, there was a significant interaction between measurements and groups, \( F(1, 42) = 12.75, p = 0.001, \eta^2_p = 0.23 \). The simple effect analysis revealed that there was significantly difference between the two groups on the second measurement (\( M_{\text{violation}} = 1.98, M_{\text{control}} = 1.51, p = 0.003 \)), but not on the first measurement (\( M_{\text{violation}} = 1.49, M_{\text{control}} = 1.60, p = 0.506 \)).

These results suggest that the violation group had lower scores on the trust evaluation and behavior than the control group; emotions were worse in the violation group than the control group. Thus, these results confirm the effectiveness of the trust violation manipulation.

The adapted Ultimatum game

Table 2 presents the mean ACCs and RTs on each condition. For ACCs, the main effect of trustworthiness cue was significant, \( F(1, 42) = 66.78, p < 0.01, \eta^2_p = 0.61 \). There was lower ACCs in the condition of untrustworthy cue than the trustworthy cue. The main effect of emotional face was also significant, \( F(1, 42) = 35.51, p < 0.01, \eta^2_p = 0.46 \). There was lower accuracies in the angry face condition than the happy face. There was no significant main or interaction effects (\( ps > 0.09 \)).

For RTs, the main effects of cue, \( F(1, 42) = 148.47, p < 0.01, \eta^2_p = 0.78 \), and group were significant, \( F(1, 42) = 196.65, p < 0.010, \eta^2_p = 0.82 \). The interaction between cue and emotion was significant, \( F(1, 42) = 25.67, p < 0.010, \eta^2_p = 0.38 \). The simple effect analysis revealed that there was longer RTs in the untrustworthy cue than the trustworthy cue, and longer RTs in the angry face condition than the happy face. There was no significant main or interaction effects (\( ps > 0.455 \)).

Table 2 Mean and standard deviations for The adapted Ultimatum game by group

<table>
<thead>
<tr>
<th>Groups</th>
<th>Trustworthy happy</th>
<th>Trustworthy angry</th>
<th>Untrustworthy happy</th>
<th>Untrustworthy angry</th>
</tr>
</thead>
<tbody>
<tr>
<td>RTs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Violation</td>
<td>547.45 (59.42)</td>
<td>643.45 (85.87)</td>
<td>657.21 (67.07)</td>
<td>702.78 (87.86)</td>
</tr>
<tr>
<td>Control</td>
<td>529.93 (63.84)</td>
<td>624.77 (83.60)</td>
<td>637.57</td>
<td>686.45</td>
</tr>
<tr>
<td>ACCs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Violation</td>
<td>0.96 (0.04)</td>
<td>0.91 (0.07)</td>
<td>0.90 (0.06)</td>
<td>0.87 (0.07)</td>
</tr>
<tr>
<td>Control</td>
<td>0.96 (0.05)</td>
<td>0.91 (0.07)</td>
<td>0.88 (0.07)</td>
<td>0.85 (0.07)</td>
</tr>
</tbody>
</table>

**ERP Results**

Trustworthiness Cue-P3: On the P3 mean amplitudes, the interaction effect between trustworthiness cue
and group was significant, $F (1,42) = 11.73, p = 0.001, \eta^2_p = 0.22$, power (1-\beta ) = 1.00. The simple effect analysis revealed that the violation group had larger amplitudes under the untrustworthy cue ($M = 3.24 \mu V, SD = 1.73$) than the trustworthy cue ($M = 2.86 \mu V, SD = 1.61, p = 0.043$). However, the control group had smaller amplitudes under the untrustworthy cue condition ($M = 3.04 \mu V, SD = 1.90$) than the trustworthy cue ($M = 3.50 \mu V, SD = 1.85, p = 0.008$). There was no any significant main effects ($ps > 0.678$) (See Figure 5).

Fig. 5. Trustworthiness cue-locked P3 results. (A) Grand-average ERP waveforms in different cues at CP1, CPZ, CP2, CP4, P1, PZ, P2, P4 and P6 for the violation (up) and control group (bottom). (B) topographical maps assessed between 420 and 470 ms following trustworthiness cue for each group. (C) Mean P3 amplitudes under different trustworthiness cues for two groups. Error bars represent standard errors of the means. An asterisk represents $p<0.05$; two asterisks represent $p<0.010$.

Emotional face-P3: On the P3 mean amplitudes, the main effects of trustworthiness cue, $F (1,42) = 37.87, p<0.001, \eta^2_p = 0.47$, and emotional face were significant, $F (1,42) = 5.14, p = 0.029, \eta^2_p = 0.11$. The interaction between group and emotion was significant, $F (1,42) = 4.81, p = 0.034, \eta^2_p = 0.10$, power (1-\beta ) = 0.99. The simple effect analysis revealed that the violation group had significant higher amplitudes for angry face ($M = 10.13 \mu V, SE = 1.24$) than the happy face ($M = 8.78 \mu V, SE = 1.15, p = 0.004$). And there was no significant difference between emotion on the control group ($M_{happy} = 8.98 \mu V, SE = 1.05; M_{anger} = 9.00 \mu V, SE = 1.14, p = 0.958$) (See Figure 6).

There were no significant main or interaction effects, except the interaction between trustworthiness cue and emotional face. The interaction between trustworthiness cue and emotional face was significant, $F (1,42) = 8.39, p = 0.006, \eta^2_p = 0.17$. The simple effect analysis revealed that there was significant higher amplitudes for angry face ($M = 10.56 \mu V, SE = 0.84$) than the happy face ($M = 9.26 \mu V, SE = 0.82, p = 0.002$) in the trustworthy cue, not the untrustworthy cue ($M_{happy} = 8.51 \mu V, SE = 0.77; M_{anger} = 8.57 \mu V, SE = 0.88, p = 0.859$).
Fig. 6. Emotional face-locked P3 results. (A) Grand-average ERP waveforms under different emotional faces at CP1, CPZ, CP2, CP4, P1, PZ and P2 for each group. (B) topographical maps assessed between 360 and 650 ms following emotional face for each group. (C) Mean P3 amplitudes under emotional faces for each group. Error bars represent standard errors of the means. Two asterisks represent \( p < 0.010 \); n.s. represents nonsignificant.

Emotional face-N450: On the N450 mean amplitudes, the main effect of cue was significant, \( F(1, 42) = 51.60, p < 0.001, \eta^2 = 0.55 \). There was smaller amplitudes in the untrustworthy condition (\( M_{\text{untrustworthy}} = 1.85 \mu V \)) than the trustworthy condition (\( M_{\text{trustworthy}} = 4.11 \mu V \)). There was no other significant main or interaction effects, \( ps > 0.192 \).

Emotional face-SP: On the SP mean amplitudes, only interactions between group and trustworthiness, \( F(1,42) = 7.39, p = 0.009, \eta^2 = 0.15 \), and among group, trustworthiness cue, and emotional face was significant, \( F(1,42) = 6.73, p = 0.013, \eta^2 = 0.14 \). Other main or interaction effects were all insignificant, \( ps > 0.261 \). For analysis this three interaction further, group (violation, control) \( \times \) trustworthiness cue (trustworthy, untrustworthy) was performed separately on angry and happy face. On the angry face condition, the main and interaction effects were all insignificant, \( ps > 0.423 \). On the happy condition, the interaction effect was significant, \( F(1,42) = 11.14, p = 0.002, \eta^2 = 0.21 \), power \( (1 - \beta) = 1.00 \). Τηε σιμπλε εφφετ αναλψσις ρεςειεδ τηατ τηε ιολατιον γρουπ ηαδ σμαλλερ αμπλιτυδες (\( M = 4.38 \mu V, \Sigma\Delta = 4.65 \)) τηαν τηε οντρυ-τατρωρηψ ςυε ςονδιτιον. There was no significant difference between groups under the trustworthy cue condition (\( M_{\text{violation}} = 5.94 \mu V, SD = 3.91 \); \( M_{\text{control}} = 5.71 \mu V, SD = 2.94, p = 0.827 \))(See Figure 7).

We correlated SP amplitudes with RTs on the condition of trustworthy and happy face, and found that the two showed a significant negative correlation \( (r = -0.459, p = 0.002) \).
Discussion

The present study aimed to explore how trust violations influence emotional interference control in basic cognitive level (emotional Stroop) and interpersonal interactions (cued Ultimatum game). First, the results indicate that trust violations impaired negative interference control in both static and interpersonal interactions. More specifically, compared to the control group, the violation group had smaller SP amplitudes in the negative interference condition than the no interference condition during the emotional Stroop paradigm. Additionally, the violation group also had smaller SP amplitudes when the benign cues are interfered with by threat information than when they are not interfered with.

The impaired ability to control negative interference can be explained by dual competition model (Pessoa, 2009). Firstly, previous studies have found that smaller SP amplitudes always indicate a lower ability to control interference (McNeely et al., 2003; West, 2004; Xiang et al., 2018). For example, McNeely et al. (2003) observed that patients with schizophrenia displayed heightened Stroop interference in their behavior, along with a lack of a conflict SP. These results suggest that patients with schizophrenia may have an impaired conflict resolving process. Similarly, in the present study, there was a positive correlation between SP amplitudes and RTs. Secondly, according to the dual competition model, when the interference stimuli is an angry word or unfair distribution, the violated group may allocate more resources towards task-irrelevant negative stimuli due to a negative bias. This results in a reduced amount of cognitive resources available for the conflict resolution stage. Therefore, the smaller SP amplitudes observed in the violation group than the control group suggest that trust violations impaired the ability to control negative interference.

However, contrary to our hypothesis, we also found trust violations impaired interference control when the target was an angry face. The results showed that the violation group had longer RTs in incongruent condition and smaller SP amplitudes in both congruent and incongruent conditions. According to the dual competition model, when the emotional stimuli that are relatively high in threat lead to enhanced sensory enhancement but also divert processing resources away from other cognitive mechanisms, such as cognitive control. Luo et al. (2014) asked participants to perform an N-back task with fearful and neutral faces as
stimuli. They found that on the low load condition, negative stimuli impaired the behavioral performance. Thus, emotional stimuli may not always promote cognitive processing. However, in the present study, this impairment only appeared in the emotional Stroop task. More direct evidence is still needed to verify that the threat of target stimuli is too high, rather than increasing interference leading to a decrease in conflict resolution.

Second, our findings indeed indicate that trust violations lead to the allocation of increased cognitive resources during the processing of negative stimuli. This is evident in the observation that the violation group exhibited higher P3 amplitudes when exposed to untrustworthy cues and angry faces than trustworthy cues and happy faces, respectively. These findings were not observed in the control group. The heightened allocation of resources towards negative stimuli within the violation group could be attributed to a negative bias. Firstly, it is well-established that larger P3 amplitudes correspond to greater cognitive engagement (Polich, 2007). Moreover, emotional content stimuli, particularly those of a negative nature, further amplify the P3 amplitude (Ishida et al., 2018). For example, a meta-analysis conducted by Clauss et al. (2022) uncovered a linkage between anxiety and both reflexive orienting and sustained attention towards threat. Additionally, Botelho et al. (2023) used meta-analysis to reveal that anxious individuals exhibit heightened P3 amplitudes towards stimuli congruent with negative disorders. Secondly, prior studies employing the interview method have reported that individuals who have experienced trust violations demonstrate hyper-vigilance towards negative information (Bies & Tripp, 1996). Yuan et al. (2021) further substantiated this by employing the dot probe paradigm, which revealed that trust violations augment attentional bias towards negative information. Thirdly, by synthesizing the cognitive model of trust with mood congruency, it becomes apparent that individuals who have undergone trust violations find themselves in a state of low trust and adverse emotional disposition. Consequently, they adopt a heightened vigilance and enhance their maintenance of attention towards potential threats as a protective measure against being misled (Bower, 1981; Koster et al., 2005; Mayo, 2015; Schul et al., 2004). Therefore, the augmented P3 amplitudes observed in our study suggest that the violation group allocates additional resources towards processing threat-related stimuli. Notably, this augmented allocation is not limited to negative facial expressions but also extends to distrust cues.

What’s more, there was no significant influence of the trust violation on the N450 amplitudes under the threat interference, which suggested that the trust violation did not impair the conflict monitoring during threat environment. We speculate that this may be due to the relatively weak interference effect of negative stimuli at this stage. On one hand, Ma et al. (2016) using the emotional Stroop paradigm found that negative stimuli influence the conflict-resolving phase rather than the conflict-monitoring stage. They explain that the conflict monitoring does not fall under intentional control processing, so emotional processing and conflict monitoring do not compete for cognitive resources. On the other hand, there was no significant interaction between the emotional type and congruency on the N450 amplitudes in present study. Thus, processing of negative stimuli did not compete with the conflict monitoring stage for cognitive resources, which may explain why trust violation had no effect on conflict monitoring.

Present study was the first to explore how trust violation influence interference control under emotional context. The results suggest trust violations destroy the balance between negative impulsive process and cognitive control. It enriches and expands the research on the effects of trust violations on cognitive control from the perspective of dual system processing. In the process of trust repair and psychological treatment for individuals who have experienced trust violation, we should start from the correction of cognitive bias of negative stimulus and the conflict control ability of negative stimulus, so as to effectively alleviate the cognitive damage of such individuals.

However, there were still some limitations that needed to be addressed. Firstly, in this study, only “happy” and “angry” were used as interfering words, which may have weakened interference effects as the number of trials increased. Future studies should include additional words in both the positive and negative categories to enrich the two types of “happy” and “angry” and further validate our experimental results. Secondly, the trustworthiness and untrustworthiness of faces were directly informed by the instruction, rather than based on subjects’ prior experience. Although previous studies have suggested that direct communication
can successfully induce trustworthiness, this form of direct communication may be somewhat inconsistent with the reality (Fouragnan et al., 2013). Therefore, future studies can address this point and increase the ecological validity of the experimental results.

Conclusion

Our current study addressed the relationship between trust violations and emotional interference control. The findings suggested that (a) trust violations impaired the resolution of negative interference and (b) increased cognitive resources towards threatening stimuli. The impaired positive interference control requires further verification. Together, these results offer valuable insights into how trust violations influence emotional cognitive control from both a bottom-up and top-down control perspective.

Reference


