OXFORD

Social Cognitive and Affective Neuroscience, 2022, **17**, 590–597

DOI: https://doi.org/10.1093/scan/nsab126
Advance Access Publication Date: 25 January 2022
Original Manuscript

Neural correlates of successful costly punishment in the Ultimatum game on a trial-by-trial basis

Patrick Mussel,[1] Martin Weiß,[2] Johannes Rodrigues,[3] Hauke Heekeren,[1] and Johannes Hewig[3]

[1]Department of Psychology, Freie Universität Berlin, Berlin 14195, Germany
[2]Department of Psychiatry, Psychosomatics, and Psychotherapy, University Hospital Würzburg, Würzburg 97080, Germany
[3]Department of Psychology I, Julius Maximilians University Würzburg, Würzburg 97070, Germany
Correspondence should be addressed to Patrick Mussel, Department of Psychology, Freie Universität Berlin, Habelschwerdter Allee 45, Berlin 14195, Germany.
E-mail: patrick.mussel@fu-berlin.de.

Abstract

Costly punishment describes decisions of an interaction partner to punish an opponent for violating rules of fairness at the expense of personal costs. Here, we extend the interaction process by investigating the impact of a socio-emotional reaction of the opponent in response to the punishment that indicates whether punishment was successful or not. In a modified Ultimatum game, emotional facial expressions of the proposer in response to the decision of the responder served as feedback stimuli. We found that both honored reward following acceptance of an offer (smiling compared to neutral facial expression) and successful punishment (sad compared to neutral facial expression) elicited a reward positivity, indicating that punishment was the intended outcome. By comparing the pattern of results with a probabilistic learning task, we show that the reward positivity on sad facial expressions was specific for the context of costly punishment. Additionally, acceptance rates on a trial-by-trial basis were altered according to P3 amplitudes in response to the emotional facial reaction of the proposer. Our results are in line with the concept of costly punishment as an intentional act following norm-violating behavior. Socio-emotional stimuli have an important influence on the perception and behavior in economic bargaining.

Key words: costly punishment; Ultimatum game; feedback-related negativity; emotional facial expressions; altruism

Introduction

Economic bargaining between two individuals is a highly complex social interaction. While economic theories like rational choice theory (Neumann and Morgenstern, 1944) aim to explain such behavior as a function of maximizing personal utility in terms of monetary gain, research on social, cognitive and neurophysiological variables has altered this traditional view. For example, fairness considerations (Boksem and De Cremer, 2010), emotional states (Mussel *et al.*, 2013), personality (Thielmann *et al.*, 2020), gender (Flinkenfloger *et al.*, 2017) or attractiveness (Ma *et al.*, 2015) have been found to influence economic decision-making beyond personal utility. In the present study, we contribute to this literature by investigating social-emotional variables in an extended bargaining process and their neural representation as predictors of behavior in future decisions.

The Ultimatum game (Güth *et al.*, 1982) is an often-used paradigm to investigate decision-making in an economic setting (de Quervain *et al.*, 2004; Strobel *et al.*, 2011). While research on social-cognitive and neural mechanisms typically focused on the decision of the proposer (Weiland *et al.*, 2012; Rodrigues *et al.*, 2015) or the responder (Sanfey *et al.*, 2003; Polezzi *et al.*, 2008), Mussel *et al.* (2018) extended the interaction process between the two partners. After the decision of the responder, the proposer indicated how he or she felt considering the decision of the responder by sending a socio-emotional cue (a picture with an emotional facial expression). Acceptance rates were altered for offers from proposers who responded in a characteristic way according to the decision of the responder. On the neural level, smiling facial expressions after acceptance elicited a reward positivity in the time frame of the N2 component, indicating that events are better than expected (Gehring and Willoughby, 2002; Holroyd and Coles, 2002; Hajcak *et al.*, 2005; Holroyd *et al.*, 2008; Baker and Holroyd, 2011). Interestingly, the sad compared to neutral facial expression after rejection also elicited a reward positivity. As negative compared to positive emotional facial expressions usually elicit a feedback-related negativity (FRN) (Miltner *et al.*, 1997), this effect was labeled as reversed FRN. It can be interpreted that the rejection of an unfair offer indicates costly punishment (Henrich *et al.*, 2006; Rodrigues *et al.*, 2020), and the sad facial expression following rejection indicated that punishment was successful. Thus, the smile upon acceptance indicated honored reward and the sad face upon rejection successful punishment. Both are accompanied by a reward positivity in the FRN and higher acceptance rates in upcoming trials.

Received: 10 February 2021; **Revised:** 20 September 2021; **Accepted:** 4 November 2021

© The Author(s) 2022. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Higher acceptance rate for proposers who respond with a smile upon acceptance, compared to a neutral facial expression, was replicated in studies using human faces (Weiß *et al.*, 2019b, 2020), but only partly for emojis (Weiß *et al.*, 2019a,b). The reversed FRN effect was not replicated in studies using a block-wise design (Weiß *et al.*, 2019a,b). In those studies, proposers reacted with a contingent facial feedback in response to the decision of the responders across all trials within a block. Thus, the feedback was highly expectable and may thus not have been processed as surprising (Alexander and Brown, 2011) or better or worse than expected (Holroyd and Coles, 2002).

In the present study, we extended this line of research by investigating the effects of successful punishment on a trial-by-trial basis (Osinsky *et al.*, 2012). We also used a block-wise design. However, rather than creating types of proposers reacting with a fixed pattern according to the decision of the responder, we implemented probabilistic contingencies and investigated effects on decision-making. We investigated acceptance rates in trial $n+1$ as well as neural responses following the emotional facial expression (N170, FRN and P3). We expected to replicate the reversed FRN for sad compared to neutral faces. Our main hypothesis is that successful punishment (sad compared to neutral after rejection) as well as rewarded non-punishment (smile compared to neutral after acceptance), including their neural representations, predict higher acceptance rates in trial $n+1$.

Method

The present study was pre-registered. The pre-registration, the data and all scripts are available via https://osf.io/8cj69.

Participants

An a priori sample size of $N=59$ was estimated for a medium effect of partial $\eta^2=0.06$, $\alpha=0.05$ and $\beta=0.95$ (Strobel *et al.*, 2011; Mussel *et al.*, 2013, 2018), see the pre-registration. Eventually, 58 individuals (43 female; 13 male, 2 diverse; mean age = 25.7 years, s.d. = 7.6) participated for course credit. Additionally, they were told that they could gain more money during the Ultimatum game, depending on their task behavior. Since, unknown to them, they played against the computer (see below) they finally received an additional payout of 2.00 euros to keep any frustration about the deception as low as possible. All participants gave written informed consent. The protocol was approved by the local ethics committee of the Department of Education and Psychology of the Freie Universität Berlin.

Experimental design

After arrival, participants received a general instruction about the experiments and they filled in informed consent. Next, three pictures were taken from each participant, each with either a smiling, neutral or sad facial expression. These pictures were used in Task 2 (see below). Additionally, to make the cover story as realistic as possible, they were told that the offers they see in Task 3 (our main task) are from participants who participated in the experiment earlier, and that their pictures would be used accordingly in the future. However, pictures were taken from a standardized set (Langner *et al.*, 2010) and offers were predefined. The pictures were deleted after the experiment, following a debriefing of the participants.

Next, they were given a reinforcement learning task as a reference task to measure neural responses to facial stimuli in a non-costly-punishment setting (Task 1). Due to space restrictions, this task is described in detail in Supplementary Online Material.

Following this task, participants played a modified Ultimatum game in the role of the proposer (Task 2). The purpose of this task was mainly to familiarize the participant with the main experiment and to enhance the plausibility of the cover story. On each of 10 trials, participants made a decision on how many cents (0 to 5 from a total of 10) to offer to his or her partner (see Figure 1A). The decision was visualized as a pie chart. Next, the partner decided to accept or reject the offer. The participants were told that decisions were made by former participants. Unknown to the participant, decisions were predefined with a 100% probability to accept for an offer of 5 cents, continually decreasing to 0% for an offer of 0 cent. Next, the participant had the option to react to the decision of the partner by sending a smiling, sad or neutral facial expression of himself or herself. The pictures contained the images of the participants that were made prior to the experiment. The chosen facial expression was briefly shown as a confirmation.

In Task 3, the main task, participants played the modified Ultimatum game in the role of the receiver (see Figure 1B). There were eight blocks, each played against a different proposer (four female and four male proposers), presented in randomized order. Each block consisted of 48 trials (6 offer levels presented 8 times each), also presented in randomized order (approximately 35 min). Each trial started with the offer of the proposer, ranging from 0 to 5 cents from a total of 10 cents. The offer was displayed as a pie chart. The participants decided to either accept or reject the offer, and the decision was briefly confirmed. Next, the participants saw a facial expression in response to their decision. After an accepted offer, either a neutral or a smiling facial expression was shown with a probability of 50% each, respectively. After a rejected offer, either a neutral or a sad facial expression was shown with a probability of 50% each, respectively. As noted above, participants were told that offers and facial expressions in response to the decision were collected from other participants who played the Ultimatum game earlier (as they did when playing the Ultimatum game in the role of the proposer).

All stimuli were presented on a 17″ screen with a grey background. Stimulus presentation and response recordings were controlled by PsychoPy v2020.1.3 (Peirce, 2008). During the task, participants were seated in a comfortable chair with a distance about 70 cm between the head and the screen. Each of the face

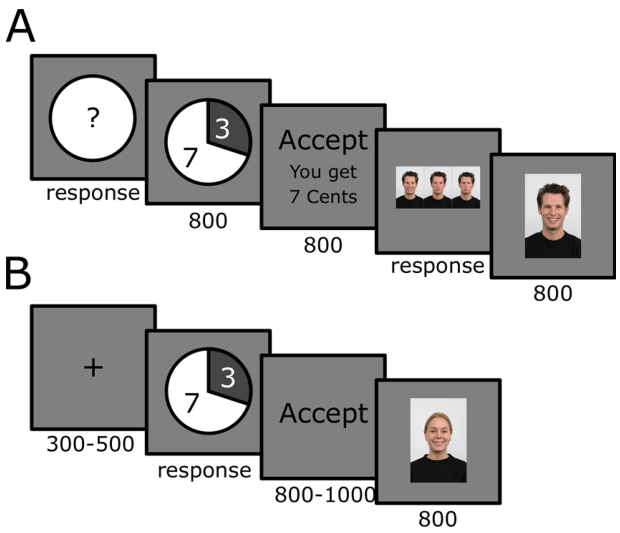

Fig. 1. Task line for the Ultimatum game played in the role of the proposer (A) and receiver (B). Numbers below the stimuli are presentation times in milliseconds.

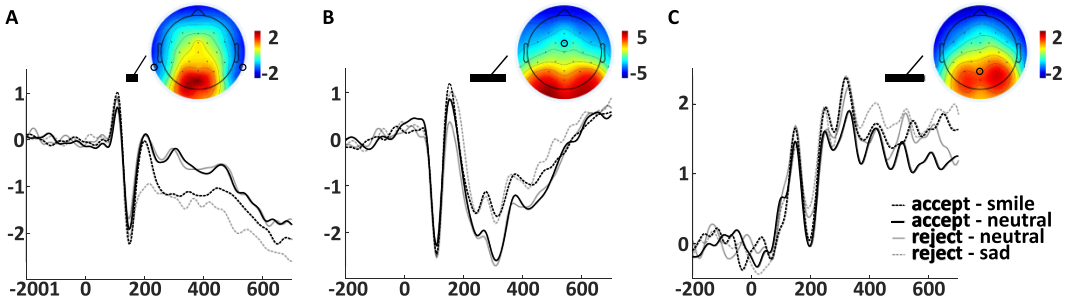

Fig. 2. ERP following the presentation of the emotional facial stimulus. (A) N170 at TP9/TP10, topoplot averaged in the time range from 140 to 176 ms. (B) FRN at FCz, topoplot averaged in the time range from 224 to 344 ms. (C) P3 at Pz, topoplot averaged in the time range from 456 to 588 ms.

pictures was 10 cm high and 6.65 cm wide, resulting in a visual angle of about 14.2° × 9.5°. The pie charts had a diameter of 2.5 cm (3.6° visual angle).

EEG recordings and analyses

While subjects performed the Ultimatum Game, electroencephalogram (EEG) (analog bandpass: 0.1–80 Hz, sampling rate: 250 Hz) was recorded from 31 scalp sites according to the 10-20 system (Fp1, Fp2, F9, F7, F3, Fz, F4, F8, F10, FC5, FC1, FC2, FC6, T7, C3, C4, T8, TP9, CP1, CP2, TP10, P7, P3, Pz, P4, P8, PO9, O1, O2, PO10 and Iz), using Ag/AgCl electrodes and a BrainAmpDC amplifier (Brain Products GmbH, Gilching, Germany). During recording, impedances were kept below 10 kΩ and electrodes were referenced to FCz. Data were processed offline, using MATLAB R2018b (MathWorks, Natick, MA) and the Toolbox EEGLAB v2019.1 (Delorme and Makeig, 2004). First, data were re-referenced to the average across scalp electrodes and electrode FCz was reinstated. Data were then epoched from −800 to 1200 ms around the target stimulus (presentation of the facial expression or offer) and baseline-corrected (baseline from −200 to 0 ms). Channels (excluding FP1 and FP2) containing artifacts were rejected according to joint probability using a criterion of $z = 3.29$. Next data were high-pass filtered with 1 Hz. A first independent component analysis (ICA) was performed for trial rejection. Trials exceeding the criterion of $z = 3.29$ for joint probability or kurtosis (on all channels excluding FP1 and FP2) were excluded (1% of the trials). A second ICA was performed on the remaining trials. Artifacts were detected using the MARA plugin (Winkler *et al.*, 2014). The ICA solution was written to the unfiltered data, and components containing artifacts were automatically rejected. Next, excluded channels were interpolated, and the data were low-pass filtered at 20 Hz (Rodrigues *et al.*, 2021).

ERP quantification

We investigated neural correlates of the decision-making process according to the presentation of emotional facial expressions in Task 3. For each stimulus, we quantified the N170, FRN, P3 amplitudes and theta power for each participant and each trial (see Figure 2).

The N170 is an event-related potential, occurring ∼170 ms after visual processing of human faces at occipito-temporal electrode positions (Bentin *et al.*, 1996; Eimer, 2000; Rossion *et al.*, 2000). It originates in the fusiform gyri (Pizzagalli *et al.*, 2002) and is sensitive to the facial-emotional expression of faces (Hinojosa *et al.*, 2015). We quantified the N170 as the mean amplitude in the time window between 140 and 176 ms according to the grand average at electrodes TP9 and TP10 and according to visual inspection of the ERP. The voltage at the two electrodes was subsequently averaged.

The FRN is a negative deflection in the event-related potential, ∼200–350 ms following negative compared to positive feedback at fronto-central electrode positions (Miltner *et al.*, 1997). It is elicited due to a phasic decrease in dopaminergic signaling in basal ganglia, followed by a disinhibition of apical dendrites of the motor neurons of the anterior cingulate cortex (Debener *et al.*, 2005). More positive amplitudes have been interpreted against the background of reinforcement learning as temporal difference error (Sutton and Barto, 1998), that is events that are better than expected (Holroyd and Coles, 2002; Baker and Holroyd, 2011). We quantified the FRN as the mean amplitude in the interval between 224 and 344 ms according to the grand average and according to the visual inspection of the ERP.[1]

The P3 component typically peaks between 350 and 600 ms and has a positive maximum over parietal electrode sides (Sutton *et al.*, 1965). It has been found to be sensitive to the magnitude of reward (Yeung and Sanfey, 2004), to the motivational relevance of stimuli (Nieuwenhuis *et al.*, 2005) and, according to more recent studies, also to the valence of the reward, with more positive amplitudes for positive compared to negative stimuli (Bellebaum and Daum, 2008; Kreussel *et al.*, 2012). Processes reflected by the P3 have been interpreted as contributing to behavioral adjustment to stimuli from the environment (Bouret and Sara, 2004; Dayan and Yu, 2006). The neural source of the P3 component is less well known and is probably distributed over different regions of the cortex, most likely including the temporal-parietal junction and the locus coeruleus-norepinephrine system (San Martín, 2012). The P3 was quantified as the mean amplitude between 456 and 588 ms at electrode Pz according to the grand average and according to visual inspection of the ERP.[2]

Theta oscillations have been shown to reflect ongoing cognitive processes related to different cognitive tasks, including working memory, executive control and short-term memory load (Kahana *et al.*, 1999; Jensen and Tesche, 2002; Rutishauser *et al.*, 2010; Anguera *et al.*, 2013) Theta power is sensitive to percepts associated with reward and punishment and has thus been proposed as an alternative indicator of feedback processing (Cohen *et al.*, 2008). Theta power was obtained by performing a wavelet-analysis based on Cohen (2014) using the function wavelet_power_2 (Rodrigues *et al.*, 2021). The frequency

[1] Visual inspection of the lower and upper time limits of the ERP components, rather than mean around the peak, was chosen due to the asymmetrical shape of the FRN (see Figure 2).

[2] The early peak at 324 ms was identified as the negative dipole of the N2. See also Supplementary Figure S2 for the ERP of the reference task, for which the same time range was used.

band was set to 4–8 Hz using 3.5 cycles. Data were transformed to decibel by computing 10 times the common logarithm of the power. We quantified the average power between 100 and 500 ms following the stimulus at FCz and subtracted the baseline power between −300 and −100 ms relative to the stimulus. To our surprise, theta power was not sensitive to feedback valence. As such, neither fair compared to unfair offers in the Ultimatum game ($F = 0.02$, $P = 0.88$) nor smiling compared to neutral facial feedback ($F = 0.001$, $P = 0.98$) elicited a corresponding response. Thus, modulation of the feedback response in the context of successful punishment is meaningless as, in the present study, the measure was not sensitive to feedback *per se*. Given these results and recent evidence that FRN and theta are functionally dissociated (Rodrigues *et al.*, 2020; Wang *et al.*, 2020), we refrain from conducting further analyses on theta power.

Statistical analysis

Data were analyzed with linear mixed-effects models and generalized mixed-effects models (for binomial variables; i.e. Mixed Effects Logistic Regressions) in R-Studio 1.3.959 on R 3.6.3 using the packages lme4 (Bates *et al.*, 2014) and ggplot2 (Wickham, 2016). Decision to accept or reject an offer in the Ultimatum game was analyzed on trial level with a generalized linear mixed model with fixed effects 'offer size' and z-standardized 'trial number' and random intercept for 'participants'. Psychophysiological responses to the facial feedback were analyzed with a linear mixed-effects model with the fixed effects 'decision of the participant' (accept *vs.* reject) and 'facial feedback' (more positive, i.e. smile after accept or neutral after reject, compared to more negative, i.e. neutral after accept and sad after reject) and random intercept for 'participants'. Sequential effects of the facial feedback on decision-making in the next trial ($n + 1$) were analyzed using a generalized linear mixed-effect model with fixed effects 'decision in trial n' (accept *vs.* reject) and 'facial feedback' (more positive, i.e. smile after accept or neutral after reject, compared to more negative, i.e. neutral after accept and sad after reject) and random intercept for 'participants'. For brain-behavior relations, we extended this statistical model by including z-standardized neural responses to the facial stimulus in trial n as additional fixed effect and random slopes for z-standardized neural responses to account for potential differences between participants in the relation between electrophysiological indicators and subsequent decisions (separately for amplitudes of the FRN and the P3).

Results

When playing the modified Ultimatum game as a proposer, participants offered on average 3.7 cents. If their offer was accepted, they responded with a smiling facial expression in 91% of trials, and with a neutral expression in the remaining 9%. A sad facial expression was never chosen. If their offer was rejected, they responded with a sad facial expression in 58% and with a neutral expression in 40% of the trials. A smiling expression was chosen in only 2% of the trials.

When playing the Ultimatum game as a responder, the average number of trials were 110, 106, 82, and 82 trials for the conditions of accept-smile, accept-neutral, reject-neutral and reject-sad, respectively. For decision to accept or reject an offer in the Ultimatum game, higher offers resulted in higher acceptance rates ($F = 945$, $df = 5$, $P < 0.001$): for offers rising from 0 to 5 (out of 5) cents, acceptance rates were 15%, 21%, 36%, 77%, 93% and 99%, respectively. A significant effect for trial number ($F = 127$, $df = 1$, $\beta = 0.25$, $P < 0.001$) and the interaction between trial number and offer size ($F = 4.5$, $df = 5$, $P < 0.001$) indicated that acceptance rates increased during the game (estimate 0.25), especially for offers of 3 cents or less.

For N170 amplitudes following the facial feedback, results revealed a significant interaction between facial feedback and decision ($F = 14.5$, $df = 1$, $\beta = 0.36$, $P < 0.001$). The pattern of the effect can be interpreted as arousal effect, with more negative amplitudes for the emotional (accept-smile; reject-sad) compared to neutral (accept-neutral; reject-neutral, all $\beta > 0.20$, $t > 1.9$, $P < 0.05$) facial feedback. A similar effect was found for the reference task 1 (see Supplementary Online Material). No other effects were significant.

For the FRN, we also found a significant interaction between facial feedback and decision ($F = 81$, $df = 1$, $\beta = -1.80$, $P < 0.001$). Interestingly, both the sad facial expression after rejection and the smiling facial expression after acceptance elicited a feedback positivity which was significantly different from the neutral facial expression after both rejection and acceptance ($P < 0.001$, see Figures 2B and 3A). In contrast to the reference task, where sad facial expressions elicited more negative values compared to smiling facial expressions (see Supplementary Online Material), there was no significant difference between these two expressions in the context of the Ultimatum ($P = 0.81$), yielding support for a reversed FRN that might indicate successful punishment. No other effects revealed significance. Direct comparisons between the reference task and the main task can be found in the Supplemental Online Material A.

We found no effect of face or decision on P3 amplitudes. In the reference task, sad compared to smiling facial expressions elicited stronger P3 amplitudes (see Supplemental Online Material A), which may reflect behavioral adjustment. In the context of the Ultimatum game, there was no significant difference between these facial expressions reflecting honored reward and successful punishment. However, we note that descriptively results were in the expected direction (higher P3 amplitudes following successful punishment and honored reward, compared to the neutral facial expressions, respectively). Additionally, exploratory post hoc analyses that we report in Supplemental Material Online B show that the effect was significant when controlling for dynamic states during the game. Particularly, it can be interpreted that the effect is larger in trials in which participants are actively engaged in the game.

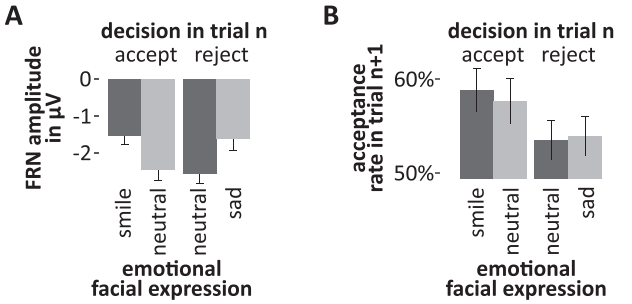

Fig. 3. Effect of emotional facial expression of the proposer (smile, neutral and sad) according to the decision of the responder in trail n on (A) the FRN following the presentation of the emotional facial stimulus; (B) acceptance rated in trial $n + 1$. Error bars indicate standard error of the mean.

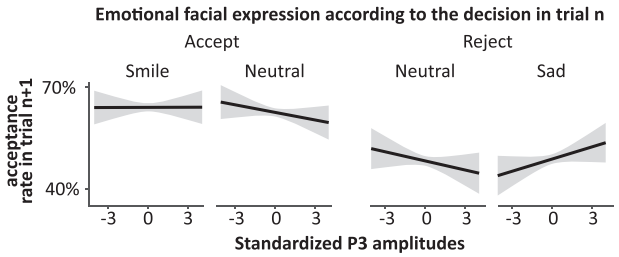

Fig. 4. Acceptance rate in trial $n+1$ as a function of standardized P3 amplitudes following the emotional facial feedback in trial n. The sad facial expression after rejection indicates successful costly punishment, the neutral facial expression after rejection non-successful punishment. Shaded areas represent the 95% confidence interval.

Brain-behavior relations

Regarding sequential effects on decision-making in trial $n+1$, we found a main effect of decision in trial n ($F=29.3$, $df=1$, $\beta=0.21$, $P<0.001$). Higher acceptance rates in trial n predicted higher acceptance rates in trial $n+1$. Descriptively, pressing 'accept' two times in a row is more likely (36%) compared to switching from 'accept' to 'reject' or vice versa (both 21%) or pressing 'reject' two times in a row (22%).[3] Our hypotheses posited that honored reward (smile after accept) and successful punishment (sad after reject) compared to non-honored reward (neutral after acceptance) and non-successful punishment (neutral after rejection) leads to higher acceptance rates in trail $n+1$. While there was a tendency in the expected direction (see Figure 3B), the corresponding interaction between decision in trial n and facial feedback did not reach significance ($F=2.3$, $df=1$, $\beta=-.09$, $P=0.13$).

In our brain-behavior analyses, we found no significant effect of the FRN on decision-making on trial $n+1$, nor for any interaction with decision in trail n or facial feedback. For amplitudes in the P3, we found a significant main effect ($F=4.5$, $df=1$, $\beta=-0.03$, $P=0.03$), indicating that higher amplitudes on the P3 in trial n predicted lower acceptance rates in trial $n+1$. This effect was qualified by a three-way interaction between amplitudes of the P3, decision in trail n and facial feedback ($F=4.9$; $df=1$, $\beta=-.13$, $P=0.03$). As illustrated in Figure 4, the negative effect of P3 on decision-making in trial $n+1$ (higher P3 amplitudes predicting lower acceptance rates) was due to trials of non-honored reward (neutral after acceptance) as well as non-successful punishment (neutral after rejection). On the contrary, the effect vanishes for honored reward (smiling after acceptance) and is even reversed for successful punishment (sad after rejection).

Discussion

Economic bargaining is a complex social interaction. In line with this premise, we confirmed an influence of socio-emotional variables on economic decision-making. As a novelty of the present study, we combined the socio-emotional reaction of the proposer in an extended Ultimatum game with sequence effects in trial-by-trial experiments (Osinsky *et al.*, 2012). As key findings of this study, our results (i) replicate the reversed FRN effect, (ii) resolve a discrepancy in the literature regarding successful punishment effects in block-wise designs and (iii) show that decision-making in the upcoming trial is predicted by the neural representation of successful punishment.

The reversed FRN effect as an indicator of successful punishment was first reported by Mussel *et al.* (2018). The study implemented proposers with different identities (the specific emotional facial expressions according to the decision of the responder), which were indicated at the beginning of each trial in the form of a picture of the proposer with a neutral facial expression. Results revealed a main effect between the smiling identity (smiling upon acceptance, neutral facial expression upon rejection) compared to a neutral identity (neutral facial expression following both acceptance and rejection). However, in two studies using a block-wise design, this effect was not supported (Weiß *et al.*, 2019a,b). In the latter two studies, the picture of the proposer with a neutral facial expression at the beginning of a trial was omitted; rather, receivers played a whole block against a proposer with a certain identity who showed a contingent facial expression upon the decision of the responder. Thus, the reaction of the proposer was easy to learn after a few trials, and thus, fully predictable, which might explain the lack of a reward positivity effect. In the present study, participants also played a full block with the same proposer. However, emotional facial feedback was probabilistic (50%) and, thus, not predictable. We found a significant reversed FRN effect suggesting that not the block-wise design *per se*, but the lack of prediction error was responsible for the conflicting results in the past. The lacking reversed FRN effect might thus be due to the fact that the feedback was fully anticipated (Schultz *et al.*, 1997).

The reversed FRN effect provides support for the concept of costly punishment. The rejection of an unfair offer can thus be interpreted as an intentional act to punish the proposer. The sad facial expression of the proposer upon rejection might signal that the punishment was understood, or the proposer feels sorry. The reward positivity elicited by this emotional facial cue shows that it was the desired and intended outcome and that punishment was successful (see Fehr and Gächter, 2002; de Quervain *et al.*, 2004; Strobel *et al.*, 2011; Mothes *et al.*, 2016).

We found that acceptance rates in the next trial were predicted by the neurophysiological representation of successful punishment. In line with existing evidence (Mussel *et al.*, 2018), processes occurring in the time frame of the P3 (rather than earlier components) were relevant for predicting subsequent behavior. Particularly, for our expected effect of higher acceptance rates after successful and lower acceptance rates after non-successful punishment, we only found a non-significant tendency in the expected direction. However, P3 amplitudes according to successful and non-successful punishment predicted acceptance rates in trial $n+1$ in the expected direction. Particularly, higher compared to lower P3 amplitudes were followed by decreased acceptance rates after non-honored reward and non-successful punishment and by increased acceptance rates after successful costly punishment. This also suggests that emotional facial expressions only impact subsequent behavior when the stimulus is actually interpreted as a signal for successful punishment, as indicated by high compared to low P3 amplitudes.

We note that there is a discrepancy in the reported effects regarding the component that reflects the effects of successful punishment and honored reward. Whereas the emotional facial expression had an impact on the FRN, amplitudes in the time range and topography of the P3 predicted subsequent behavior. In contributing to resolving this discrepancy, we note the following: first, there was a tendency of an effect of emotional facial expressions on P3 amplitudes which became significant when

[3] This effect might reflect acceptance rates >50% on average, increasing acceptance rates during the game or response patterns, such as inconsiderate acceptance of multiple offers in a row.

controlling for (what we interpreted as) current engagement in the task. The latter results, described in detail in the supplemental material, are exploratory post hoc analyses that must be treated with caution. Yet, they provide initial evidence that the P3 component was also affected by the facial expressions, even though in a more complicated way than expected. Second, our results are in line with prior research. Mussel *et al.* (2018) found, in the same vein, an effect of facial expression on the FRN, whereas P3 amplitudes (in their study following the picture of the proposer with a neutral facial expression) predicted subsequent acceptance rates.

In integrating this pattern of results, we propose the following description of the temporal course underlying the processing of the emotional feedback. In the time frame of the N170, facial cues are interpreted regarding their emotional facial expression, as mirrored in equivalent patterns in the Ultimatum game and the reference task. In the time course of the N2, these cues are evaluated regarding their valence within their specific context, as evident from altered patterns in the Ultimatum game compared to the reference task. Finally, in the time frame of the P3, processes relating to behavioral adjustment occur as evident in the prediction of behavior in the next trial (Bouret and Sara, 2004; Dayan and Yu, 2006). This implies a complex interplay of multiple neural structures and across time, which leads to differential effects on different ERP components.

As a limitation of the present study, we first mention that our sample was not representative for the general population. The higher percentage of female participants might have affected our results, eventually in interaction with the gender of the proposer. As the small number of male participants did not allow for the analysis of gender-specific effects, future research is needed to investigate generalization to the general population. Second, to make the feedback realistic, we did not implement a full design with all facial expressions (smile, neutral and sad) following both acceptance and rejection. Third, we used pre-defined offers and stimuli from a pre-tested set of facial expressions, rather than offers and pictures from other participants. We included Task 2 to enhance the cover story yet cannot rule out that some participants might have suspected that they did not play against a participant. Fourth, the incentive for the participants might have been too small to guarantee high engagement, which might have dampened some of the effects. Yet, we were still able to confirm our hypotheses, which support a reasonable stability of the findings.

In sum, our study supports the view of costly punishment as an intentional behavior that can be rewarding. Future decision-making depends on whether punishment is perceived as successful or non-successful. This stresses the importance of feedback also in the context of punishment for the punisher, and the rewarding notion may partly explain the costly rule and law reinforcement behavior in society although no personal materialistic gain is involved.

Acknowledgements

We thank Katharina Trachte and Sebastian Heß for their assistance in collecting the data.

Funding

We acknowledge support by the Open Access Publication Fund of the Freie Universität Berlin.

Conflict of interest

The authors declare no competing financial interests.

Supplementary data

Supplementary data are available at *SCAN* online.

References

Alexander, W.H., Brown, J.W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience*, **14**(10), 1338.

Anguera, J.A., Boccanfuso, J., Rintoul, J.L., *et al.* (2013). Video game training enhances cognitive control in older adults. *Nature*, **501**(7465), 97-+.

Baker, T.E., Holroyd, C.B. (2011). Dissociated roles of the anterior cingulate cortex in reward and conflict processing as revealed by the feedback error-related negativity and N200. *Biological Psychology and Aging*, **87**(1), 25–34.

Bates, D., Mächler, M., Bolker, B., Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**(1), 1–48.

Bellebaum, C., Daum, I. (2008). Learning-related changes in reward expectancy are reflected in the feedback-related negativity. *European Journal of Neuroscience*, **27**(7), 1823–35.

Bentin, S., Allison, T., Puce, A., Perez, E., McCarthy, G. (1996). Electrophysiological studies of face perception in humans. *Journal of Cognitive Neuroscience*, **8**(6), 551–65.

Boksem, M.A.S., De Cremer, D. (2010). Fairness concerns predict medial frontal negativity amplitude in ultimatum bargaining. *Social Neuroscience and Biobehavioral Reviews of Modern Physics*, **5**(1), 118–28.

Bouret, S., Sara, S.J. (2004). Reward expectation, orientation of attention and locus coeruleus-medial frontal cortex interplay during learning. *European Journal of Neuroscience*, **20**(3), 791–802.

Cohen, M.X., Ridderinkhof, K.R., Haupt, S., Elger, C.E., Fell, J. (2008). Medial frontal cortex and response conflict: evidence from human intracranial EEG and medial frontal cortex lesion. *Brain Research*, **1238**, 127–42.

Cohen, M.X. (2014). *Analyzing Neural Time Series Data: Theory and Practice*. Cambridge, MA: The MIT Press.

Dayan, P., Yu, A.J. (2006). Phasic norepinephrine: a neural interrupt signal for unexpected events. *Network (Bristol, England)*, **17**(4), 335–50.

de Quervain, D.J.F., Fischbacher, U., Treyer, V., *et al.* (2004). The neural basis of altruistic punishment. *Science*, **305**(5688), 1254–8.

Debener, S., Ullsperger, M., Siegel, M., Fiehler, K., Von Cramon, D.Y., Engel, A.K. (2005). Trial-by-trial coupling of concurrent electroencephalogram and functional magnetic resonance imaging identifies the dynamics of performance monitoring. *Journal of Neuroscience*, **25**(50), 11730–7.

Delorme, A., Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, **134**(1), 9–21.

Eimer, M. (2000). The face-specific N170 component reflects late stages in the structural encoding of faces. *Neuroreport*, **11**(10), 2319–24.

Fehr, E., Gächter, S. (2002). Altruistic punishment in humans. *Nature*, **415**(6868), 137–40.

Flinkenfloger, N., Novin, S., Huizinga, M., Krabbendam, L. (2017). Gender moderates the influence of self-construal priming on fairness considerations. *Frontiers in Psychology*, **8**(14), 503.

Gehring, W.J., Willoughby, A.R. (2002). The medial frontal cortex and the rapid processing of monetary gains and losses. *Science*, **295**(5563), 2279–82.

Güth, W., Schmittberger, R., Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization Science*, **3**(4), 367–88.

Hajcak, G., Holroyd, C.B., Moser, J.S., Simons, R.F. (2005). Brain potentials associated with expected and unexpected good and bad outcomes. *Psychophysiology*, **42**(2), 161–70.

Henrich, J., McElreath, R., Barr, A., *et al.* (2006). Costly punishment across human societies. *Science*, **312**(5781), 1767–70.

Hinojosa, J., Mercado, F., Carretié, L. (2015). N170 sensitivity to facial expression: a meta-analysis. *Neuroscience and Biobehavioral Reviews of Modern Physics*, **55**, 498–509.

Holroyd, C.B., Pakzad-Vaezi, K.L., Krigolson, O.E. (2008). The feedback correct-related positivity: sensitivity of the event-related brain potential to unexpected positive feedback. *Psychophysiology*, **45**, 688–97.

Holroyd, C.B., Coles, M.G.H. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, **109**(4), 679–709.

Jensen, O., Tesche, C.D. (2002). Frontal theta activity in humans increases with memory load in a working memory task. *European Journal of Neuroscience*, **15**(8), 1395–9.

Kahana, M.J., Sekuler, R., Caplan, J.B., Kirschen, M., Madsen, J.R. (1999). Human theta oscillations exhibit task dependence during virtual maze navigation. *Nature*, **399**(6738), 781–4.

Kreussel, L., Hewig, J., Kretschmer, N., Hecht, H., Coles, M.G.H., Miltner, W.H.R. (2012). The influence of the magnitude, probability, and valence of potential wins and losses on the amplitude of the feedback negativity. *Psychophysiology*, **49**(2), 207–19.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H.J., Hawk, S.T., van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. *Cognition and Emotion*, **24**(8), 1377–88.

Ma, Q., Hu, Y., Jiang, S., Meng, L. (2015). The undermining effect of facial attractiveness on brain responses to fairness in the Ultimatum game: an ERP study. *Frontiers in Neuroscience*, **9**, 77.

Miltner, W.H., Braun, C.H., Coles, M.G. (1997). Event-related brain potentials following incorrect feedback in a time-estimation task: evidence for a 'generic' neural system for error detection. *Journal of Cognitive Neuroscience*, **9**(6), 788–98.

Mothes, H., Enge, S., Strobel, A. (2016). The interplay between feedback-related negativity and individual differences in altruistic punishment: an EEG study. *Cognitive, Affective and Behavioral Neuroscience*, **16**(2), 276–88.

Mussel, P., Goritz, A.S., Hewig, J. (2013). The value of a smile: facial expression affects ultimatum-game responses. *Judgment and Decision Making*, **8**(3), 381–5.

Mussel, P., Hewig, J., Weiß, M. (2018). The reward-like nature of social cues that indicate successful altruistic punishment. *Psychophysiology*, **55**(9), e13093.

Neumann, J., Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

Nieuwenhuis, S., Aston-Jones, G., Cohen, J.D. (2005). Decision making, the P3, and the locus coeruleus–norepinephrine system. *Psychological Bulletin*, **131**(4), 510–32.

Osinsky, R., Mussel, P., Hewig, J. (2012). Feedback related potentials are sensitive to sequential order of decision outcomes in a gambling task. *Psychophysiology*, **49**, 1579–89.

Peirce, J.W. (2008). Generating stimuli for neuroscience using psychopy. *Frontiers in Neuroinformatics*, **2**, 10.

Pizzagalli, D.A., Lehmann, D., Hendrick, A.M., Regard, M., Pascual-Marqui, R.D., Davidson, R.J. (2002). Affective judgments of faces modulate early activity (not similar to 160 ms) within the fusiform gyri. *NeuroImage*, **16**(3), 663–77.

Polezzi, D., Daum, I., Rubaltelli, E., *et al.* (2008). Mentalizing in economic decision-making. *Behavioural Brain Research*, **190**(2), 218–23.

Rodrigues, J., Ulrich, N., Hewig, J. (2015). A neural signature of fairness in altruism: a game of theta? *Social Neuroscience*, **10**(2), 192–205.

Rodrigues, J., Liesner, M., Reutter, M., Mussel, P., Hewig, J. (2020). It's costly punishment, not altruistic: low midfrontal theta and state anger predict punishment. *Psychophysiology*, **15**, 660449.

Rodrigues, J., Weiß, M., Hewig, J., Allen, J.J.B. (2021). EPOS: EEG processing open-source scripts. *Frontiers in Neuroscience*, **15**, 663.

Rossion, B., Gauthier, I., Tarr, M.J., *et al.* (2000). The N170 occipito-temporal component is delayed and enhanced to inverted faces but not to inverted objects: an electrophysiological account of face-specific processes in the human brain. *Neuroreport*, **11**(1), 69–72.

Rutishauser, U., Ross, I.B., Mamelak, A.N., Schuman, E.M. (2010). Human memory strength is predicted by theta-frequency phase-locking of single neurons. *Nature*, **464**(7290), 903–U116.

San Martín, R. (2012). Event-related potential studies of outcome processing and feedback-guided learning. *Frontiers in Human Neuroscience*, **6**(304), 1–17.

Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E., Cohen, J.D. (2003). The neural basis of economic decision-making in the Ultimatum game. *Science*, **300**(5626), 1755–8.

Schultz, W., Dayan, P., Montague, P.R. (1997). A neural substrate of prediction and reward. *Science*, **275**(5306), 1593–9.

Strobel, A., Zimmermann, J., Schmitz, A., *et al.* (2011). Beyond revenge: neural and genetic bases of altruistic punishment. *NeuroImage*, **54**(1), 671–80.

Sutton, R.S., Barto, A.G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Sutton, S., Braren, M., Zubin, J., John, E.R. (1965). Evoked-potential correlates of stimulus uncertainty. *Science*, **150**(3700), 1187–8.

Thielmann, I., Spadaro, G., Balliet, D. (2020). Personality and prosocial behavior: a theoretical framework and meta-analysis. *Psychological Bulletin*, **146**(1), 30–90.

Wang, Y., Cheung, H., Yee, L.T.S., Tse, C.Y. (2020). Feedback-related negativity (FRN) and theta oscillations: different feedback signals for non-conform and conform decisions. *Biological Psychology*, **153**, 107880.

Weiland, S., Hewig, J., Hecht, H., Mussel, P., Miltner, W. (2012). Neural correlates of fair behavior in interpersonal bargaining. *Social Neuroscience*, **7**, 537–51.

Weiß, M., Gutzeit, J., Rodrigues, J., Mussel, P., Hewig, J. (2019a). Do emojis influence social interactions? Neural and behavioral responses to affective emojis in bargaining situations. *Psychophysiology*, **56**(4), e13321.

Weiß, M., Mussel, P., Hewig, J. (2019b). The value of a real face: differences between affective faces and emojis in neural processing

and their social influence on decision-making. *Social Neuroscience*, **15**(3), 255–68.

Weiß, M., Mussel, P., Hewig, J. (2020). Smiling as negative feedback affects social decision-making and its neural underpinnings. *Cognitive, Affective and Behavioral Neuroscience*, **20**, 160–71.

Wickham, H. (2016). *Ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

Winkler, I., Brandl, S., Horn, F., Waldburger, E., Allefeld, C., Tangermann, M. (2014). Robust artifactual independent component classification for BCI practitioners. *Journal of Neural Engineering*, **11**, 035013.

Yeung, N., Sanfey, A.G. (2004). Independent coding of reward magnitude and valence in the human brain. *Journal of Neuroscience*, **24**(28), 6258–64.