



AFRL-AFOSR-JP-TR-2018-0033

Using Brain-State Information to facilitate Conditioned Attitude Formation

Trevor Penney
NATIONAL UNIVERSITY OF SINGAPORE

04/20/2018
Final Report

DISTRIBUTION A: Distribution approved for public release.

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ IOA
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Executive Services, Directorate (0704-0188). Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.					
1. REPORT DATE (DD-MM-YYYY) 20-04-2018		2. REPORT TYPE Final		3. DATES COVERED (From - To) 30 Sep 2014 to 31 Mar 2017	
4. TITLE AND SUBTITLE Using Brain-State Information to facilitate Conditioned Attitude Formation				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER FA2386-14-1-0018	
				5c. PROGRAM ELEMENT NUMBER 61102F	
6. AUTHOR(S) Trevor Penney				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) NATIONAL UNIVERSITY OF SINGAPORE 21 LOWER KENT RIDGE ROAD SINGAPORE, 119077 SG				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) AOARD UNIT 45002 APO AP 96338-5002				10. SPONSOR/MONITOR'S ACRONYM(S) AFRL/AFOSR IOA	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S) AFRL-AFOSR-JP-TR-2018-0033	
12. DISTRIBUTION/AVAILABILITY STATEMENT A DISTRIBUTION UNLIMITED: PB Public Release					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT Humans use facial features and emotional expressions to judge whether others can be trusted to engage in fair social and economic exchanges. These initial judgments can be changed via learning and are affected by the outcome of direct economic exchanges such as those modeled by laboratory trust games. Here, we used a modified repeated trust game in which participants viewed facial photographs of an investment partner (the Trustee), decided an amount to allocate to the Trustee for investment, and then learned the proportion of the investment gain the Trustee returned to the participant. These outcomes defined the partner as Trustworthy, Moderately Untrustworthy, or Untrustworthy. We examined whether a change in trustworthiness elicited a physiological arousal response, whether electroencephalographic (EEG) correlates of attitude change previously identified in the literature could serve as markers of learned trustworthiness, and whether a data-driven multivariate machine learning approach to the EEG data could reveal biomarkers not previously identified in the literature. We also tested whether a second-order conditioning procedure in which Trustee photographs used in the repeated trust game (CS1 stimuli) were paired with novel photos (CS2 stimuli) changed attitudes toward the individuals represented in the CS2 stimuli and whether such putative changes in attitude also manifest in physiological measures. An initial EEG experiment (Exploratory experiment) provided data for an analysis of potential biomarkers of trust behavior.					
15. SUBJECT TERMS Electroencephalography, Cognitive Processes, Social Psychology, Cognitive Neuroscience					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES 67	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			AHLERS, KRISTOPHER
Unclassified	Unclassified	Unclassified	SAR		19b. TELEPHONE NUMBER (Include area code) 315-227-7009

Final Report for AOARD Grant FA 2386-14-1-0018

Using Brain-State Information to Facilitate Conditioned Attitude Formation
March 31st, 2018

Principal Investigator: Trevor B. Penney

Current Address and Affiliation:

Department of Psychology, The Chinese University of Hong Kong
Shatin, New Territories, Hong Kong SAR
penney@cuhk.edu.hk
trevorbpennney@gmail.com

Office tel. +852 3943 9220

Mobile tel. +852 9685 9660

Fax. +852 2603 5019

Affiliation of PI during project period

Department of Psychology, National University of Singapore
9 Arts Link, Singapore 117 570

Co-PI: YEN Shih Cheng

Department of Electrical and Computer Engineering
National University of Singapore
Singapore

Co-PI: Michelle SEE

Department of Psychology
National University of Singapore
Singapore

Period of Performance: 09/30/2014 – 03/29/2017

Abstract/Executive Summary:

Humans use facial features and emotional expressions to judge whether others can be trusted to engage in fair social and economic exchanges. These initial judgments can be changed via learning and are affected by the outcome of direct economic exchanges such as those modeled by laboratory trust games. Here, we used a modified repeated trust game in which participants viewed facial photographs of an investment partner (the Trustee), decided an amount to allocate to the Trustee for investment, and then learned the proportion of the investment gain the Trustee returned to the participant. These outcomes defined the partner as Trustworthy, Moderately Untrustworthy, or Untrustworthy. We examined whether a change in trustworthiness elicited a physiological arousal response, whether electroencephalographic (EEG) correlates of attitude change previously identified in the literature could serve as markers of learned trustworthiness, and whether a data-driven multivariate machine learning approach to the EEG data could reveal biomarkers not previously identified in the literature. We also tested whether a second-order conditioning procedure in which Trustee photographs used in the repeated trust game (CS1 stimuli) were paired with novel photos (CS2 stimuli) changed attitudes toward the individuals represented in the CS2 stimuli and whether such putative changes in attitude also manifest in physiological measures. An initial EEG experiment (Exploratory experiment) provided data for an analysis of potential biomarkers of trust behavior. Biomarkers identified in this stage were then subjected to confirmatory analysis using data obtained in a replication of the first experiment with new participants (Confirmatory experiment).

Behavioral Results

In both the Exploratory and Confirmatory datasets, repeated association with positive trust game outcomes resulted in increased perceived trustworthiness, as indicated by larger transfer sums in a one-shot trust game. Additionally, the economic exchange outcomes affected subjective ratings of trustworthiness, warmth, global attitudes toward the person, attractiveness, and competence. Importantly, when participants were aware of the contingency between Trustee face and game outcome, the second-order conditioning procedure affected performance in a one-shot trust game in that participants sent larger sums to CS2 faces that had been paired with trustworthy CS1 faces. In addition, it affected subjective judgments of trustworthiness, global attitudes, valence, warmth, attractiveness, and competence. Self-reported arousal was also increased for Untrustworthy and Trustworthy faces, but skin conductance measures of physiological arousal did not differ among conditions.

Electrophysiological Results

The late positive potential (LPP) event related potential (ERP) response recorded during the trust game was larger for Trustworthy faces, as compared to Untrustworthy faces, in both the Exploratory and Confirmatory datasets. Machine-learning analysis of the ERP data revealed that a model trained on

data recorded at posterior electrodes could discriminate perceived face trustworthiness above chance and discriminate among conditions for the new data recorded in the confirmatory experiment. Hence, these neural signals contained information that robustly encoded trustworthiness information learned about faces. Multivariate analysis of EEG power revealed that classifiers trained on beta band power at posterior sites discriminated between Untrustworthy and Trustworthy faces in both the Exploratory and Confirmatory datasets. For the second-order conditioning task, some channels discriminated between conditions, but classifiers trained with these data did not perform above chance on the confirmatory dataset.

In sum, the behavioral results indicate that robust changes in perceived trustworthiness, and a range of social judgments can be achieved through first-order (i.e., Trust game) and second-order conditioning procedures. Moreover, ERP and EEG markers indicate a participant's trust categorization of interaction partners following first-order conditioning, but are unable to accurately predict categorization behavior following second-order conditioning. Overall, the evidence indicates that ERP/EEG biomarkers offer promise as a tool for predicting trust attitudes/behavior, but that the sensitivity of these measures is context/task dependent.

Introduction

Humans use facial features (Oosterhof & Todorov, 2008) and emotional expressions (Boone & Buck, 2013; Tortosa et al., 2013) to judge whether others can be trusted to engage in fair social and economic exchanges. However, these initial judgments are dynamic and can be changed via learning (Chang et al., 2010). Notably, they are affected by the outcome of direct economic exchanges such as those modeled by laboratory trust games (Berg et al., 1995). The primary goals of the present work were to investigate how trustworthiness judgments can change through learning and to attempt to identify biological markers of this learning (i.e., electroencephalographic and skin conductance measures).

First, we attempted to replicate earlier results from laboratory trust games, but also investigated whether trust game outcomes could affect attitudes and judgments relevant to social interactions such as warmth, competence, attractiveness, and dominance (Fiske et al., 2007; Todorov et al., 2015). To this end, we used a modified repeated trust game (Berg et al., 1995; Johnson & Mislin, 2011), in which participants viewed facial photographs of an investment partner (hereafter termed the Trustee), decided an amount to allocate to the Trustee for investment, and then learned the proportion of the investment gain the Trustee returned to the participant. The magnitude of the returned amount served as an indicator of the Trustee's trustworthiness in the sense that a Trustee who returned nothing would, by definition, be less trustworthy than a Trustee who returned some portion of the gain. As the game involved repeated trials with the same set of Trustees, the investment amount the participant contributed across trials to a particular Trustee revealed the development of trust in that person during the game. In addition, subjective ratings were obtained after game completion to assess whether and how the trust behavior exhibited by a particular Trustee changed the perceived trustworthiness and associated social judgments about that Trustee.

Second, we tested the extent to which Evaluative Conditioning can change attitudes toward potential interaction partners. Evaluative Conditioning (De Houwer et al., 2001; Olson & Zanna, 1993) is similar to Classical/Pavlovian Conditioning in that the learning process involves repeated pairings of a valenced stimulus (unconditioned stimulus - US) with an initially neutral stimulus (NS). This pairing changes the attitude to the NS, usually making it more similar to the attitude held toward the US (Houwer et al., 2001; Sweldens et al., 2010). Following successful conditioning, the neutral stimulus is referred to as a conditioned stimulus or CS (in this report we use the term CS to refer to the initially neutral stimuli both before and after successful conditioning). In Classical Conditioning paradigms, an established CS can be used to condition a new CS, a process called second-order conditioning (e.g., Rescorla, 1980; Walther, 2002). There is some evidence that a similar second-order conditioning process can occur with Evaluative Conditioning (e.g., Hofmann et al., 2010; Walther, 2002), but it

remains unclear how robust these effects are. Moreover, whether second-order conditioning can change social judgments that are not tightly linked to economic behavior in a direct interaction is unknown. We also examined the role of explicit awareness of the relation between the CS and US because it is considered an important moderating variable in Evaluative Conditioning (Baeyens et al., 1990; Pleyers et al., 2007; Hofmann et al., 2010).

To examine second-order conditioning, we paired the Trustee faces seen on Day 1 with new Trustee faces presented on a subsequent day (Day 2). After conditioning, participants' judgments were recorded to determine whether a transfer of attitudes and change in social judgments had occurred. These new Trustee faces were also used in a one-shot trust game to determine whether potential new attitudes altered the participant's economic behavior toward the Trustees.

Third, we attempted to identify physiological markers of trustworthiness and associated judgments. Pursuing this goal comprised three elements. First, we examined whether a change in trustworthiness elicited a physiological arousal response. Second, we examined neural correlates of attitude change previously identified in the literature and assessed whether they could serve as markers of learned trustworthiness. Third, we applied a data-driven multivariate machine learning approach (e.g., Haxby et al., 2001; Haynes & Rees, 2006) to the electroencephalographic (EEG) data in an effort to discover biomarkers not previously identified in the literature. These elements are addressed in turn below.

i. Physiological Arousal

Physiological arousal, whether measured indirectly through behavioral ratings (Bradley & Lang, 1994) or directly through skin conductance measurements (Boucsein, 2012; Critchley et al., 2000), indexes the preparation of cognitive and motor responses to motivationally relevant stimuli in the environment (Frith, 1983; Flykt et al., 2007). These measures have been associated with differences in attitudes (Oxley et al., 2008), and may constitute potentially useful markers for attitude change. We recorded arousal responses during both the Day 1 trust game and the Day 2 second-order conditioning procedure, and assessed whether changes in trustworthiness correlated with arousal responses.

ii. Neural Correlates of Attitude Change

The late positive event-related potential (LPP) has been repeatedly linked to attitudes (Cacciopo et al., 1996; Olofsson et al., 2008) and the motivational relevance of a stimulus (Schupp et al., 2000). More recently, it has been associated with trustworthiness judgments, with evidence for an enhanced LPP elicited by untrustworthy facial features (Marzi et al., 2014; Yang et al., 2011). Together these findings point to the LPP as a possible candidate marker for learned trustworthiness.

iii. Multivariate Machine Learning

Multivariate Machine Learning approaches have been used successfully with EEG to identify neural correlates of pain perception (Shulz et al., 2012) and decision-making (Bode et al., 2012). Machine learning classifiers can identify features that discriminate among conditions (Chung et al., 2015), resulting in the discovery of novel markers that would otherwise not be suggested by prior knowledge. Consequently, they provide predictions for new data, which allows evaluation of whether the discovered markers are stable across individuals in the population. Here, machine learning approaches were applied to both event-related EEG data and the oscillatory power index of the EEG.

Two EEG experiments were conducted. The first experiment provided data for an exploratory analysis of potential biomarkers of trust behavior. Putative biomarkers identified in this stage were then subjected to confirmatory analysis using data from the second experiment (confirmatory study). This Exploratory – Confirmatory approach was selected in light of substantial recent evidence that many highly influential results in the behavioral sciences are not particularly robust.

Exploratory Study

Methods

Participants

We recruited 36 ethnically Chinese participants from the student body of the National University of Singapore through announcements on a student web portal. Participants were excluded whenever they had missing data from one of the two test sessions, failed to follow task directions, or excessive EEG artifact resulted in too few usable trials. Specifically, 4 participants were excluded because of technical issues during data recording, 1 for instruction non-compliance, and 2 because more than 50% of trials were rejected in any one condition. This resulted in a final sample of 29 participants (8 males), who had a mean age of 22 years ($SD = 2$); 28 of the participants were right-handed according to self-report.

Stimuli

Face stimuli comprised 12 color photographs of male and female ethnically Asian faces obtained from the Chicago face database (Ma et al., 2015). The models had emotionally neutral facial expressions, appeared on a uniform white background wearing plain gray shirts, and all external features (e.g., hairstyle and clothing) were retained. Stimuli were initially selected based on the norming data provided in Ma et al. (2015). Specifically, we selected 60 images that were rated within 1 point from midscale (7-point scale) on Trustworthiness, and for which facial emotion intensity was judged lower

than midscale on six emotions (anger, happiness, sadness, surprise, disgust, fear). In addition, 80 ethnically Chinese participants (40 females) from the National University of Singapore, who did not take part in the EEG experiments, also rated these stimuli on Trustworthiness (9-point scale). These ratings covered a narrow range (3.0 - 4.2) and were lower than those of the norming data (in scale-range normalized values, $M = 0.33$ and $SD = 0.05$ versus $M = 0.46$ and $SD = 0.06$; $t(38) = 7.44$, $p < .001$). A subset of these images was selected to create two sets of 6 stimuli each, which were used as either CS1 or CS2 stimuli. There were no significant differences in Trustworthiness ratings between the sets (Set 1: $M = 3.7$ and $SD = 0.4$; Set 2: $M = 3.6$ and $SD = 0.4$; $t(9.9) = 0.58$, $p = .575$). For all experimental sessions except the contingency test, the presented images subtended $8.44^\circ \times 10^\circ$ degrees of visual angle.

Design and Procedure

Day 1

Initial Conditioning: Trust Game.

The main experimental session on Day 1 comprised presentation of a CS1 followed by a trust game on 50% of the trials. The outcome of the trust game, which varied in reward value, served as the US. This strategy has been used successfully to study learning of trustworthiness in repeated trust games.

In the game, the participant was the first mover/investor with each Trustee represented by a CS1 face. On each trial, the participant transferred a portion of a \$10 initial stake to the indicated Trustee, in increments of \$1, and kept the rest. In the version of the game used here, participants were not permitted to defect so had to allocate at least \$1 for transfer on each trial. This ensured that they would always experience the scheduled Trustee behavior, which was necessary to have full control of the conditioning procedure. The amount allocated by the participant was multiplied by 4 and transferred to the Trustee with the Trustee's response immediately presented to the participant. Three types of Trustee Response defined the conditions of interest: (1) an Untrustworthy condition, in which the Trustee did not return anything, (2) a Moderately Untrustworthy condition, in which the Trustee returned the initial allocation only (e.g., \$5), but none of the earnings, and (3) a Trustworthy condition in which the Trustee returned twice the allocation (e.g., \$10). Each Trustee's behavior was fixed throughout the session.

The game procedure was explained to the participants. They were told that although they were not playing with actual people in real-time, Trustee behavior mimicked responses given in a real game by the person in the photograph and that a participant's earnings for a given trial was the sum of the amount they decided to keep and the amount returned by the Trustee. They were reminded that their goal was to maximize their earnings from each trial, and were told that the earnings from a few

random trials selected at the end of the experiment would contribute to their final participant payment. Specifically, these earnings would be pooled with their Day 2 trust game earnings, and a fraction of the total would be given to them at the end of the experiment on Day 2.

Each game trial unfolded as follows. CS1 presentation was followed by a fixation cross, and then a screen appeared indicating that \$10 was available for allocation. They entered their allocation decision by moving a cursor along a scale displayed at the bottom of the screen, and confirmed their choice by pressing the space bar. The initial cursor position on each trial was random. The participant's decision was briefly presented on the next screen, followed by the outcome screen presenting the Trustee's decision (i.e., to keep all, to return the initial investment, or to share), along with the corresponding sums the Trustee was keeping and the participant was receiving. The trial ended with a fixation cross. Participants played 20 trust games with each CS1/Trustee. The trial timeline is presented in Figure 1.

For 50% of the trials, CS1s were presented without a subsequent trust game. These shorter duration non-game trials were included so that there were sufficient CS1 presentations for EEG analysis, but without making the experiment too long. There were 80 CS1 presentations per condition and 240 trials in total. At the beginning of the session, participants engaged in one trust game with each of the CS1s. This constituted an initial conditioning step, which ensured that the first presentation of all CS1s was followed by a game, thereby reducing variability in learning among conditions. To control the distribution of conditions over time during the rest of the session, we randomized the position of each condition over groups of 12 trials, making each condition equally likely in that interval. In addition, a given condition was never presented more than two times in succession. For each condition, there were two CS1s, 1 male face and 1 female face. CS1 assignment to each condition was counterbalanced across participants. There were four mandatory 30 s breaks during the session.

Participant attentiveness was ensured and measured by requiring them to press the space bar whenever an upside-down face appeared on the screen. Approximately 8% of trials comprised this target face, which was not presented at other times in the experiment. Before the session, participants practiced this task and the trust game on 8 trials with stimuli that were not included in the main experiment.

Behavioral Ratings.

At the end of the trust game session, participants exited the experiment room and washed their hair to remove the EEG electrode gel. Afterwards, they returned to the experiment room and rated each of the presented CS1s on 8 dimensions using a series of 9-point scales. The dimensions were Trustworthiness, Dominance, Attractiveness, Competence, Warmth, Arousal, Valence, Attitude, and Subjective Ambivalence. Most dimensions were rated using two items, but Arousal and Valence were rated with one MANIKIN item each (Bradley & Lang, 1994) and Subjective Ambivalence was rated

with three items (Priester & Petty, 1996). In earlier studies, Trustworthiness was typically measured using a single item (e.g., Oosterhof & Todorov, 2008), however, we decided to use a second item, a “reliability” judgment, to determine whether it could provide a more stable assessment of the Trustworthiness construct.

Day 2

Second-order Conditioning.

In the Day 2 session, participants were presented with CS2 faces either alone or paired with a CS1 face that had been presented the day before. Fifty percent of the trials were CS2-CS1 conditioning trials. On these trials, the CS2 face was presented for 1.5 s and immediately followed by a CS1 face for 1.5 s. The CS2-CS1 presentation was followed by a fixation cross. A given CS2 was always paired with the same CS1. Hence, the condition a CS2 belonged to was determined by the condition the CS1 face belonged to on Day 1. As on Day 1, CS2-CS1 pairings defined three Trustee conditions: Untrustworthy, Moderately Untrustworthy, and Trustworthy. There were two CS2s in each condition, a male face and a female face, which were paired with the CS1 of corresponding gender. Participants also performed the same target detection task as on Day 1.

The close temporal proximity of the CS2 and CS1 images in paired trials could lead to smearing of the EEG signal during processing, so we did not analyze these trials. Instead, to evaluate the EEG response to CS2s we examined the trials in which they were presented alone, but preceded and followed by a fixation cross. Across the session, each CS2 was presented alone 40 times. This resulted in 80 trials per condition, for a total of 240 CS2 Alone trials and 240 CS2-CS1 trials. As on Day 1, condition randomization was performed over 12 trials, equally mixing CS2 Alone and CS2-CS1 trials and the experimental session started with 6 paired trials that functioned as an initial exposure to the CS2-CS1 pairings. CS2 assignment to each condition was counterbalanced across participants. For each participant one set of faces was used for the CS1s while the other was used for the CS2s, but across participants all available faces appeared as both CS1 and CS2 stimuli. The experiment started with a practice block of 12 trials on stimuli not used in the main experiment. There were four mandatory 20 s breaks.

Behavioral Ratings.

Trust Game

As on Day 1, after the session ended participants completed a set of ratings. First, they engaged in a trust game. This trust game was identical to the Day 1 game with the exception that (1) it was one-shot, (2) participants were allowed to defect, and (3) there was no feedback on the Trustees’ behavior. Other than these three points, participants received the same instructions as for the Day 1 trust game. The

goal of this test was to provide a direct measure of trustworthiness. It aimed at assessing how trustworthy each CS2 was perceived to be, by measuring how much the participant would send in a one-shot game given the possibility to defect without consequence. Previous studies have shown a strong correlation between offers in such games and explicit face trustworthiness ratings (e.g., van 't Wout, 2008; Stirrat et al., 2010). This direct assessment of trustworthiness was followed by a series of explicit judgments identical to those presented on Day 1. Participants performed the game, then completed ratings of the CS2s, before going through the same procedure with the CS1s.

Contingency Awareness Test

After completing the ratings, participants were tested for their awareness of CS2 - CS1 contingencies. Each CS2 was presented on the left side of the screen once, while the 6 possible CS1s were simultaneously presented on the right side. Participants were asked to identify the CS1 that had been presented immediately after the CS2 during the second-order conditioning session. CS2 order of presentation, and CS1 positions on screen were random across participants.

Implicit Attitude Test

Next, participants completed an Affect Misattribution Procedure test (Payne et al., 2005). It assessed implicit attitudes toward each CS2 and CS1 face. The procedure was as described in Payne et al., (2005), with the exception that meaningless geometric shapes were used in place of Chinese characters, because the latter are meaningful for participants literate in Chinese and would interfere with the procedure. Each face was tested 3 times in random order.

Questionnaires. Next, participants completed questionnaires assessing Propensity to Trust (Evans & Revelle, 2008), Behavioral Inhibition, and Behavioral Activation (Carver & White, 1994), Demographics, and a funneled debriefing questionnaire. This was followed by a short questionnaire investigating participants' reasons for their judgments (used as a demand characteristic check). Participants were then given their earnings for the games and reimbursed for their time.

Skin Conductance

Electro-dermal activity (EDA) was recorded using a Biosemi Active2 system. Data were recorded from two flat Ag–AgCl electrodes of 10 mm diameter placed at the distal phalanges of digits II and III of the hand that was not used for responding. Procedures followed common recommendations (Roth et al., 2012). Data were resampled at 16 Hz, and analyzed by Continuous Decomposition Analysis using the Ledalab toolbox version 3.4.8 (Benedek, & Kaernbach, 2010). The EDA signal was decomposed into a tonic component and a phasic component, that is thereby free of baseline fluctuations. Responses were scored as the maximal phasic response between 1 and 6 seconds post stimulus onset. All participants displayed a response in at least 50% of the trials. Trials without a response were

excluded from analysis. A log transformation was applied to the resulting data. Three participants were excluded because of technical issues with EDA recording.

EEG

Recording

The EEG signal was recorded from 64 Ag/AgCl electrodes mounted in an elastic cap following the modified 10/20 system. The electro-oculogram was recorded from three electrodes placed at the left and right outer canthus and below the dominant eye. The EEG data were recorded with a Biosemi Active2 system at a sampling rate of 1024 Hz. The system uses a common mode sense electrode as recording reference and the signal is low-pass filtered using a digital sinc filter (order 5, -3dB cutoff at $1/5^{\text{th}}$ the sampling rate).

Raw EEG preprocessing

Raw EEG data were pre-processed with EEGLAB version 13.5.4b (Delorme and Makeig, 2004). The data were downsampled to 512 Hz, high-pass filtered (pass band edge = 0.2 Hz, -6dB cutoff = 0.1 Hz, transition band = 0.2Hz, 8449 points Hamming windowed sinc FIR filter), and cut into epochs time-locked to each image and extending from -2.5 s to 4 s after stimulus onset. Epochs were visually examined and those with excessive muscle or movement noise were discarded. Data were then subjected to an Independent Components Analysis (ICA) using the AMICA toolbox version 15 (Palmer et al., 2007). ICA weights and sphering matrices were derived from the original data after low-pass filtering at 1 Hz to remove slow drift, and resulting matrices were then adjoined to the original, non-filtered, epochs. Resulting components that captured eye movement, eye blink, and heart signal artifacts were visually identified based on weight topography and time course and eliminated from the data. Data were then back-projected to constitute ICA-pruned epochs for further ERP and EEG analyses.

ERP preprocessing

For ERP analysis, ICA-pruned epochs were further examined and trials containing excessive movement artifacts were discarded. Data were then low-pass filtered (pass-band edge = 20 Hz, -6 dB cutoff = 22.5 Hz, transition band = 5 Hz, 339 points Hamming windowed sinc FIR filter), re-epoched to -200 ms to +1500 ms peri-stimulus, baseline-corrected (200 ms pre-stimulus), and re-referenced to averaged mastoids. Before statistical analysis, mean amplitude was computed for each time window of interest by trial and participant. The average number of trials by condition was 64 (SD=8; range = 43–80) for Day 1, and 63 (SD=8; range = 44–75) for Day 2.

EEG Power preprocessing

For EEG power analysis, ICA-pruned data were re-referenced to average reference, baseline-corrected

(-1.5s to 0 s pre-stimulus), and epochs with excessive artifacts visually identified and rejected. The average remaining number of trials by conditions was 66 (SD = 8; range = 44 – 80) for Day 1 and 63 (SD = 8; range = 43 – 80) for Day 2. Power was computed using a continuous decomposition as implemented in the “Morlet wavelet” procedure of EEGLAB. We used a set of wavelets that captured frequencies from 1 Hz to 80 Hz in steps of 1 Hz ($c = 2^{-16}$, increasing linearly). This covered all frequency bands from theta to high gamma. Wavelet decomposition was run on 512 Hz data to benefit from maximal precision in estimation, but parameters were estimated every 1/64 s leading to an output at a temporal resolution of 64 Hz which covered the peri-stimulus interval from – 800 ms to 2000 ms. EEG power was computed from the wavelet transform at each frequency and time point.

Univariate Statistical Analysis

We used different analysis strategies for the Exploratory and Confirmatory datasets. Although some of the analyses applied to the Exploratory dataset were motivated by hypotheses derived from the literature, the overall approach was exploratory because we investigated effects in the data that we had not initially hypothesized. Hence, in the following we present all analyses as exploratory for easier presentation of the results. For all tests, we report effect size estimates, confidence intervals around these estimates, and p -values. We did not perform null hypothesis testing on these data because the risk of Type I error would be greatly inflated, and attempts to correct for multiple comparisons would reduce statistical power, which is undesirable for an exploratory study. We did, however, choose a criterion to determine candidates for interesting effects to be tested in the Confirmatory dataset.

For estimated effects, we deemed that there was sufficient evidence for an effect to be tested in the Confirmatory dataset when its 95 % CI excluded zero. For effects evaluated by model comparison, we relied on p -values because this allows use of a more intuitive cutoff than alternative model comparison measures. We used the traditional cutoff of .05, but interpreted it as providing weak evidence for a true effect. We chose this threshold because we believe most researchers would expect an effect associated with that level of evidence to be explored further in the Confirmatory dataset. We also provide p -values for estimated effects, but we consider these values as a tool to measure the relative degree of evidence for competing effect, and therefore they should not be interpreted as they would be for confirmatory research.

Confirmatory Analyses

The planned tests performed on the Confirmatory dataset were those identified in the Exploratory dataset. That is to say, all tests performed were planned before data analyses started on the Confirmatory dataset, and the data in the Confirmatory dataset did not influence which tests were conducted. No modifications were made to the analyses other than those used for dealing with model convergence issues when they arose.

Mixed Model specification

Statistical analyses were performed in R 3.3.2 (R core team, 2016) from the Microsoft R Open distribution with all packages from a CRAN repository snapshot taken on 2016-11-01. Data were analyzed using linear mixed effect models (LMM) fitted with the lme4 package (Bates, Maechler, Bolker, & Walker, 2015). Effects of interest were tested by model comparison using an ANOVA strategy with Type III sum of squares. For selected models, contrasts of parameter estimates and their 95% confidence intervals were computed using the lmerTest packages (Kuznetsova, Brockhoff, & Christensen, 2015). For all tests, p -values and confidence intervals were computed using degrees of freedom derived by Satterthwaite approximation. Reported effect sizes were derived from the models and are reported in original units with their 95% confidence interval. For model comparison, a maximum likelihood criterion was used for fitting. For analyses that required unbiased variance estimation, a restricted maximum likelihood criterion was used.

For behavioral, SCR, and ERP analyses we modeled a maximal random effect structure with intercept and all slope variances and covariances (Barr et al., 2013). Hence, unless otherwise noted, we modeled random intercept and slopes by participant and also modeled crossed random effects by item (Baayen et al., 2008). This allowed us to analyze data from single CS items in regression analysis, and ensured that our findings would generalize to CSs not included in the study. For behavior and SCR, we modeled random intercept and all slopes by CS. For ERP data, however, the fixed effect part of the model was typically larger, and the resulting random effect structure was too complex to be supported by the data. We consequently adopted a more flexible strategy and modeled a random intercept by item with a random intercept by channel when appropriate (Bates et al., 2015; Payne et al., 2015).

When models failed to converge, we simplified the random effects structure (Bates et al., 2015). This was done after attempting to restart optimization from earlier estimates or re-parameterizing the model by releveling problematic regressors. When this failed, we omitted random effects with lower variance one by one, until convergence was reached. In our data, variances of omitted effects were typically estimated near zero.

Event Related Potential (ERP) Analyses

For both Day 1 and Day 2, we analyzed the late positive potential (LPP) in its centro-posterior and frontal aspects (Cacciopo et al., 1996), in the left and right hemisphere, from Central (C1, C3, C5; C2, C4, C6), Posterior (P1, P3, P5; P2, P4, P6), and Frontal (AF7, F3, F5; AF8, F4, F6) channels. We included a frontal region because visual examination of the ERP grand-average indicated a frontal aspect of the late positive potential response to Trustworthiness at the anterior left and right sites (Cunningham, et al., 2005). Following earlier work (Thiruchselvam et al., 2011), we analyzed the

average LLP amplitude in early (500–900 ms) and late time windows (900–1500 ms). When the statistical modeling strategy differed between analyses, details are provided in the Results section.

Multivariate Classification of EEG

General strategy

We trained classifiers for each EEG channel, which could then use temporal patterns of activity across whole trials to discriminate conditions. An alternative approach is to train classifiers at each time point, while using scalp spatial pattern as discriminative features. While this is useful to track moment-to-moment detection accuracy, it did not fit our goal because spatial information is coarse in EEG, and carries limited information. However, EEG affords high temporal resolution, and we wanted to leverage on the rich information that is available in the time dimension to derive accurate discrimination.

The EEG power data were analyzed according to the same principles in that we attempted to use all available temporal information for discrimination. Here, classifiers were trained at each channel and each frequency. This allowed us to identify channels and frequency bands that could reliably discriminate among conditions using temporal information in the EEG signal.

For both analyses, classifiers were trained on data pooled across participants. In multivariate classification studies of neural signals, classifiers are typically trained and their performance evaluated on a single participant's data (Etzel et al., 2013; Haxby et al., 2014). While the identified patterns are participant-specific responses, or codes, that are robust for prediction within-participants, they might differ from participant to participant and may not be representative of the general population. Here, we sought instead to identify patterns that would emerge over several participants, and would likely allow discrimination of responses from new participants.

Classification

Classification was performed under MATLAB R2015a (v. 8.5.0.197613; The MathWorks Inc., Natick, MA) using a Support Vector Machine (SVM) algorithm optimized for linear classification as implemented in the Liblinear toolbox v. 2.11 (L2-regularized, L1-loss SVM; Fan et al., 2008; Chiang et al., 2016). The c parameter was set to 1 for the majority class. To correct for imbalances in trial numbers between conditions, the value of c for the minority class was increased by a factor equal to the imbalance ratio (Akbari et al., 2004; Veropoulos et al., 1999). We opted for this strategy instead of oversampling the minority conditions because class-specific adjustment of the c parameter has been shown to be asymptotically equivalent to oversampling (Cawley & Talbot, 2001), while being computationally less expensive. Overall imbalance ratios between condition pairs were small (max ratios: Day 1, ERP = 0.52 %, EEG = 1.26 %; Day 2, ERP = 0.27 %, EEG = 1.26 %).

Classification was performed independently at each electrode, using time points as features. This returned a topographic map of classification performance. Classification performance on the training set was evaluated using k-fold cross-validation. Each fold contained the trials from one participant, which allowed testing of classifier performance on new participants. To do so, we trained classifiers on trials from all but one participant, and tested classification on trials from the remaining participant. To increase the signal to noise ratio, each participant's data were averaged over groups of 5 consecutive trials (Bode et al., 2012). Three sets of classifier were trained, one for each pairwise comparison of the Trustee variable. They classified the following conditions, Untrustworthy vs. Trustworthy, Untrustworthy vs. ModeratelyUntrustworthy, and ModeratelyUntrustworthy vs. Trustworthy. Classification performance was measured using the area under the receiver operating characteristic curve (AUC) computed over trials from raw decision values. This measure integrates the false positive and false negative rates, and can thereby account for biases in classification that could emerge from class imbalance in the testing set (Huang et al., 2015).

Classification Statistical Analysis

For each pairwise comparison we determined whether classification performance was statistically different from chance (i.e., $AUC = .5$). First, we computed the average cross-validation AUC across participants, and the associated confidence intervals using the influence curve approach (LeDell et al., 2015). This allowed us to derive p -values for each classifier. Because we trained one classifier per channel for the ERP data, and one per channel and frequency for the EEG, using these raw estimates would entail a high false positive risk. To protect against this, we further controlled for the false discovery rate at 5% (BH, 1995). For the EEG classification, we additionally applied a cluster threshold under the assumption that because of the spatial and frequency smearing of the recording and processing steps, effects due to noise should be spread over several electrodes on the scalp and should cover several frequency bins. For individual channels, we first excluded clusters that spanned less than three frequency bins, and further excluded clusters that extended over less than two neighboring electrodes.

For electrodes and frequencies that showed evidence of discrimination power based on statistical tests, we additionally confirmed the results using permutation tests. The goal was to assess how the performance of the classifiers compared with classifiers trained on data with randomly shuffled labels from which no class information could be extracted reliably. For each pairwise comparison identified using a test, a distribution of discrimination performance under shuffled labels was constructed using 1000 permutations. All labels were permuted for each participant individually, to retain the same stratification as the original data (Etzel & Braver, 2013). The AUC was computed for each permutation, and the resulting distribution was used to determine the statistical significance of the

AUC values derived from the original dataset. This procedure was restricted to channels and frequencies from which significant discrimination was found. Analyzing the whole dataset using permutation was computationally too expensive to be feasible.

Predictors of Learning Rate

During the Day 1 session participant responses to face stimuli in the game were recorded for each reinforced trial. This allowed derivation of the rate at which their judgment of the faces changed over time (i.e., learning rate) using a formal model of classical conditioning (Rescorla & Wagner, 1972). We aimed at determining whether the individual learning rate could be predicted from ERP and EEG signals. The formal model expresses the associative strength V of a CS toward the US on trial t as a function of the learning rate α , and the maximum achievable associative strength parameterized by λ : $V(t) = V(t-1) + \alpha[\lambda(t) - V(t-1)]$. The amount sent in the trust game was taken as a direct measure of associative strength V of a CS, we therefore expressed V , α and λ in these units. A model was fit to each participant's reinforced trial data by finding the value of α in the interval $[0,1]$ that minimized a least square criterion. We used the Limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm with bound constraints (L-BFGS-B; Byrd et al., 1995) as implemented in the `optim` function in R. Maximum associative strength λ was set to 1 for the Untrustworthy condition, and to 10 for the Trustworthy condition. Initial associative strength V was set to the mid-range value of 4.5. After individual α values were derived, participants were categorized as fast and slow learners based on a median split. This was done for each condition separately in order to account for the fact that participant learning rate could differ as a function of individual sensitivity to reward and punishment (Gray, 1987). Participants recorded for the Confirmatory dataset were split into two groups based on the Exploratory dataset median value.

Classification performance was evaluated using a leave-one-out approach. Decision-value output-by-classifier for all of the five-trial averages of a participant were summed to derive a participant's decision score. This was equivalent to a voting procedure count. All individual participant decision scores were then used together with actual group membership to build a ROC curve and compute discrimination performance using the AUC. To test for which channels this classification performance differed from the chance value of .5, single participant labels were randomly shuffled, and the AUC analysis performed repeatedly to build a permutation sample under the null (10 000 replicates). One-tailed permuted p -values derived from this procedure then entered a FDR correction with a 5 % FDR criterion.

Prospective Power Analysis

After analysis of the Exploratory dataset, but before starting acquisition of the Confirmatory dataset, we performed a prospective power analysis for some of the effects identified in the Exploratory

dataset. The goal was to determine the sample size that would give sufficient power to detect a true effect of the size estimated in the Exploratory dataset. Because there is no agreed upon standardized effect size measure for LMMs, we used a parametric bootstrapping approach (Bolker, 2008). This approach rests on the fact that data simulated from fitted models have characteristics of the original data as estimated from the model, including effect sizes for fixed effects, and associated variability estimated by random effects and residual variances. Repeated bootstrap sampling from the model should yield the long-run variability that one would encounter through repeated sampling from the underlying population, assuming the model is the correct one. The proportion of significant tests among test runs on these samples then constitutes an estimate of prospective power for the sample size, effect size, and associated variability under study. The analysis was performed using the *simR* package (version 1.0.2; Green & MacLeod, 2016) and in-house code implementing the tests of interest.

The power estimation was performed for the two primary effects of interest. First, for the behavioral data we examined power for detecting the difference in sum transferred between Trustworthy and Untrustworthy conditions when contingency was remembered. For the EEG data, we examined power for the effect of LPP in predicting the sum transferred on Day 2, when contingency was remembered. For both effects we computed the sample size required for 80% power for the test relevant parameters with $\alpha = .05$. For each test, power was computed for two effect sizes, the original effect size, and a smaller, lower bound effect size. The effect sizes used for power computation are justified below separately for behavior and EEG. The strategy was as follows. First and when required, the original fitted models were modified to change the estimated effect size to the target effect sizes. Then 2000 parametric bootstrap samples were drawn at each of five sample sizes ($N = 30, 36, 42, 48, 52$). Tests were then performed on each sample and the average prospective power computed as the proportion of statistically significant tests at $\alpha = .05$ across all bootstrap samples of a given sample size.

RESULTS

EXPLORATORY STUDY

Rating Variable Factor Structure

We recorded data for variables of interest in three rating sessions (CS1 Day1, CS1 Day 2, CS2 Day 2). As we expected some variables to be highly correlated, we investigated their relationship with principle components analysis (PCA) and oblimin rotation. Most measures had high loading on their respective variables. There was a problem with the “reliability” item included to measure Trustworthiness. It had low loading on all factors, and did not load with the other Trustworthiness item. We therefore followed earlier studies and analyzed Trustworthiness based on a single item.

CS1 - Trust game

Sum Transferred.

On Day 1, there was an effect of TrusteeType on the mean amount transferred to CS1 ($F(2, 29.8) = 203.55, p < .001$). Untrustworthy Trustees received less money than ModeratelyUntrustworthy and Trustworthy Trustees (Untrustworthy vs. ModeratelyUntrustworthy: $b = -3.2, CI_{95\%} = [-4.27, -2.12], t(29) = -6.09, p < .001$; Untrustworthy vs. Trustworthy: $b = -6.92, CI_{95\%} = [-7.62, -6.22], t(29.9) = -20.17, p < .001$) and ModeratelyUntrustworthy Trustees received less than Trustworthy Trustees ($b = -3.73, CI_{95\%} = [-4.80, -2.65], t(30.1) = -7.09, p < .001$).

On Day 2, there was an effect of TrusteeType on the mean amount transferred to CS1 ($F(2, 29.5) = 233.49, p < .001$). Untrustworthy Trustees received less money than ModeratelyUntrustworthy and Trustworthy Trustees (Untrustworthy vs. ModeratelyUntrustworthy: $b = -3.75, CI_{95\%} = [-5.17, -2.32], t(29.6) = -5.37, p < .001$; Untrustworthy vs. Trustworthy: $b = -8.9, CI_{95\%} = [-9.76, -8.04], t(28) = -21.19, p < .001$), and ModeratelyUntrustworthy Trustees received less than Trustworthy Trustees (ModeratelyUntrustworthy vs. Trustworthy: $b = -5.15, CI_{95\%} = [-6.57, -3.74], t(29.5) = -7.45; p < .001$).

CS1 Ratings

Effects of TrusteeType on ratings for both days were tested using LMMs with factors TrusteeType and Day.

Perceived Trustworthiness.

A significant effect of TrusteeType was observed ($F(2, 27.2) = 116.86, p < .001$), but there was no effect of Day ($F(1, 10.3) = 2.6, p = .145$) or a TrusteeType x Day interaction ($F(2, 19.3) = 0.77$).

Trustees in the Untrustworthy condition were rated as less trustworthy than Trustees in the ModeratelyUntrustworthy and Trustworthy conditions (Untrustworthy vs. ModeratelyUntrustworthy: $b = -2.83, CI_{95\%} = [-3.55, -2.11], t(26.2) = -8.1, p < .001$; Untrustworthy vs. Trustworthy: $b = -5.65, CI_{95\%} = [-6.42, -4.88], t(27.9) = -14.99, p < .001$). Trustees in the ModeratelyUntrustworthy condition were rated as less trustworthy than those in the Trustworthy condition (ModeratelyUntrustworthy vs. Trustworthy: $b = -2.82, CI_{95\%} = [-3.51, -2.13], t(22.9) = -8.5, p < .001$).

Attitudes. There was an effect of TrusteeType on global Attitudes toward CS1 ($F(2, 26.1) = 89.04, p < .001$), no effect of Day on Attitudes ($F(1, 14.2) = 0.03, p = .868$), but a significant TrusteeType x Day interaction ($F(2, 18.4) = 8.5, p = .002$). Significant effects were observed for all pairwise tests on

both days (all $ps < .001$). Day 1 effects were greater than Day 2 effects (Day 1: Untrustworthy vs. ModeratelyUntrustworthy: $b = -2.74$, $CI_{95\%} = [-3.49, -2]$, $t(25.6) = -7.61$; Untrustworthy vs. Trustworthy: $b = -5.3$, $CI_{95\%} = [-6.11, -4.48]$, $t(27.7) = -13.34$; ModeratelyUntrustworthy vs. Trustworthy: $b = -2.55$, $CI_{95\%} = [-3.27, -1.83]$, $t(22.9) = -7.31$; Day 2: Untrustworthy vs. ModeratelyUntrustworthy: $b = -1.91$, $CI_{95\%} = [-2.6, -1.22]$, $t(18.3) = -5.82$; Untrustworthy vs. Trustworthy: $b = -4.26$, $CI_{95\%} = [-5.04, -3.47]$, $t(27.8) = -11.12$; ModeratelyUntrustworthy vs. Trustworthy: $b = -2.34$, $CI_{95\%} = [-3.12, -1.57]$, $t(22.4) = -6.28$, all $ps < .001$).

Warmth. There was a significant effect of TrusteeType ($F(2,22) = 60.99$, $p < .001$), no effect of Day ($F(1,11.6) = 0.14$, $p = .714$), and a significant TrusteeType x Day interaction ($F(2,30.7) = 4.56$, $p = .018$). Trustees in the Untrustworthy condition were rated as less warm than those in the ModeratelyUntrustworthy and Trustworthy conditions and those in the ModeratelyUntrustworthy condition were rated less warm than those in the Trustworthy condition. Day 1 effects were greater than Day 2 effects (Day 1: Untrustworthy vs. ModeratelyUntrustworthy: $b = -1.79$, $CI_{95\%} = [-2.45, -1.13]$, $t(13.5) = -5.84$; Untrustworthy vs. Trustworthy: $b = -4.27$, $CI_{95\%} = [-5.07, -3.47]$, $t(27.9) = -10.94$; ModeratelyUntrustworthy vs. Trustworthy: $b = -2.48$, $CI_{95\%} = [-3.34, -1.62]$, $t(17.3) = -6.1$; Day 2: Untrustworthy vs. ModeratelyUntrustworthy: $b = -1.49$, $CI_{95\%} = [-2.12, -0.86]$, $t(19.5) = -4.93$; Untrustworthy vs. Trustworthy: $b = -3.54$, $CI_{95\%} = [-4.38, -2.7]$, $t(26.3) = -8.7$; ModeratelyUntrustworthy vs. Trustworthy: $b = -2.05$, $CI_{95\%} = [-2.99, -1.11]$, $t(19) = -4.56$).

Attractiveness. There were significant effects of TrusteeType ($F(2,18.9) = 12.94$, $p < .001$), and Day ($F(1,31.5) = 7.35$, $p = .011$), but no significant TrusteeType x Day interaction ($F(2,32.2) = 1.32$, $p = .281$). Untrustworthy condition Trustees were viewed as less attractive than both ModeratelyUntrustworthy and Trustworthy condition Trustees (Untrustworthy vs. ModeratelyUntrustworthy: $b = -1.16$, $CI_{95\%} = [-1.81, -0.51]$, $t(15) = -3.82$, $p = .002$; Untrustworthy vs. Trustworthy: $b = -1.73$, $CI_{95\%} = [-2.48, -0.98]$, $t(22.9) = -4.76$, $p < .001$), but little evidence for an attractiveness difference between Trustees in the ModeratelyUntrustworthy and Trustworthy conditions (ModeratelyUntrustworthy vs. Trustworthy: $b = -0.57$, $CI_{95\%} = [-1.24, 0.1]$, $t(19.1) = -1.77$, $p = .093$).

Competence. There was an effect of TrusteeType ($F(2,25.5) = 8.73$, $p = .001$), no effect of Day ($F(1,26.8) = 0.59$, $p = .448$), and no TrusteeType x Day interaction ($F(2,19.7) = 0.99$, $p = .390$). Trustees in the Trustworthy condition were considered more competent than those in the Untrustworthy condition (Untrustworthy vs. Trustworthy: $b = -1.29$, $CI_{95\%} = [-1.94, -0.64]$, $t(25.1) = -4.1$, $p < .001$) and ModeratelyUntrustworthy condition (ModeratelyUntrustworthy vs. Trustworthy: $b = -0.96$, $CI_{95\%} = [-1.62, -0.31]$, $t(26.4) = -3.02$, $p = .006$). There was no difference between Trustees in the Untrustworthy and ModeratelyUntrustworthy conditions (Untrustworthy vs.

ModeratelyUntrustworthy: $b = -0.33$, $CI_{95\%} = [-0.81, 0.15]$, $t(24.3) = -1.42$, $p = .168$).

Ambivalence. There was an effect of TrusteeType ($F(2,22.1) = 6.13$, $p = .008$), no effect of Day ($F(1,14.4) = 0.44$, $p = .518$) and no TrusteeType x Day interaction ($F(2,36.1) = 1.33$, $p = .277$). Ambivalence was greater in the Untrustworthy and ModeratelyUntrustworthy conditions as compared to the Trustworthy condition (Untrustworthy vs. Trustworthy: $b = 0.7$, $CI_{95\%} = [0.17, 1.23]$, $t(15.6) = 2.81$, $p = .013$; ModeratelyUntrustworthy vs. Trustworthy: $b = 0.76$, $CI_{95\%} = [0.25, 1.27]$, $t(25.1) = 3.07$, $p = .005$), but no difference between the Untrustworthy and ModeratelyUntrustworthy conditions ($b = -0.06$, $CI_{95\%} = [-0.57, 0.46]$, $t(24.9) = -0.23$, $p = .821$).

Dominance. The effect of TrusteeType ($F(2, 23.7) = 1.24$, $p = .306$) and the TrusteeType x Day interaction ($F(2,27.9) = 0.03$, $p = .967$) were not significant. There was a significant effect of Day ($F(1, 21.7) = 6.4$, $p = .019$) with slightly lower dominance ratings on Day 1 as compared to Day 2 ($b = -0.34$, $CI_{95\%} = [-0.63, -0.05]$, $t(20.6) = -2.48$, $p = .022$).

Arousal. There was an effect of TrusteeType ($F(2,28.2) = 5.85$, $p = .008$), and an effect of Day ($F(1,28.1) = 5.58$, $p = .025$), but no TrusteeType x Day interaction ($F(2,28.2) = 0.41$, $p = .670$). Follow-up analyses indicated increased arousal for the Trustworthy as compared to the ModeratelyUntrustworthy condition (ModeratelyUntrustworthy vs. Trustworthy: $b = -1.27$, $CI_{95\%} = [-2.13, -0.41]$, $t(27.9) = -3.01$, $p = .006$, $p < .001$), but no evidence for a difference between ModeratelyUntrustworthy and Untrustworthy (Untrustworthy vs. ModeratelyUntrustworthy: $b = 0.61$, $CI_{95\%} = [-0.08, 1.29]$, $t(26.9) = 1.81$, $p = .081$), or between Trustworthy and Untrustworthy (Untrustworthy vs. Trustworthy: $b = -0.66$, $CI_{95\%} = [-1.7, 0.38]$, $t(27.8) = -1.31$, $p = .202$).

Valence. There was an effect of TrusteeType ($F(2,26.6) = 41.05$, $p < .001$), no effect of Day ($F(1,15.7) = 0.18$, $p = .673$), but a TrusteeType x Day interaction ($F(2,21.2) = 3.52$, $p = .048$). Valence ratings were lower in the Untrustworthy condition as compared to the ModeratelyUntrustworthy and Trustworthy conditions, and lower for ModeratelyUntrustworthy than Trustworthy conditions. Effects were stronger for Day 1 than Day 2.

Day 1: Untrustworthy vs. ModeratelyUntrustworthy: $b = -1.75$, $CI_{95\%} = [-2.42, -1.09]$, $t(23.8) = -5.44$; Untrustworthy vs. Trustworthy: $b = -4.3$, $CI_{95\%} = [-5.24, -3.36]$, $t(29) = -9.37$; ModeratelyUntrustworthy vs. Trustworthy: $b = -2.55$, $CI_{95\%} = [-3.43, -1.67]$, $t(23.4) = -5.97$; Day 2: Untrustworthy vs. ModeratelyUntrustworthy: $b = -1.32$, $CI_{95\%} = [-2.05, -0.6]$, $t(22) = -3.77$; Untrustworthy vs. Trustworthy: $b = -3.47$, $CI_{95\%} = [-4.45, -2.49]$, $t(28) = -7.24$; ModeratelyUntrustworthy vs. Trustworthy: $b = -2.15$, $CI_{95\%} = [-2.99, -1.31]$, $t(26) = -5.25$, all $ps < .001$).

CS2 - CS1 Contingency Awareness

For the CS1 - CS2 contingency awareness test, participants had an average accuracy of 39.1% (CI_{95%} [26.9, 51.2]), with strong evidence that it differed from chance performance of 17%. However, there was high inter-participant variability (see Figure 2), so we examined whether explicit awareness of CS1 - CS2 contingency changed as a function of TrusteeType. It is possible that the contingency is remembered better for Untrustworthy as compared to Trustworthy CS1 stimuli. A generalized mixed effect logistic model was fit to raw accuracy data and modeled a fixed effect of TrusteeType condition, and random intercept and slope by participant, and random intercept by CS1. A likelihood ratio test against the intercept-only model did not support such an effect ($\chi^2(2) = 5.48, p = .065$).

CS2 Behavioral Responses

Effects of TrusteeType and Contingency Awareness on the behavioral responses to CS2 were examined. Specifically, we tested whether explicit awareness of the CS1 - CS2 contingency was necessary to observe an effect of TrusteeType on the amount transferred to the CS2 and on the ratings of the CS2 on various characteristics. Effects should be reflected in an interaction between TrusteeType and Contingency Awareness because enhanced contingency awareness should lead to decreased scores for Untrustworthy CS2s, but to increased scores for Trustworthy CS2s. To examine this effect, we fitted behavioral response models comprising fixed effects of Trust, Contingency Awareness performance, and their interaction.

Sum Transferred in Day 2 Trust Game

There was a significant main effect of CS1 TrusteeType on the sum transferred to paired CS2s ($F(2, 22.4) = 5.79, p = .009$), but no significant main effect of Contingency Awareness, $F(1, 16.2) = 0.28, p = .603$. Although the evidence for a TrusteeType x Contingency Awareness interaction was weak ($F(2, 16.7) = 2.82, p = .088$), because of our hypothesis we further investigated effects separately for items with and without contingency awareness (see Figure 4). There was no evidence for an effect of CS1 TrusteeType on the sum transferred in the absence of contingency awareness (Untrustworthy vs. ModeratelyUntrustworthy: $b = -0.67, CI_{95\%} = [-2.1, 0.77], t(29.1) = -0.95$; Untrustworthy vs. Trustworthy: $b = -0.54, CI_{95\%} = [-1.7, 0.62], t(25.4) = -0.96$; ModeratelyUntrustworthy - Trustworthy: $b = 0.13, CI_{95\%} = [-1.13, 1.38], t(39.1) = 0.2$; all $ps > .350$). For items for which contingency was remembered, however, there was evidence of larger transfer amounts in the Trustworthy as compared to the Untrustworthy condition (Untrustworthy vs. Trustworthy: $b = -3.26, CI_{95\%} = [-5.49, -1.03], t(14.3) = -3.13, p = .007$). There was no evidence for a difference between ModeratelyUntrustworthy and Untrustworthy conditions (Untrustworthy vs. ModeratelyUntrustworthy: $b = -1.54, CI_{95\%} = [-3.22, 0.14], t(19.9) = -1.92, p = .070$), or between ModeratelyUntrustworthy and Trustworthy conditions

(ModeratelyUntrustworthy vs. Trustworthy: $b = -1.72$, $CI_{95\%} = [-4.18, 0.74]$, $t(14.8) = -1.49$, $p = .158$).

Subjective Ratings

Perceived Trustworthiness. There was no evidence of an effect of CS1 TrusteeType on CS2 perceived trustworthiness ($F(2,7.8) = 2.28$, $p = .167$) or an effect of Contingency Awareness ($F(1,14.5) = 1.12$, $p = .308$). However, CS1 TrusteeType interacted with Contingency Awareness ($F(2,18.1) = 6.52$, $p = .007$). There was no evidence for an effect of CS1 TrusteeType on ratings in the absence of Contingency Awareness (Untrustworthy vs. ModeratelyUntrustworthy: $b = -0.3$, $CI_{95\%} = [-1.52, 0.92]$, $t(8) = -0.57$; Untrustworthy vs. Trustworthy $b = 0.2$, $CI_{95\%} = [-0.71, 1.11]$, $t(9.3) = 0.49$; ModeratelyUntrustworthy vs. Trustworthy: $b = 0.5$, $CI_{95\%} = [-0.39, 1.39]$, $t(12.3) = 1.22$; all $ps > .246$). However, for items for which contingency was remembered, there was evidence for higher trustworthiness ratings for the Trustworthy as compared to the Untrustworthy condition (Untrustworthy vs. Trustworthy: $b = -1.76$, $CI_{95\%} = [-2.96, -0.57]$, $t(12.3) = -3.21$, $p = .007$), but not for other comparisons (Untrustworthy vs. ModeratelyUntrustworthy: $b = -1.11$, $CI_{95\%} = [-2.61, 0.4]$, $t(7.6) = -1.71$, $p = .127$; ModeratelyUntrustworthy vs. Trustworthy: $b = -0.66$, $CI_{95\%} = [-2.01, 0.7]$, $t(6.4) = -1.16$; $ps > .127$).

Attitudes. There were no significant effects of CS1 TrusteeType ($F(2,7.8) = 2.28$, $p = .167$) or Contingency Awareness ($F(1,38.7) = 1.12$, $p = .297$) on global attitudes toward the CS2. However, there was a TrusteeType x Contingency Awareness interaction ($F(2,18.1) = 6.52$, $p = .007$). There was no evidence for an effect of CS1 TrusteeType on ratings in the absence of contingency awareness (Untrustworthy vs. ModeratelyUntrustworthy: $b = 0.04$, $CI_{95\%} = [-0.63, 0.71]$, $t(11.2) = 0.13$; Untrustworthy vs. Trustworthy: $b = 0.05$, $CI_{95\%} = [-0.55, 0.64]$, $t(14.3) = 0.17$; ModeratelyUntrustworthy - Trustworthy: $b = 0.01$, $CI_{95\%} = [-0.48, 0.49]$, $t(20.3) = 0.03$; all $ps > .870$). However, for items for which contingency was remembered, there was evidence for higher attitude ratings for the Trustworthy as compared to the Untrustworthy condition (Untrustworthy vs. Trustworthy: $b = -1.13$, $CI_{95\%} = [-1.96, -0.29]$, $t(15.4) = -2.88$, $p = .011$), and the ModeratelyUntrustworthy condition (ModeratelyUntrustworthy vs. Trustworthy: $b = -0.87$, $CI_{95\%} = [-1.72, -0.02]$, $t(8.2) = -2.35$, $p = .046$). There was no difference between Untrustworthy and ModeratelyUntrustworthy conditions (Untrustworthy vs. ModeratelyUntrustworthy: $b = -0.26$, $CI_{95\%} = [-1.18, 0.67]$, $t(9.1) = -0.62$, $p = .548$).

Warmth. There was no evidence for an effect of CS1 TrusteeType ($F(2,17.9) = 1.81$, $p = .192$) or Contingency Awareness ($F(1,30.7) = 2.25$, $p = .144$) on CS2 perceived warmth. However, there was weak evidence for a TrusteeType x Contingency Awareness interaction ($F(2,17) = 6.12$, $p = .010$). There was no evidence for an effect of CS1 TrusteeType on ratings in the absence of contingency awareness (Untrustworthy vs. ModeratelyUntrustworthy: $b = -0.22$, $CI_{95\%} = [-1.34, 0.9]$, $t(8.3) = -0.45$;

Untrustworthy vs. Trustworthy: $b = 0.25$, $CI_{95\%} = [-0.47, 0.96]$, $t(7.6) = 0.8$; ModeratelyUntrustworthy vs. Trustworthy: $b = 0.46$, $CI_{95\%} = [-0.31, 1.24]$, $t(12) = 1.3$; all $ps > .217$). However, for items for which contingency was remembered, there was strong evidence for higher ratings for the Trustworthy as compared to the Untrustworthy condition (Untrustworthy vs. Trustworthy: $b = -1.07$, $CI_{95\%} = [-1.76, -0.38]$, $t(24.3) = -3.19$, $p = .004$). There were no differences between the Untrustworthy and ModeratelyUntrustworthy ($b = -0.41$, $CI_{95\%} = [-1.38, 0.56]$, $t(13.9) = -0.9$) or between the ModeratelyUntrustworthy and Trustworthy conditions ($b = -0.66$, $CI_{95\%} = [-1.51, 0.18]$, $t(15.4) = -1.66$, $ps > .116$).

Ambivalence. There was no effect of Contingency Awareness on ambivalence toward CS2s ($F(1,18.5) = 0.01$, $p = .938$), but there was weak evidence for an effect of CS1 TrusteeType ($F(2,9.7) = 6.12$, $p = .019$), and strong evidence for a CS1 TrusteeType x Contingency Awareness interaction ($F(2,11.6) = 9.36$, $p = .004$). There was no evidence for an effect of CS1 TrusteeType on ratings in the absence of contingency awareness (Untrustworthy vs. ModeratelyUntrustworthy: $b < 0.01$, $CI_{95\%} = [-0.6, 0.59]$, $t(8.9) = -0.01$; Untrustworthy vs. Trustworthy: $b = -0.03$, $CI_{95\%} = [-0.62, 0.57]$, $t(9.2) = -0.1$; ModeratelyUntrustworthy vs. Trustworthy: $b = -0.02$, $CI_{95\%} = [-0.78, 0.74]$, $t(6.3) = -0.08$; all $ps > .924$). However, for items for which contingency was remembered, ambivalence ratings were significantly lower for the Trustworthy as compared to the Untrustworthy condition (Untrustworthy vs. Trustworthy: $b = 1.62$, $CI_{95\%} = [0.75, 2.49]$, $t(8.3) = 4.27$, $p = .002$), and for the ModeratelyUntrustworthy as compared to the Untrustworthy condition (Untrustworthy vs. ModeratelyUntrustworthy: $b = 1.08$, $CI_{95\%} = [0.17, 1.98]$, $t(5.9) = 2.92$, $p = .027$). There was no significant difference between Trustworthy and ModeratelyUntrustworthy conditions ($b = 0.54$, $CI_{95\%} = [-0.56, 1.65]$, $t(6.1) = 1.2$, $p = .273$).

Attractiveness, Dominance, Competence, Valence, Arousal. For these variables there was no evidence for effects of TrusteeType (all $ps > .083$), Contingency Awareness (all $ps > .471$), or for TrusteeType x Contingency Awareness interactions (all $ps > .162$).

Skin Conductance Response (SCR)

CS1 response Day 1

There was no evidence for an effect of TrusteeType on SCR responses to CS1 ($F(2,24.2) = 0.31$, $p = .736$). Results are presented in Figure 6.

CS2 response Day 2

There was no evidence for an effect of TrusteeType on SCR responses to CS2 ($F(2,149.6) = 1.43$, $p = .243$). Results are presented in Figure 6.

ERP Univariate Analyses

Late Positive Potential (LPP) response to CSI

For the posterior Late Positive Potential (LPP), Linear Mixed Models including the effects of TrusteeType, Region (central/parietal), and Laterality (left/right), and all higher order interactions were created separately for the early and late time-windows. For the frontal LPP, TrusteeType, Laterality and their interaction were included in the model.

Posterior LPP.

Early Time Window

In the early time-window, there were effects of Region ($F(1,33.7) = 26.13, p = 10^{-5}$), which reflected a larger LPP at central sites as compared to parietal sites ($b = 1.5, CI_{95\%} = [0.87, 2.14], t(27.4) = 4.83, p < .001$), and Laterality ($F(1,20.8) = 5, p = .036$). Moreover, the TrusteeType x Laterality interaction ($F(2,121.5) = 4.1, p = .019$) was significant, but not the three-way interaction including Region ($F(2,192.7) = 0.08, p = .919$), which suggests a stable effect across these regions. Tests of contrasts provided evidence for effects at the left hemisphere electrodes only, in which the LPP was significantly enhanced in the Trustworthy as compared to the Untrustworthy condition (Untrustworthy vs. Trustworthy: $b = -0.91, CI_{95\%} = [-1.66, -0.15], t(27.8) = -2.46, p = .02$), other contrasts involving the Moderately Untrustworthy condition did not suggest any effect (Untrustworthy vs. ModeratelyUntrustworthy: $b = -0.34, CI_{95\%} = [-0.76, 0.08], t(28.9) = -1.65, p = .110$; ModeratelyUntrustworthy vs. Trustworthy: $b = -0.57, CI_{95\%} = [-1.34, 0.21], t(27.5) = -1.5, p = .144$). No contrasts in the right hemisphere were significant (Untrustworthy vs. ModeratelyUntrustworthy: $b = -0.07, CI_{95\%} = [-0.55, 0.41], t(28.9) = -0.31$; Untrustworthy vs. Trustworthy: $b = -0.4, CI_{95\%} = [-1.11, 0.31], t(27.9) = -1.15$; ModeratelyUntrustworthy vs. Trustworthy, $b = -0.33, CI_{95\%} = [-1.01, 0.36], t(27.7) = -0.98; ps > .258$).

Late Time Window

In the late time-window, at posterior and central sites neither the main effect of TrusteeType ($F(2,28) = 1.42, p = .258$), nor the TrusteeType x Laterality ($F(2, 104.7) = 2.58, p = .080$), TrusteeType x Region ($F(2, 1925.3) = 1.3, p = .271$), TrusteeType x Laterality x Region ($F(2, 1925.3) = 0.1, p = .903$) interactions, or any other effects (all $ps > .115$), were statistically significant. ERP traces and scalp topographies are presented in Figures 7a and 7b, respectively.

Frontal LPP. In both the early (all $ps > .618$) and late time-windows (all $ps > .276$), there were no significant effects involving the factors TrusteeType or Laterality.

LPP response to CS2

For the LPP response to CS2 we used the same model as in the LPP response to CS1 analysis.

Posterior LPP

In the early time window, the only significant effects were for Laterality ($F(1,25.4) = 16.86, p < .001$) and Region ($F(1, 31.2) = 25.01, p = 10^{-4}$), which reflected a larger LPP at central sites, (Central vs. Parietal, $b = 1.28, CI_{95\%} = [0.74, 1.82], t(30.4) = 4.85, p < 10^{-15}$), and at right sites (Left vs. Right, $b = -0.66, CI_{95\%} = [-1.01, -0.31], t(22.4) = -3.9, p = .001$). This suggests that while a lateralized posterior LPP emerged, it was not sensitive to second-order conditioning. In the late time window, no effects were significant (all $ps > .074$).

Frontal LPP

Visual examination of the frontal LPP response to CS2 suggested that its temporal distribution was unstable over time. To assess these impressions, we examined its amplitude across six 100 ms long time windows covering the interval from 850 to 1450 ms post-stimulus onset. This yielded evidence for an effect of TrusteeType in the 1150 – 1250 ms time window only ($F(2, 28) = 4.5, p = .020$; all other time-windows $ps > .071$). The LPP was larger in the Untrustworthy as compared to the Trustworthy condition (Untrustworthy vs. Trustworthy: $b = 1.08, CI_{95\%} = [0.29, 1.86], t(26.6) = 2.81, p = .009$). There was also weak evidence for a larger LPP in the ModeratelyUntrustworthy as compared to the Trustworthy condition ($b = 0.9, CI_{95\%} = [-0.01, 1.81], t(27.2) = 2.03, p = .052$), and no evidence for a significant difference between the Untrustworthy and ModeratelyUntrustworthy conditions ($b = 0.18, CI_{95\%} = [-0.73, 1.08], t(27.2) = 0.4, p = .691$). These results suggest that this ERP component indexes trustworthiness-by-association in CS2s. There was no main effect of Laterality ($F(1,29.3) = 1.35, p = .255$) and TrusteeType did not interact with Laterality ($F(2, 70.5) = 0.14, p = .866$). Frontal LPP traces and scalp topographies are presented in Figures 7c and 7d, respectively.

Because the TrusteeType effect on the behavioral measures was a function of contingency awareness, we also investigated whether TrusteeType effects on the LPP were also a function of contingency awareness. We added Contingency Awareness and all higher order interactions to the earlier model, along with all random slopes by participants. Results revealed no evidence that the LPP amplitude was sensitive to Contingency Awareness alone ($F(1,25) = 0.01, p = .930$) or in interaction with TrusteeType ($F(2, 29.9) = 0.55, p = .584$). Results are presented in Figure 7d.

Frontal LPP and Behavior Prediction

We investigated whether the LPP response to CS2 predicted the behavioral response to each CS2 recorded in the one-shot trust game that followed the EEG session. We first tested whether LPP predicted the sum transferred to the CS2 in the trust game using a Linear Mixed Model of the sum

transferred with a fixed effect of the mean LPP response in left and right ROIs, Laterality, and their interactions. Random effects included slope for mean LPP response, Laterality, their interaction, and intercept by subject, and an intercept by CS2. There was no evidence for a predictive effect of LPP amplitude alone ($F(1, 26.5) = 0.41, p = .527$) or in interaction with Laterality ($F(1, 51.2) = 0.10, p = .755$).

Given the behavioral evidence that the sum transferred in the Day 2 game was a function of Contingency Awareness, we tested the possibility that a relationship between LPP and the sum transferred was also function of Contingency Awareness by adding a main effect of Contingency Awareness, and all higher order interactions. Intercept by channel was removed from the model because it was estimated at zero, and prevented model convergence. There was some evidence for an interaction of Contingency Awareness and LPP amplitude in predicting the sum transferred to CS2s ($F(1, 30.7) = 4.63, p = .039$), but no evidence for other effects (all $ps > .414$). Follow-up models fitted separately for items with and without contingency awareness provided weak evidence that when CS2 contingency was remembered, increasing LPP amplitude predicted decreasing sum transferred ($b = -0.226, CI_{95\%} [-0.437, -0.014], t(16.3) = 2.26, p = .038$), but not when contingency was not remembered ($b = 0.050, CI_{95\%} [-0.078, 0.179], t(22.1) = 0.81, p = .428$). Results are presented in figure 7f.

ERP Multivariate Analyses

Whole scalp SVM

Response to CS1

There was evidence that the cross-validated discrimination performance measured by AUC_{cv} differed from chance level (.5) for several condition differences and scalp locations. Differences between Untrustworthy and Trustworthy conditions were captured at central and temporal regions (C1, CP1, CPz, T7) with average channel AUC_{cv} ranging from 0.557 (T7, $CI_{95\%} [0.516, 0.597], p_{\text{perm}} = .011$) to 0.585 (CP1, $CI_{95\%} [0.545, 0.626]; p_{\text{perm}} < .001$). Differences between ModeratelyUntrustworthy and Trustworthy conditions were captured over frontal (F1, Fz, FCz, F2, FC2, F6) and parietal regions (P1, PO3, CP6, P6), AUC ranged from 0.557 (CP6, $CI_{95\%} [0.517, 0.597], p_{\text{perm}} < .003$) to 0.588 (Fz, $CI_{95\%} [0.547, 0.628], p_{\text{perm}} < .001$). No ERP data discriminated between Untrustworthy and ModeratelyUntrustworthy (max AUC at FT8: 0.549, $CI [0.509, 0.589], p_{\text{perm}} = .013$).

ERP predictors of CS1 Learning Rate

Whole scalp classification based on CS1 trial response revealed that electrode channels Oz (AUC = .857, $CI_{95\%} [0.615, 0.97], p_{\text{perm}} < .001$) and PO3 (AUC = .800, $CI_{95\%} [0.586, 0.924], p_{\text{perm}} = .002$) discriminated between fast and slow learners for the Trustworthy condition. No electrodes predicted

the learning rate for Untrustworthy items above chance and also passed a 5% FDR criterion. A maximum AUC of 0.726 was observed at P5 ($CI_{95\%}$ [0.484, 0.896], $p_{\text{perm}} = .019$, q-value = .112).

Response to CS2

Whole scalp SVM.

Cross-validated discrimination performance measured by AUC indicated that differences between the Untrustworthy and Trustworthy conditions were captured at the F1 channel with average channel AUC of 0.569 ($CI_{95\%}$ [0.528, 0.61], $p_{\text{perm}} = .003$). Differences between Untrustworthy and ModeratelyUntrustworthy conditions were captured over left posterior sites (O1, Iz, Oz, PO7, P9, Cz) with AUC ranging from 0.559 (Cz, $CI_{95\%}$ [0.519, 0.6], $p_{\text{perm}} = .013$) to 0.592 (Oz, $CI_{95\%}$ [0.551, 0.632], $p_{\text{perm}} = .001$). Differences between ModeratelyUntrustworthy and Trustworthy conditions were captured over right frontal and parietal sites (Fz, F4, AF4, CP6) and left posterior sites (PO7, O1) with AUC ranging from 0.559 (PO7, $CI_{95\%}$ [0.518, 0.6], $p_{\text{perm}} = .007$) to 0.570 (CP6, $CI_{95\%}$ [0.529, 0.611], $p_{\text{perm}} < .001$).

EEG Multivariate Analyses

Response to CS1

Examination of discrimination performance of CS1 revealed that information discriminating between Untrustworthy and Trustworthy conditions was present in the beta frequency band (14 – 22 Hz) at a left posterior cluster (CP1, P3, P5, PO7), with cluster AUCs ranging from 0.566 ($CI_{95\%}$ [0.526, 0.605]; $p_{\text{perm}} = 0.009$) to 0.601 ($CI_{95\%}$ [0.562, 0.640]; $p_{\text{perm}} < .001$). Information discriminating between the ModeratelyUntrustworthy and Trustworthy conditions was present at analogous right hemisphere electrode sites in two clusters. The first was in the high alpha and low beta bands (11 – 16 Hz; PO4, PO8) with cluster AUC from 0.568 ($CI_{95\%}$ [0.528, 0.607]; $p_{\text{perm}} = .005$) to 0.588 ($CI_{95\%}$ [0.549, 0.627]; $p_{\text{perm}} < .001$). The second was in the beta band (20 – 24 Hz; P1, Pz), with cluster AUC from 0.570 ($CI_{95\%}$ [0.530, 0.610]; $p_{\text{perm}} = .001$) to 0.585 ($CI_{95\%}$ [0.545, 0.625]; $p_{\text{perm}} < .001$). No clusters spanning more than 3 Hz and 2 channels discriminated between the Untrustworthy and ModeratelyUntrustworthy conditions (max AUC = 0.585, $CI_{95\%}$ [0.528, 0.607]).

EEG predictors of CS1 Learning Rate

Whole scalp-classification on CS1 did not reveal any channel or frequency that could reliably discriminate between fast and slow learners. Permutation tests were performed based on 1000 replicates. Although several statistical tests returned above-chance results that passed a 5% FDR criterion, no clusters spanned at least 3 frequency bins and 2 neighboring channels, suggesting that the significant tests likely emerged from noise.

Response to CS2

Classification analysis on CS2 did not reveal any information in the power data that could discriminate conditions above chance. Although several tests returned positive results that passed the 5% FDR criterion for comparisons of Untrustworthy versus Trustworthy, and Moderately Untrustworthy versus Trustworthy, no clusters spanned at least 3 frequency bins and 2 neighboring channels, suggesting that the significant tests likely emerged from noise.

Prospective Power Analysis

Behavior

We computed prospective statistical power for the primary result on Day 2: The difference in sum transferred to Trustworthy and Untrustworthy CS2 images for which contingency was correctly remembered. We based the power computation on this effect because it was the smallest behavioral effect of interest. Prospective power was computed for the observed effect size ($b = -3.26$), and for the effect size corresponding to the lower bound of the 95% CI (i.e., $b = -1.03$). Power was computed for tests at $\alpha = .05$. The lower bound effect size was considered a theoretically meaningful effect to examine because it reflected about a \$1 difference in sum transferred. Simulation results indicated that for the original effect size a sample size of 36 or greater would achieve power of 80%. For the lower bound, power was weak at around 65 %. The power curve quickly reached a plateau. This is likely a consequence of effects being modeled with both Participants and Items as random effects, and therefore further power could only be achieved by increasing both Participant sample size and Item sample size. Increasing the latter was not feasible because it would have affected the structure and duration of the experiment.

ERPs

For ERPs, we determined the sample size required to have 80% power at $\alpha = .05$ for the test of LPP prediction of sum transferred on Day 2, when contingency was remembered. We computed power for the estimated effect ($b = -0.226$, $CI_{95\%} [-0.437, -0.014]$), as well as a lower bound effect. For the lower bound effect we did not use the lower limit of the CI, because it was too close to zero to be of theoretical interest. Over the average within-participant range of LPP amplitudes measured in the Exploratory Data, this represented a change in sum transferred of only \$0.08 between conditions (2.7% of the average sum transferred). Instead of this value, we used an effect size that was about half the observed effect, -0.1, and that led to a change between conditions of \$0.57 (corresponding to a change of 19.4% of mean sum transferred). For the original effect size, 80% power at $\alpha = .05$ could be achieved with 48 participants. For the lower bound effect size, power was extremely low (e.g., at $N = 54$, power = 30.06%, $CI_{95\%} [29.16, 30.97]$), it is therefore unlikely that a true effect would be

detected if it were in this range.

Confirmation Experiment

Methods

Unless otherwise specified, the methods were identical to those of the exploratory study.

Participants

The target sample size for the Confirmation Experiment was 48 participants. We recorded data from 55 participants. Participants were excluded whenever they had missing data from one of the two sessions. Five participants were excluded because of technical issues during data acquisition, and 1 because he did not complete the behavioral post-tests. At the data processing stage, we excluded 1 participant whose behavior indicated she did not comply with trust game instructions, and 2 participants because more than 50% of the EEG trials were rejected in any one condition. When a participant was rejected replacement data were recorded. However, because of time constraints two participants could not be replaced. The final analysis included data from 46 participants (12 males). Participant mean age was 22 years ($SD = 1$), all but one were right-handed.

EEG and ERP preprocessing

Preprocessing steps and settings were identical to those used for the Exploratory Dataset. There were no findings of EEG power response to CS2s in the Exploratory Dataset, so the corresponding Confirmatory data were not examined. For ERPs, the average number of trials per condition was 61 ($SD = 9$; range = 42 – 80) for Day 1, and 60 ($SD = 8$; range = 40 – 79) for Day 2. For EEG, the average number of trials per condition was 61 ($SD = 9$; range = 42 – 80) for Day 1.

SCR

No participants were excluded from the analysis.

Multivariate Classification

We used the Confirmatory Data to evaluate the classification performance of the SVM classification models trained on the Exploratory Data. Discrimination accuracy was determined by computing the AUC in the same manner as for the Exploratory Data, bootstrap confidence intervals (10 000 replicates) were used in place of parametric confidence intervals because they were less expensive to compute in this setting. In addition, we computed the probability of observing each AUC value by chance. Models were trained repeatedly on the Exploratory Dataset with shuffled labels, and their

performance on the Confirmatory Dataset recorded to build a distribution of AUC when no class information was available. This provides a more stringent indicator of performance than the CI, as it is based on the likelihood of observing a given AUC value assuming no information is available for discrimination, and performance is due to chance.

Results

Behavior

Sum transferred to CS1

On Day 1, there was a significant effect of TrusteeType on the mean sum transferred ($F(2,45) = 402.46, p < .001$). Participants transferred more money to Trustees in the Trustworthy condition than those in the Untrustworthy and ModeratelyUntrustworthy conditions, and more to those in the ModeratelyUntrustworthy condition than the Untrustworthy condition (Untrustworthy vs. ModeratelyUntrustworthy: $b = -3.69, CI_{95\%} = [-4.56, -2.83], t(48.6) = -8.57$; Untrustworthy vs. Trustworthy: $b = -7.04, CI_{95\%} = [-7.55, -6.54], t(45) = -28.08$; ModeratelyUntrustworthy vs. Trustworthy: $b = -3.35, CI_{95\%} = [-4.14, -2.56], t(45.8) = -8.53$; all $ps < 10^{-15}$).

The effect of TrusteeType carried on to the Day 2 trust game, (Trust $F(2,44.1) = 174.17, p < 10^{-15}$). Participants transferred more money to Trustees in the Trustworthy condition than those in the Untrustworthy and ModeratelyUntrustworthy conditions, and more to those in the ModeratelyUntrustworthy condition than the Untrustworthy condition (Untrustworthy vs. ModeratelyUntrustworthy: $b = -3.45, CI_{95\%} = [-2.29, -4.6], t(44.4) = -6.03$; Untrustworthy vs. Trustworthy: $b = 8, CI_{95\%} = [-8.88, -7.12], t(42.5) = 18.36$; ModeratelyUntrustworthy vs. Trustworthy: $b = -4.55, CI_{95\%} = [-5.7, -3.41], t(43.8) = -8.02$; all $ps < .001$).

CS1 Subjective Ratings

Perceived Trustworthiness. The main effect of TrusteeType was significant ($F(2,42.6) = 176.19, p < 10^{-15}$), but the main effect of Day ($F(1,26.5) = 0.03, p = .857$), and the TrusteeType x Day interaction ($F(2,25.5) = 0.06, p = .943$), were not. Unsurprisingly, Trustees in the Untrustworthy condition were perceived as less trustworthy than those in the Trustworthy and ModeratelyUntrustworthy conditions, moreover, those in the ModeratelyUntrustworthy condition were perceived as less trustworthy than those in the Trustworthy condition (Untrustworthy vs. ModeratelyUntrustworthy: $b = -2.74, CI_{95\%} = [-3.38, -2.09], t(37.8) = -8.62$; Untrustworthy vs. Trustworthy: $b = -5.33, CI_{95\%} = [-5.91, -4.75], t(43.6) = -18.57$; ModeratelyUntrustworthy vs. Trustworthy: $b = -2.59, CI_{95\%} = [-3.22, -1.96], t(40.7) = -8.3$, all $ps < 10^{-15}$).

Attitudes. The effect of TrusteeType was significant ($F(2,34.8) = 146.67, p < 10^{-15}$), as was the TrusteeType x Day interaction ($F(2,20.5) = 9.57, p = .001$), but the main effect of Day ($F(1,16.2) = 2.63, p = .124$) was not. On both days, participants had less positive attitudes toward Trustees in the Untrustworthy condition as compared to the ModeratelyUntrustworthy and Trustworthy conditions, and were less positive toward those in the ModeratelyUntrustworthy than the Trustworthy condition. Moreover, the effects were greater on Day 1 than Day 2 (Day 1: Untrustworthy vs. ModeratelyUntrustworthy: $b = -2.48, CI_{95\%} = [-2.99, -1.98], t(39.4) = -9.92$; Untrustworthy vs. Trustworthy: $b = -4.86, CI_{95\%} = [-5.44, -4.27], t(32.6) = -16.88$; ModeratelyUntrustworthy vs. Trustworthy: $b = -2.37, CI_{95\%} = [-3.01, -1.74], t(25.6) = -7.7$); Day 2: Untrustworthy vs. ModeratelyUntrustworthy: $b = -1.74, CI_{95\%} = [-2.34, -1.14], t(19.8) = -6.06$; Untrustworthy vs. Trustworthy: $b = -3.75, CI_{95\%} = [-4.33, -3.17], t(41.7) = -12.98$; ModeratelyUntrustworthy vs. Trustworthy: $b = -2.01, CI_{95\%} = [-2.58, -1.43], t(23.5) = -7.16$, all $ps < .001$)).

Warmth. The main effect of TrusteeType was significant ($F(2,33)=84.53, p<10^{-12}$), as was the TrusteeType x Day interaction ($F(2,24.4)=3.68, p=.040$), but the main effect of Day was not ($F(1,16.2) = 0.07, p = .798$). On both days, participants viewed Trustees in the Untrustworthy condition as less warm than Trustees in the ModeratelyUntrustworthy and Trustworthy conditions, and those in the ModeratelyUntrustworthy condition as less warm than those in the Trustworthy condition. Moreover, the effects were greater on Day 1 than Day 2 (Day 1: Untrustworthy vs. ModeratelyUntrustworthy: $b = -2, CI_{95\%} = [-2.51, -1.49], t(38.8) = -7.95$; Untrustworthy vs. Trustworthy: $b = -3.87, CI_{95\%} = [-4.51, -3.23], t(35.7) = -12.23$; ModeratelyUntrustworthy vs. Trustworthy: $b = -1.86, CI_{95\%} = [-2.59, -1.13], t(27.1) = -5.25$; Day 2: Untrustworthy vs. ModeratelyUntrustworthy: $b = -1.42, CI_{95\%} = [-2.01, -0.82], t(20) = -4.95$; Untrustworthy vs. Trustworthy: $b = -3.08, CI_{95\%} = [-3.74, -2.42], t(40.2) = -9.45$; ModeratelyUntrustworthy vs. Trustworthy: $b = -1.67, CI_{95\%} = [-2.3, -1.03], t(24.8) = -5.39$, all $ps < .001$).

Attractiveness. The main Effect of TrusteeType was significant ($F(2,20.5) = 33.47, p < 10^{-6}$), as was the main effect of Day ($F(1,30.5) = 16.35, p < .001$), with lower ratings on Day 1 ($b = -0.41, CI_{95\%} = [-0.62, -0.2], t(27) = -3.98, p < .001$). The TrusteeType x Day interaction was not significant ($F(2,33.8) = 0.2, p = .819$). Participants viewed Trustees in the Untrustworthy condition as less attractive than Trustees in the ModeratelyUntrustworthy and Trustworthy conditions, and those in the ModeratelyUntrustworthy condition as less attractive than those in the Trustworthy condition (Untrustworthy vs. ModeratelyUntrustworthy: $b = -0.9, CI_{95\%} = [-1.32, -0.48], t(16.5) = -4.52$; Untrustworthy vs. Trustworthy: $b = -2.12, CI_{95\%} = [-2.66, -1.58], t(33.9) = -7.93$; ModeratelyUntrustworthy vs. Trustworthy: $b = -1.22, CI_{95\%} = [-1.78, -0.66], t(16.7) = -4.61$, all $ps < .001$).

Competence. The main effect of TrusteeType was significant ($F(2,17.8) = 22.52, p = 10^{-5}$), but effect of Day ($F(1,25.9) = 2.07, p = .162$) and the TrusteeType x Day interaction ($F(2,11.5) = 0.14, p = .874$), were not. Participants viewed Trustees in the Untrustworthy condition as less competent than Trustees in the ModeratelyUntrustworthy and Trustworthy conditions, and those in the ModeratelyUntrustworthy condition as less competent than those in the Trustworthy condition (Untrustworthy vs. ModeratelyUntrustworthy: $b = -0.66, CI_{95\%} = [-0.16, -1.16], t(12.5) = -2.86, p = .014$; Untrustworthy vs. Trustworthy: $b = -1.93, CI_{95\%} = [-2.56, -1.31], t(27.5) = -6.35, p < 10^{-16}$; ModeratelyUntrustworthy vs. Trustworthy: $b = -1.27, CI_{95\%} = [-1.75, -0.8], t(21.1) = -5.57, p < 10^{-16}$).

Ambivalence. The main effect of TrusteeType was significant ($F(2,21.4) = 19.71, p = 10^{-5}$), but the main effect of Day ($F(1,38.7) = 1.73, p = .196$) and the TrusteeType x Day interaction ($F(2,8.5) = 0.35, p = .715$), were not. Trustees in the Untrustworthy and ModeratelyUntrustworthy conditions were viewed more ambivalently than those in the Trustworthy condition (Untrustworthy vs. Trustworthy: $b = 0.84, CI_{95\%} = [0.42, 1.27], t(24.4) = 4.12, p < .001$; ModeratelyUntrustworthy vs. Trustworthy: $b = 1.3, CI_{95\%} = [0.84, 1.77], t(24.1) = 5.83, p < 10^{-15}$), but not from each other (Untrustworthy vs. ModeratelyUntrustworthy: $b = -0.46, CI_{95\%} = [-0.97, 0.05], t(14) = -1.93, p = .074$).

Dominance. There main effect of TrusteeType was significant ($F(2,27.5) = 7.49, p = .003$), but the main effect of Day ($F(1,38.7) = 2.06, p = .159$) and the TrusteeType x Day interaction ($F(2,16.8) = 0.73, p = .495$) were not. Dominance ratings were higher for Trustees in the Untrustworthy condition as compared to the Trustworthy and ModeratelyUntrustworthy conditions (Untrustworthy vs. Trustworthy: $b = 0.82, CI_{95\%} = [0.12, 1.52], t(38.1) = 2.38, p = .022$; Untrustworthy vs. ModeratelyUntrustworthy: $b = 1.05, CI_{95\%} = [0.47, 1.63], t(20) = 3.77, p = .001$). There was no significant difference between the Trustworthy and ModeratelyUntrustworthy conditions (ModeratelyUntrustworthy vs. Trustworthy: $b = -0.23, CI_{95\%} = [-0.76, 0.31], t(35.5) = -0.85, p = .398$).

Arousal. There were significant main effects of TrusteeType ($F(2,16.6) = 11.49, p = .001$) and Day ($F(1,31.6) = 6.23, p = .018$), with higher arousal ratings on Day 1 ($b = 0.38, CI_{95\%} = [0.06, 0.7], t(29.4) = 2.45, p = .02$), but the TrusteeType x Day interaction was not significant ($F(2,15.1) = 0.48, p = .626$). Arousal ratings were higher for Trustees in the Untrustworthy and Trustworthy conditions as compared to the ModeratelyUntrustworthy condition (Untrustworthy vs. ModeratelyUntrustworthy, $b = 1.11, CI_{95\%} = [0.54, 1.68], t(12.6) = 4.23, p = .001$; Trustworthy vs. ModeratelyUntrustworthy: $b = 0.85, CI_{95\%} = [1.48, 0.22], t(21.3) = -2.8, p = .011$), with no significant difference in the ratings between the Trustworthy and Untrustworthy conditions (Untrustworthy vs. Trustworthy: $b = 0.26, CI_{95\%} = [-0.43, 0.95], t(39.4) = 0.76, p = .454$).

Valence. The main effect of TrusteeType was significant ($F(2,39) = 122.99, p < 10^{-15}$), as was the TrusteeType x Day interaction ($F(2,24) = 11.4, p < .001$), but the main effect of Day ($F(1,22) = 0.01, p = .931$) was not. On both days, participants rated the valence of Trustees in the Untrustworthy condition as lower than Trustees in the ModeratelyUntrustworthy and Trustworthy conditions, and those in the ModeratelyUntrustworthy condition as lower than those in the Trustworthy condition. Moreover, the effects were greater on Day 1 than Day 2 (Day 1: Untrustworthy vs. ModeratelyUntrustworthy: $b = -2.46, CI_{95\%} = [-3.12, -1.8], t(38.4) = -7.52$; Untrustworthy vs. Trustworthy: $b = -5.06, CI_{95\%} = [-5.73, -4.39], t(40.8) = -15.28$; ModeratelyUntrustworthy vs. Trustworthy: $b = -2.6, CI_{95\%} = [-3.26, -1.94], t(30.5) = -8.01$); Day 2: Untrustworthy vs. ModeratelyUntrustworthy: $b = -2, CI_{95\%} = [-2.66, -1.35], t(24.7) = -6.3$; Untrustworthy vs. Trustworthy: $b = -3.71, CI_{95\%} = [-4.33, -3.1], t(36.2) = -12.18$; ModeratelyUntrustworthy vs. Trustworthy: $b = -1.71, CI_{95\%} = [-2.25, -1.17], t(32.1) = -6.41$, all $ps < 10^{-15}$).

CS2 - CS1 Contingency Awareness

In the test of CS1-CS2 contingency awareness, participants had an average accuracy of 43.1%, $CI_{95\%} [33.0, 53.3]$, which was significantly different from chance ($t(45) = 8.56, p < 10^{-10}$). Results are presented Figure 11. We investigated whether explicit awareness of CS1 - CS2 contingency changed as a function of TrusteeType. A likelihood ratio test against the intercept-only model was not significant ($\chi^2(2) = 0.39, p = .98$).

CS2 Behavioral Responses

Sum transferred in Day 2 trust game. There was a direct effect of CS1 TrusteeType on the sum transferred to CS2s ($F(2,18.4) = 6.98, p = .006$), but no significant main effect of Contingency Awareness ($F(2,14.6) = 2.2, p = .147$) or, unlike in the Exploratory Dataset, no significant interaction between TrusteeType and Contingency Awareness ($F(2,14.6) = 2.2, p = .147$). However, following the Exploratory Dataset findings we performed tests on parameters by contingency. In absence of Contingency Awareness, there was no significant effect of TrusteeType on sum transferred (Untrustworthy vs. ModeratelyUntrustworthy: $b = -0.4, CI_{95\%} = [-1.36, 0.57], t(11.6) = 0.9, p = .389$; Untrustworthy vs. Trustworthy: $b = -0.44, CI_{95\%} = [-1.45, 0.57], t(16.1) = -0.92, p = .373$; ModeratelyUntrustworthy vs. Trustworthy: $b = -0.04, CI_{95\%} = [-0.97, 0.89], t(46.9) = -0.09, p = .930$). For Trustees for which contingency was remembered, there was a significantly larger sum transferred for the ModeratelyUntrustworthy condition as compared to the Untrustworthy condition (Untrustworthy vs. ModeratelyUntrustworthy: $b = -1.64, CI_{95\%} = [-2.68, -0.61], t(20.1) = 3.31, p = .004$). The difference between Trustworthy and Untrustworthy approached significance (Untrustworthy vs. Trustworthy: $b = -1.48, CI_{95\%} = [-3.03, 0.07], t(11.6) = -2.09, p = .06$), but there was no significant difference between ModeratelyUntrustworthy and Trustworthy conditions

(ModeratelyUntrustworthy vs. Trustworthy: $b = 0.16$, $CI_{95\%} = [-1.72, 2.05]$, $t(10.1) = 0.19$, $p = .851$).

CS2 Subjective Ratings

Trustworthiness. There was no main effect of CS1 TrusteeType ($F(2,41) = 2.05$, $p = .142$) or Contingency Awareness ($F(1,141.6) = 0.37$, $p = .543$) on CS2 perceived Trustworthiness, but the TrusteeType x Contingency Awareness interaction was significant ($F(2,168.6) = 4.42$, $p = .014$). CS1 TrusteeType did not affect ratings in the absence of contingency awareness (Untrustworthy vs. ModeratelyUntrustworthy: $b = 0.29$, $CI_{95\%} = [-0.33, 0.9]$, $t(61.5) = 0.92$; Untrustworthy vs. Trustworthy: $b = 0.11$, $CI_{95\%} = [-0.54, 0.77]$, $t(65.7) = 0.35$; ModeratelyUntrustworthy vs. Trustworthy: $b = -0.17$, $CI_{95\%} = [-0.85, 0.51]$, $t(69.4) = -0.5$; all $ps > .359$). However, for items for which contingency was remembered, trustworthiness ratings were higher for Trustees in the Trustworthy as compared to the Untrustworthy conditions (Untrustworthy vs. Trustworthy: $b = -1.15$, $CI_{95\%} = [-1.89, -0.41]$, $t(80.5) = -3.1$, $p = .003$), other differences were not significant (Untrustworthy vs. ModeratelyUntrustworthy: $b = -0.71$, $CI_{95\%} = [-1.41, 0]$, $t(81) = -1.98$, $p = .051$; ModeratelyUntrustworthy vs. Trustworthy: $b = -0.44$, $CI_{95\%} = [-1.21, 0.32]$, $t(83.7) = -1.15$, $p = .253$).

Attitudes. The main effect of CS1 TrusteeType ($F(2,19) = 4.05$, $p = .034$), and the CS1 TrusteeType x Contingency Awareness interaction ($F(2,16.2) = 4.39$, $p = .03$) were significant, but the main effect of Contingency Awareness was not ($F(1,23.6) = 0.09$, $p = .764$). CS1 TrusteeType did not affect ratings in absence of contingency awareness (Untrustworthy vs. ModeratelyUntrustworthy: $b = 0.12$, $CI_{95\%} = [-0.51, 0.76]$, $t(10) = 0.43$, $p = .675$; Untrustworthy vs. Trustworthy: $b = -0.03$, $CI_{95\%} = [-0.62, 0.56]$, $t(14.3) = -0.11$, $p = .917$; ModeratelyUntrustworthy vs. Trustworthy: $b = -0.15$, $CI_{95\%} = [-0.77, 0.47]$, $t(21.1) = -0.51$, $p = .617$). However, for items for which contingency was remembered, attitude ratings were significantly increased for the ModeratelyUntrustworthy and Trustworthy conditions as compared to the Untrustworthy condition (Untrustworthy vs. ModeratelyUntrustworthy: $b = -0.62$, $CI_{95\%} = [-1.2, -0.04]$, $t(21.7) = -2.22$, $p = .037$; Untrustworthy vs. Trustworthy: $b = -1.14$, $CI_{95\%} = [-1.84, -0.44]$, $t(13) = -3.51$, $p = .004$). There was no significant difference between the ModeratelyUntrustworthy and Trustworthy conditions (ModeratelyUntrustworthy vs. Trustworthy: $b = -0.52$, $CI_{95\%} = [-1.29, 0.26]$, $t(16) = -1.42$, $p = .176$).

Warmth. The main effects of CS1 TrusteeType ($F(2,20.8) = 1.93$, $p = .170$) and Contingency Awareness ($F(1,25.1) = 0.04$, $p = .851$) were not significant, but the CS1 TrusteeType x Contingency Awareness interaction was significant ($F(2,22.7) = 5.61$, $p = .01$). CS1 TrusteeType did not affect ratings in absence of contingency awareness (Untrustworthy vs. ModeratelyUntrustworthy: $b = 0.41$, $CI_{95\%} = [-0.13, 0.96]$, $t(44.4) = 1.53$, $p = .132$; Untrustworthy vs. Trustworthy: $b = 0.15$, $CI_{95\%} = [-0.43, 0.73]$, $t(33.5) = 0.51$, $p = .612$; ModeratelyUntrustworthy vs. Trustworthy: $b = -0.27$, $CI_{95\%} = [-0.86, 0.33]$, $t(57.3) = -0.9$, $p = .372$). For items for which contingency was remembered, ratings were higher

for the Trustworthy and ModeratelyUntrustworthy conditions as compared to the Untrustworthy condition (Untrustworthy vs. Trustworthy: $b = -1.2$, $CI_{95\%} = [-2.28, -0.13]$, $t(12.3) = -2.43$, $p = .031$; Untrustworthy vs. ModeratelyUntrustworthy: $b = -0.9$, $CI_{95\%} = [-1.59, -0.2]$, $t(22.7) = -2.66$, $p = .014$). There was no significant difference between the Trustworthy and ModeratelyUntrustworthy conditions (ModeratelyUntrustworthy vs. Trustworthy: $b = -0.31$, $CI_{95\%} = [-1.26, 0.64]$, $t(14.1) = -0.7$, $p = .498$).

Attractiveness. The main effect of CS1 TrusteeType ($F(2,21.4) = 4.36$, $p = .026$) and the CS1 TrusteeType x Contingency Awareness interaction ($F(2,42.8) = 5.21$, $p = .009$) were significant, but the main effect of Contingency Awareness ($F(1,20.3) = 0$, $p = .945$) was not. CS1 TrusteeType did not affect ratings in the absence of contingency awareness (Untrustworthy vs. ModeratelyUntrustworthy: $b = 0.16$, $CI_{95\%} = [-0.44, 0.76]$, $t(19.3) = 0.56$, $p = .580$; Untrustworthy vs. Trustworthy: $b = -0.06$, $CI_{95\%} = [-0.62, 0.51]$, $t(24.1) = -0.21$, $p = .836$; ModeratelyUntrustworthy vs. Trustworthy: $b = -0.22$, $CI_{95\%} = [-0.77, 0.34]$, $t(51.7) = -0.79$, $p = .435$). For items for which contingency was remembered, attractiveness ratings were higher for the Trustworthy condition as compared to the ModeratelyUntrustworthy and Untrustworthy conditions (Untrustworthy vs. ModeratelyUntrustworthy: $b = -0.83$, $CI_{95\%} = [-1.55, -0.12]$, $t(26.8) = -2.39$, $p = .024$; Untrustworthy vs. Trustworthy: $b = -1.3$, $CI_{95\%} = [-2.05, -0.55]$, $t(18.5) = -3.63$, $p = .002$). There was no significant difference between the ModeratelyUntrustworthy and Untrustworthy conditions (ModeratelyUntrustworthy vs. Untrustworthy: $b = -0.47$, $CI_{95\%} = [-1.08, 0.15]$, $t(43) = -1.54$, $p = .132$).

Ambivalence. The main effects of CS1 TrusteeType ($F(2,19.6) = 2.29$, $p = .128$) and Contingency Awareness ($F(1,30.9) = 2.97$, $p = .095$), as well as the CS1 TrusteeType x Contingency Awareness interaction ($F(2,14.2) = 1.18$, $p = .335$) were not significant.

Dominance. The main effects of CS1 TrusteeType ($F(2,30.4) = 0.08$, $p = .920$) and Contingency Awareness ($F(1,24.6) = 0.32$, $p = .575$), as well as the CS1 TrusteeType x Contingency Awareness interaction ($F(2,151.8) = 0.08$, $p = .920$) were not significant.

Competence. The main effects of CS1 TrusteeType ($F(2,24.1) = 0.85$, $p = .439$) and Contingency Awareness ($F(1,27.3) = 0.39$, $p = .537$) were not significant, but the CS1 TrusteeType x Contingency Awareness interaction was ($F(2,21.6) = 4.54$, $p = .023$). CS1 TrusteeType did not affect ratings in the absence of contingency awareness (Untrustworthy vs. ModeratelyUntrustworthy: $b = 0.51$, $CI_{95\%} = [-0.08, 1.1]$, $t(14.9) = 1.85$, $p = .085$; Untrustworthy vs. Trustworthy: $b = 0.21$, $CI_{95\%} = [-0.36, 0.78]$, $t(48.9) = 0.74$, $p = .464$; ModeratelyUntrustworthy vs. Trustworthy: $b = -0.3$, $CI_{95\%} = [-0.96, 0.36]$, $t(23.9) = -0.94$, $p = .355$). For items for which contingency was remembered, ratings were higher in the Trustworthy condition as compared to the Untrustworthy condition (Untrustworthy vs.

Trustworthy: $b = -0.87$, $CI_{95\%} = [-1.7, -0.03]$, $t(14.2) = -2.21$, $p = .044$). Other pairwise differences were not significant (Untrustworthy vs. ModeratelyUntrustworthy: $b = -0.5$, $CI_{95\%} = [-1.1, 0.1]$, $t(22.8) = -1.73$, $p = .098$; ModeratelyUntrustworthy vs. Trustworthy: $b = -0.36$, $CI_{95\%} = [-1.03, 0.31]$, $t(30.4) = -1.11$, $p = .276$).

Valence. The main effect of CS1 TrusteeType ($F(2,22.6) = 4.92$, $p = .017$), and the CS1 TrusteeType x Contingency Awareness interaction ($F(2,16.3) = 7.32$, $p = .005$) were significant, but the main effect of Contingency Awareness ($F(1,23.1) = 0.13$, $p = .720$) was not. CS1 TrusteeType did not affect ratings in the absence of contingency awareness (Untrustworthy vs. ModeratelyUntrustworthy: $b = 0.15$, $CI_{95\%} = [-0.57, 0.88]$, $t(15.4) = 0.45$, $p = .659$; Untrustworthy vs. Trustworthy: $b = 0.14$, $CI_{95\%} = [-0.49, 0.76]$, $t(22) = 0.45$, $p = .659$; ModeratelyUntrustworthy vs. Trustworthy: $b = -0.02$, $CI_{95\%} = [-0.75, 0.71]$, $t(13.3) = -0.05$, $p = .960$). For items for which contingency was remembered, valence ratings were lower in the Untrustworthy compared to the ModeratelyUntrustworthy and Trustworthy conditions, and lower in the ModeratelyUntrustworthy than the Trustworthy condition (Untrustworthy vs. ModeratelyUntrustworthy: $b = -0.77$, $CI_{95\%} = [-1.52, -0.01]$, $t(17.3) = -2.15$, $p = .046$; Untrustworthy vs. Trustworthy: $b = -1.62$, $CI_{95\%} = [-2.48, -0.76]$, $t(15.4) = -4.01$, $p = .001$; ModeratelyUntrustworthy vs. Trustworthy: $b = -0.85$, $CI_{95\%} = [-1.62, -0.09]$, $t(16) = -2.36$, $p = .031$).

Arousal. The main effects of CS1 TrusteeType ($F(2,19.2) = 0.7$, $p = .510$) and Contingency Awareness ($F(1,15.8) = 0.91$, $p = .354$), as well as the Trustworthiness x Contingency Awareness interaction ($F(2,25.1) = 1.06$, $p = .362$), were not significant.

Skin Conductance Response

There was no significant effect of TrusteeType on SCR responses to CS1 on Day 1, $F(2, 1334.1) = 0.49$, $p = .614$, or to CS2 on Day 2, $F(2,4847) = 0.98$, $p = .374$).

ERP Univariate Analysis

LPP response to CS1

Posterior LPP.

In the early time-window, the TrusteeType x Laterality interaction was not significant ($F(2,78.3) = 1.12$, $p = .333$), which stands in contrast to the effect obtained for the Exploratory Data. For both the left and right hemisphere there were no significant differences between the Untrustworthy and Trustworthy conditions (Untrustworthy vs. Trustworthy left: $b = -0.37$, $CI_{95\%} = [-1.07, 0.34]$, $t(44.1) = -1.05$, $p = .3$; Untrustworthy vs. Trustworthy right: $b = -0.14$, $CI_{95\%} = [-0.84, 0.55]$, $t(44.1) = -0.42$, $p = .675$), although the effects were of the same sign as the Exploratory Data effects. However, there was a significant main effect of TrusteeType which did not emerge in the Exploratory Data ($F(2,44.8)$

= 4.92, $p = .012$). In both hemispheres, the LPP was smaller in the ModeratelyUntrustworthy than the Trustworthy condition (Left: ModeratelyUntrustworthy vs. Trustworthy: $b = -0.88$, $CI_{95\%} = [-1.5, -0.26]$, $t(44.1) = -2.85$, $p = .007$; Right: ModeratelyUntrustworthy vs. Trustworthy: $b = -0.72$, $CI_{95\%} = [-1.32, -0.13]$, $t(44.1) = -2.46$, $p = .018$), which was not significant in the Exploratory Data. In the right hemisphere only, the LPP was larger for the ModeratelyUntrustworthy than the Untrustworthy condition (Right: Untrustworthy vs. ModeratelyUntrustworthy: $b = 0.58$, $CI_{95\%} = [0.05, 1.11]$, $t(43.5) = 2.19$, $p = .034$; Left: Untrustworthy vs. ModeratelyUntrustworthy: $b = 0.51$, $CI_{95\%} = [-0.04, 1.07]$, $t(44.3) = 1.86$, $p = .069$). The only other significant effect in the overall model was an effect of Region ($F(1,48.7) = 77.76$, $p < 10^{-10}$; all other effects $ps > .078$) which reflected higher LPP at central sites as compared to parietal sites (Central vs. Parietal: $b = 1.96$, $CI_{95\%} = [1.5, 2.42]$, $t(45.7) = 8.59$, $p < 10^{-15}$) as suggested in the Exploratory Data. However, the effect of Laterality was not significant ($F(1,44.9) = 0.38$, $p = .541$).

Results in the late time-window were consistent with the Exploratory Dataset results in that there was no significant effect of TrusteeType ($F(2,44.4) = 0.8$, $p = .455$), no TrusteeType x Laterality interaction ($F(2,276.3) = 0.09$, $p = .913$), or TrusteeType x Region interaction ($F(2,3066) = 0.64$, $p = .528$) or TrusteeType x Laterality x Region interaction ($F(2,3066) = 0.25$, $p = .781$). The effect of Region was significant ($F(1,41.8) = 8.29$, $p = .006$), but there were no other significant effects (all $ps > .078$).

Frontal LPP

Consistent with the Exploratory Dataset, at frontal sites, there was no significant effect of TrusteeType ($F(2,44.8) = 2.88$, $p = .067$) nor a TrusteeType x Laterality interaction ($F(2,228.3) = 0.11$, $p = .9$). In the late time window, there were no significant effects (all $ps > .615$).

ERP response to CS2

Posterior LPP.

In the early time window, there were no significant effects (all $ps > .153$) other than effects of Region ($F(1,45.1) = 108.77$, $p = 10^{-12}$), which reflected higher LPP at central sites, (Central vs. Parietal, $b = 1.88$, $CI_{95\%} = [1.51, 2.26]$, $t(43.8) = 10.19$, $p < 10^{-15}$). This confirmed the Exploratory Dataset results indicating that a LPP emerged, but was not sensitive to second-order conditioning. Unlike in the Exploratory Dataset, there was no significant main effect of Laterality, ($F(1,38.6) = 0.05$, $p = .820$), there were no LLP amplitude differences between hemispheres (Left vs. Right, $b = -0.03$, $CI_{95\%} = [-0.31, 0.25]$, $t(35.3) = -0.22$, $p = .826$). In the late time window, the original model failed to converge, and was fitted without a random slope by participant for the higher order interaction. There was a significant effect of Region ($F(1,36.5) = 34.09$, $p = 10^{-6}$) which was absent in the Exploratory Dataset. Unplanned comparisons indicated that the LPP was enhanced at posterior as opposed to central sites (b

= 0.91, $CI_{95\%} = [0.58, 1.24]$, $t(30.4) = 5.56$, $p < 10^{-16}$). No other effects were significant (all $ps > .167$).

Frontal LPP.

The main effect TrusteeType in the 1150 – 1250 ms time window observed in the Exploratory Dataset was not significant ($F(2,44.7) = 2.87$, $p = .067$), and there was no significant difference in LPP between Untrustworthy and Trustworthy conditions ($b = 0.27$, $CI_{95\%} = [-0.41, 0.96]$, $t(43.3) = 0.8$, $p = .429$). Further, there was no enhancement of the LPP in the ModeratelyUntrustworthy as compared to the Trustworthy condition (ModeratelyUntrustworthy vs. Trustworthy: $b = -0.42$, $CI_{95\%} = [-1.08, 0.24]$, $t(44) = -1.27$, $p = .211$). In contrast to the Exploratory Dataset findings, LPP was smaller for the ModeratelyUntrustworthy than the Untrustworthy condition (Untrustworthy vs. ModeratelyUntrustworthy: $b = 0.69$, $CI_{95\%} = [0.1, 1.28]$, $t(43.7) = 2.34$, $p = .024$). No other effects were significant, TrusteeType did not interact with Laterality ($F(2,331.3) = 0.2$, $p = .820$) and there was no main effect of Laterality ($F(1,5.4) = 0$, $p = .952$).

Frontal LPP and Behavior

As for the Exploratory Dataset, there was no significant effect of LPP amplitude ($F(1,41.8) = 0.59$, $p = .446$) on sum transferred or a LPP Amplitude x Laterality ($F(1,449.3) = 0.05$, $p = .832$) interaction. A model testing whether the frontal LPP predicted the sum transferred yielded negative results. Model estimation failed to converge, and a model was fit without the random slope by participant for the higher-order interaction. There was no significant Contingency Awareness x LPP amplitude interaction ($F(1,32.1) = 0.21$, $p = .649$), and no other effects were significant (all $ps > .184$). Models fitted separately for remembered and non-remembered contingency indicated that for both, increasing LPP amplitudes did not significantly predict the sum transferred (Remembered: $b = 0.015$, $CI_{95\%} [-0.158, 0.187]$, $t(21.7) = 0.17$, $p = .864$; Not remembered: $b = 0.083$, $CI_{95\%} [-0.028, 0.193]$, $t(34.7) = 1.52$, $p = 0.137$). Results are presented Figure 16c. In agreement with the Exploratory Dataset, there was no evidence that the LPP was responsive to contingency awareness. There was no significant main effect of Contingency Awareness on the LPP ($F(1,31.9) = 0.16$, $p = .688$) or a significant Contingency Awareness x Trust ($F(2,43.2) = 2$, $p = .147$) interaction.

ERP Multivariate

Response to CS1

The SVM models trained on the Exploratory Dataset were used to classify data from the Confirmatory Dataset. For the Untrustworthy and Trustworthy conditions, the model trained on the Exploratory Data from channel CP1 could discriminate Confirmatory Data above chance ($AUC = 0.549$, $CI_{95\%} [0.513, 0.584]$, $p_{perm} = .003$). Moderately Untrustworthy and Untrustworthy conditions could be discriminated above chance by a model trained at electrode P1 ($AUC = 0.556$, $CI_{95\%} [0.525, 0.588]$, $p_{perm} = .006$). For the Untrustworthy and ModeratelyUntrustworthy conditions, the exploratory data analysis had not

indicated a channel with above chance cross-validation discrimination performance.

ERP predictors of CS1 Learning Rate

Learning rate classifiers trained on the Exploratory Dataset at channels PO7 and Oz were used to discriminate between fast and slow learners in the Confirmatory Dataset. However, on the Confirmatory Dataset these classifiers did not perform significantly above chance (Oz AUC = .606, 95 % CI [0.423, 0.761], $p_{perm} = .113$); PO7 AUC = .480, 95 % CI [0.311, 0.646], $p_{perm} = .596$).

Response to CS2

Whole scalp SVM.

For all contrasts, none of the channels that performed above chance for the Exploratory Dataset, could discriminate conditions above chance for the Confirmatory Dataset.

EEG Multivariate

Response to CS1

Multivariate analyses on CS1 confirmed the Exploratory Dataset findings that information about differences between the Untrustworthy and Trustworthy conditions and between the Untrustworthy and ModeratelyUntrustworthy conditions were encoded in the EEG power. Information discriminating between Untrustworthy and Trustworthy conditions was present in the beta frequency range (16 – 20 Hz) in 3 of the 4 electrodes in the cluster identified in the Exploratory Dataset (P3, P5, PO7), with AUC ranging from 0.535 (P5, CI_{95%} [0.508, 0.562], $p_{perm} = .009$) to 0.545 (PO7, CI_{95%} [0.509, 0.578], $p_{perm} = .005$). Information discriminating between the ModeratelyTrustworthy and Trustworthy conditions was present at previously identified channel P1 at 20 Hz yielding a discrimination performance AUC of 0.542 (CI_{95%} [0.516, 0.569]), $p_{perm} = .004$). As no clusters discriminating between the Untrustworthy and ModeratelyUntrustworthy conditions were identified in the Exploratory dataset, no such classification was conducted on the Confirmatory Dataset.

Response to CS2

Classification analysis on CS2 did not reveal any information in the power data that yielded classification of CS2 responses. Although statistical tests returned significant results for all three condition pairs, and these tests passed the FDR correction at 5% comparisons of Untrustworthy vs. Trustworthy, and ModeratelyUntrustworthy vs. Trustworthy, no clusters spanned at least 3 frequency bins and 2 neighboring electrodes, which suggests that the significant results likely emerged from random noise.

Discussion

We used an economic trust game, in conjunction with a second-order conditioning procedure, to examine the behavioral and biological markers of learned trust/mistrust as well as related social judgments and attitudes. This entailed identifying putative neural correlates of these processes and assessing the degree to which these markers allow prediction of an individual's judgments and attitudes.

Behavioral Markers of Learned Trust

We first examined how trustworthiness judgments were affected by direct economic exchange using an economic trust game. In the game, faces were repeatedly paired with one of three monetary outcomes. In both the Exploratory and Confirmatory datasets, repeated association with greater reward resulted in increased perceived trustworthiness, as indicated by larger transfer sums in a one-shot trust game. This confirms earlier findings on the dynamic nature of trustworthiness judgments (Chang et al., 2010). Additionally, it extends prior findings by showing that economic exchange outcomes affect subjective judgments of warmth and global attitudes toward people. We also observed similar, but weaker, effects on judgments of attractiveness, with decreased perceived attractiveness for faces associated with negative outcomes, and effects on competence, with increased perceived competence for faces associated with positive outcomes. This indicates that economic exchange affects social judgments that are outside of the context of the economic exchange itself. Such social judgments are crucial to social interactions, and can determine the choice of non-economic social partners (Fiske et al., 2006) or mates (Little et al., 2011). Interestingly, the trust game had a weaker effect on dominance judgments, which suggests that they are not driven by the type of behavior modeled by the trust game. This is in agreement with the facial feature literature, which indicates that dominance judgments are independent of trustworthiness judgments (Oosterhof & Todorov, 2008).

The most novel behavioral result is that trustworthiness and related social judgments could be altered through second-order conditioning when participants were aware of the contingency between CS1 face and game outcome. The effect of second-order conditioning on judgments primarily emerged between Untrustworthy and Trustworthy faces and was present in both the Exploratory and Confirmatory datasets. In the one-shot trust game, participants sent larger sums to CS2 faces that had been paired with Trustworthy CS1 faces. Moreover, subjective judgments of trustworthiness, global attitudes, valence, warmth, attractiveness, and competence were enhanced for CS2 faces that had been paired with trustworthy CS1 faces. There were no effects on dominance judgments, which is in agreement with responses to the trust game.

Finally, although self-reported arousal was increased for Untrustworthy and Trustworthy faces, which

may be linked to the goal-relevance of these images with respect to financial gains and losses, skin conductance measures of physiological arousal did not differ between conditions. This suggests that behavioral responses might have been driven primarily by attention-orienting to goal-relevant faces (Vogt et al., 2009), with limited physiological consequences.

Neural Markers of Attitude Change

A major goal of the study was to detect neural markers of attitudes and assess the extent to which they could predict behavioral outcomes.

Late Positive Potential (LPP)

The LPP response recorded during the trust game reflected the change in trustworthiness judgments induced by game outcomes. The posterior aspect of the LPP was enhanced for Trustworthy faces in both the Exploratory and Confirmatory datasets. The LPP is considered a direct marker of attitudes (Cacciopo et al., 1996; Olofsson et al., 2008) and has been characterized as reflecting the motivational relevance of a target stimulus (Bradley et al., 2007; Schupp et al., 2000). In other words, the LPP marks a goal-relevant valenced stimulus, to which greater processing resources should be allocated (Hajcak et al., 2013). The Confirmatory dataset results obtained here suggest that the LPP indexed the goal relevance of Trustworthy and Untrustworthy stimuli, as these two face categories had a substantial impact on participants' earnings. Differences in the response patterns between the Exploratory and Confirmatory datasets might be due to inter-individual moderators of goal-relevance, such as sensitivity to negative reward (Gray, 1987; Kim et al., 2015).

In the Exploratory dataset the LPP following second-order conditioning was enhanced for faces paired with more negative outcomes and the amplitude predicted the sum transferred to CS2 faces in the one-shot trust game. However, in the Confirmatory dataset the LPP response was smallest for ModeratelyUntrustworthy faces. The change in pattern of responses between the Exploratory and Confirmatory datasets may have resulted from restricting the analysis to a specific time-window, an approach likely more sensitive to changes in response latency across groups of participants.

Multivariate ERP markers

We used machine-learning techniques to find neural markers in a data-driven fashion by performing a broad search of the parameter space for discriminative features. We then tested whether these discriminative features could predict new participants' responses to the stimuli. For the trust game data, cross-validation demonstrated that a model trained on data recorded over posterior areas could discriminate perceived face trustworthiness above chance. Importantly, the model could also discriminate perceived face trustworthiness above chance using data from these channel locations in the Confirmatory dataset (i.e., new participants). This indicates that neural signals captured at parietal

electrodes contain information that robustly encodes trustworthiness information learned about faces.

Examination of feature weights suggested that these reflect processes likely linked to early markers of attentional capture and later LPP markers of goal salience (Olofsson et al., 2008).

The ability of these findings to generalize to new data might be due to the analysis leveraging on a temporally broad set of markers that encompass a 1.5 s time interval post-stimulus onset, and that cover both early perceptual and later salience related processes. An additional analysis was conducted to attempt to predict inter-individual differences in sensitivity to learning from ERP markers. This did not identify markers that could discriminate between participants learning rates derived from a popular model of associative learning (Rescola & Wagner, 1972).

For the second-order conditioning task, some channels emerged as potentially discriminating between conditions, but classifiers trained with these data did not perform above chance on the Confirmatory dataset. This likely reflects the volatility of the neural response to CS2 stimuli that was already observed in the LPP analysis, and the sensitivity to changes in temporal distribution of neural signals across individuals.

Multivariate EEG Power markers

Multivariate analysis of EEG power was performed using the same strategy as for the ERP data. Classifiers trained on beta band power at posterior sites discriminated between Untrustworthy and Trustworthy faces in both the Exploratory and Confirmatory datasets. These markers likely index processes that are independent from the ERP markers, and seem to capture information about the relative emotional valence of the stimuli. Beta oscillations have been linked to arousal (Aftanas et al., 2005) and visual attention (Escoffier et al., 2015; Wróbel et al., 2000). Hence, beta power may index attentional processes, which likely function as a proxy for discriminating valence associated to untrustworthy and trustworthy faces in economic games.

The classifier approach could not isolate channels and frequency bands that carried discriminative features for CS2 stimuli following second-order conditioning. This might be due to the variability of participants' responses, and the fact that, in absence of an explicit task, the manipulation might have been too subtle to strongly engage the kind of attentional processes that index learned trustworthiness and related attitudes.

Summary

In both the Exploratory and Confirmatory datasets, repeated association with positive trust game outcomes resulted in increased perceived trustworthiness, as indicated by larger transfer sums in a one-shot trust game. Additionally, the economic exchange outcomes affected subjective ratings of

trustworthiness, warmth, global attitudes toward the person, attractiveness, and competence. Importantly, when participants were aware of the contingency between Trustee face and game outcome, the second-order conditioning procedure affected performance in a one-shot trust game in that participants sent larger sums to CS2 faces that had been paired with trustworthy CS1 faces. In addition, it affected subjective judgments of trustworthiness, global attitudes, valence, warmth, attractiveness, and competence. Self-reported arousal was also increased for Untrustworthy and Trustworthy faces, but the skin conductance measures of physiological arousal did not differ among conditions.

The late positive potential (LPP) event related potential (ERP) response recorded during the trust game was larger for Trustworthy faces, as compared to Untrustworthy faces, in both the exploratory and confirmatory datasets. Machine-learning analysis of the ERP data revealed that a model trained on data recorded at posterior electrodes could discriminate perceived face trustworthiness above chance and discriminate among conditions for the new data recorded in the confirmatory experiment. This indicates that these neural signals contained information that robustly encoded trustworthiness information learned about faces. Multivariate analysis of EEG power revealed that classifiers trained on beta band power at posterior sites discriminated between untrustworthy and trustworthy faces in both the Exploratory and Confirmatory datasets. For the second-order conditioning task, some channels discriminated between conditions, but classifiers trained with these data did not perform above chance on the confirmatory dataset.

In sum, the behavioral results indicate that robust changes in perceived trustworthiness, and a range of social judgments can be achieved through first-order (i.e., Trust game) and second-order conditioning procedures. Moreover, ERP and EEG markers indicate a participant's trust categorization of interaction partners following first-order conditioning, but are unable to accurately predict categorization behavior following second-order conditioning. Overall, the evidence indicates that ERP/EEG biomarkers offer promise as a tool for predicting trust attitudes/behavior, but that the sensitivity of these measures is context/task dependent.

References

- Aftanas, L. I., & Pavlov, S. V. (2005). Trait anxiety impact on posterior activation asymmetries at rest and during evoked negative emotions: EEG investigation. *International Journal of Psychophysiology*, 55(1), 85–94.
- Akbani, R., Kwek, S., & Japkowicz, N. (2004). Applying support vector machines to imbalanced datasets. In *European conference on machine learning* (pp. 39–50). Springer.
- Andrzej, & others. (2000). Beta activity: a carrier for visual attention. *Acta Neurobiologiae Experimentalis*, 60(2), 247–260.
- Baeyens, F., Eelen, P., & Bergh, O. van den. (1990). Contingency awareness in evaluative conditioning: A case for unaware affective-evaluative learning. *Cognition and Emotion*, 4(1), 3–18.
- Balconi, M., & Lucchiari, C. (2006). EEG correlates (event-related desynchronization) of emotional face elaboration: a temporal analysis. *Neuroscience Letters*, 392(1), 118–123.
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278.
- Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv Preprint arXiv:1506.04967*.
- Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., & others. (2015). lme4: Linear mixed-effects models using Eigen and S4, 2014. *R Package Version*, 1–1.
- Benedek, M., & Kaernbach, C. (2010). A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods*, 190(1), 80–91.
- Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1), 122–142.
- Bode, S., Sewell, D. K., Lilburn, S., Forte, J. D., Smith, P. L., & Stahl, J. (2012). Predicting perceptual decision biases from early brain activity. *The Journal of Neuroscience*, 32(36), 12488–12498.
- Bolker, B. (2007). Stochastic simulation and power analysis. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.720.8252>
- Bolker, B. M. (2008). *Ecological models and data in R*. Princeton University Press.
- Boone, R. T., & Buck, R. (2003). Emotional expressivity and trustworthiness: The role of nonverbal behavior in the evolution of cooperation. *Journal of Nonverbal Behavior*, 27(3), 163–182.
- Boucsein, W. (2012). *Electrodermal activity*. Springer Science & Business Media.
- Bradley, M. M., Hamby, S., Löw, A., & Lang, P. J. (2007). Brain potentials in perception: Picture complexity and emotional arousal. *Psychophysiology*, 44(3), 364–373.
- Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1), 49–59.
- Cacioppo, J. T., Crites, S. L., & Gardner, W. L. (1996). Attitudes to the right: Evaluative processing is

- associated with lateralized late positive event-related brain potentials. *Personality and Social Psychology Bulletin*, 22(12), 1205–1219.
- Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS Scales. *Journal of Personality and Social Psychology*, 67(2), 319.
- Cawley, G. C., & Talbot, N. L. (2001). Manipulation of prior probabilities in support vector classification. In *Neural Networks, 2001. Proceedings. IJCNN'01. International Joint Conference on* (Vol. 4, pp. 2433–2438). IEEE.
- Chang, L. J., Doll, B. B., van 't Wout, M., Frank, M. J., & Sanfey, A. G. (2010). Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology*, 61(2), 87–105.
- Coan, J. A., & Allen, J. J. (2003). Frontal EEG asymmetry and the behavioral activation and inhibition systems. *Psychophysiology*, 40(1), 106–114.
- Coan, J. A., & Allen, J. J. (2004). Frontal EEG asymmetry as a moderator and mediator of emotion. *Biological Psychology*, 67(1), 7–50.
- Critchley, H. D., Elliott, R., Mathias, C. J., & Dolan, R. J. (2000). Neural activity relating to generation and representation of galvanic skin conductance responses: A functional magnetic resonance imaging study. *Journal of Neuroscience*, 20(8), 3033–3040.
- Cunningham, W. A., Espinet, S. D., DeYoung, C. G., & Zelazo, P. D. (2005). Attitudes to the right-and left: frontal ERP asymmetries associated with stimulus valence and processing goals. *NeuroImage*, 28(4), 827–834.
- Cuthbert, B. N., Schupp, H. T., Bradley, M. M., Birbaumer, N., & Lang, P. J. (2000). Brain potentials in affective picture processing: covariation with autonomic arousal and affective report. *Biological Psychology*, 52(2), 95–111.
- Davidson, R. J., Ekman, P., Saron, C. D., Senulis, J. A., & Friesen, W. V. (1990). Approach-withdrawal and cerebral asymmetry: Emotional expression and brain physiology: I. *Journal of Personality and Social Psychology*, 58(2), 330.
- De Houwer, J., Thomas, S., & Baeyens, F. (2001). Association learning of likes and dislikes: A review of 25 years of research on human evaluative conditioning. *Psychological Bulletin*, 127(6), 853.
- 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.
- Etzel, J. A., & Braver, T. S. (2013). MVPA permutation schemes: Permutation testing in the land of cross-validation. In *Pattern recognition in neuroimaging (prni), 2013 international workshop on* (pp. 140–143). IEEE.
- Etzel, J. A., Zacks, J. M., & Braver, T. S. (2013). Searchlight analysis: Promise, pitfalls, and potential. *NeuroImage*, 78, 261–269.
- Evans, A. M., & Revelle, W. (2008). Survey and behavioral measurements of interpersonal trust.

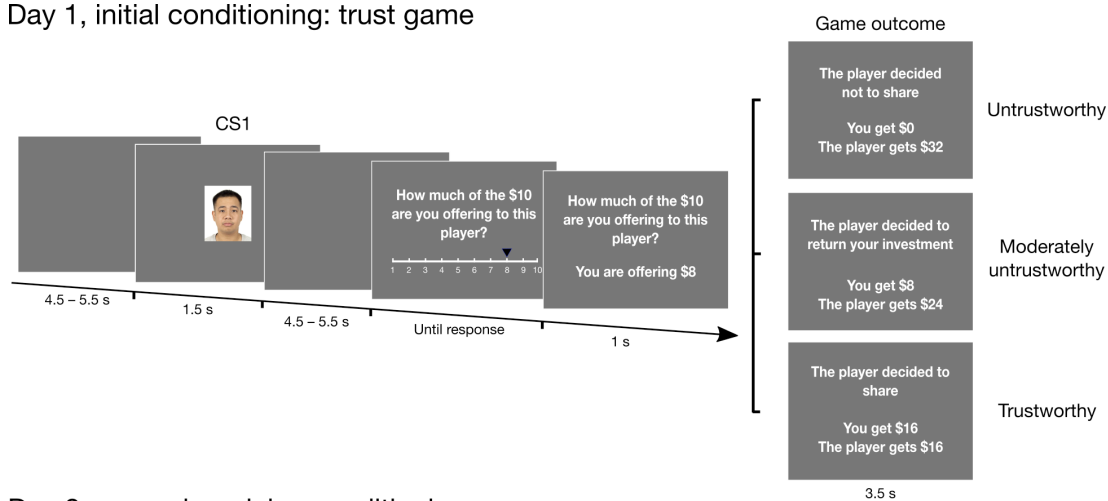
- Journal of Research in Personality*, 42(6), 1585–1593.
- Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences*, 11(2), 77–83.
- Flykt, A., Esteves, F., & Öhman, A. (2007). Skin conductance responses to masked conditioned stimuli: Phylogenetic/ontogenetic factors versus direction of threat? *Biological Psychology*, 74(3), 328–336.
- Frith, C. D., & Allen, H. A. (1983). The skin conductance orienting response as an index of attention. *Biological Psychology*, 17(1), 27–39.
- Gray, J. A. (1987). *The psychology of fear and stress* (Vol. 5). CUP Archive.
- Green, P., & MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*.
- Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annual Review of Neuroscience*, 37(1), 435–456.
- Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539), 2425–2430.
- Haynes, J.-D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7(7), 523–534.
- Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: a meta-analysis. *Psychological Bulletin*, 136(3), 390.
- Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17(3), 299–310.
- Johnson, N. D., & Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology*, 32(5), 865–889.
- Kerkhof, I., Vansteenwegen, D., Baeyens, F., & Hermans, D. (2011). Counterconditioning: an effective technique for changing conditioned preferences. *Experimental Psychology*, 58(1), 31–38.
- Kim, S. H., Yoon, H., Kim, H., & Hamann, S. (2015). Individual differences in sensitivity to reward and punishment and neural activity during reward and avoidance learning. *Social Cognitive and Affective Neuroscience*, 10(9), 1219–1227.
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). Package “lmerTest.” *R Package Version*, 2.
- LeDell, E., Petersen, M., & van der Laan, M. (2015). Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. *Electronic Journal of Statistics*, 9(1), 1583.
- Little, A. C., Jones, B. C., & DeBruine, L. M. (2011). Facial attractiveness: evolutionary based research. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*,

- 366(1571), 1638–1659.
- Marzi, T., Righi, S., Ottonello, S., Cincotta, M., & Viggiano, M. P. (2014). Trust at first sight: evidence from ERPs. *Social Cognitive and Affective Neuroscience*, 9(1), 63–72.
- Olofsson, J. K., Nordin, S., Sequeira, H., & Polich, J. (2008). Affective picture processing: An integrative review of ERP findings. *Biological Psychology*, 77(3), 247–265.
- Olson, J. M., & Zanna, M. P. (1993). Attitudes and attitude change. *Annual Review of Psychology*, 44(1), 117–154.
- Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, 105(32), 11087–11092.
- Oxley, D. R., Smith, K. B., Alford, J. R., Hibbing, M. V., Miller, J. L., Scalora, M., ... Hibbing, J. R. (2008). Political attitudes vary with physiological traits. *Science*, 321(5896), 1667–1670. <https://doi.org/10.1126/science.1157627>
- Palmer, J., Kreutz-Delgado, K., Rao, B., & Makeig, S. (2007). Modeling and estimation of dependent subspaces with non-radially symmetric and skewed densities. *Independent Component Analysis and Signal Separation*, 97–104.
- Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, 89(3), 277–293.
- Payne, B. R., Lee, C.-L., & Federmeier, K. D. (2015). Revisiting the incremental effects of context on word processing: Evidence from single-word event-related brain potentials. *Psychophysiology*, 52(11), 1456–1469.
- Pleyers, G., Corneille, O., Luminet, O., & Yzerbyt, V. (2007). Aware and (dis) liking: item-based analyses reveal that valence acquisition via evaluative conditioning emerges only when there is contingency awareness. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(1), 130.
- Priester, J. R., & Petty, R. E. (1996). The gradual threshold model of ambivalence: relating the positive and negative bases of attitudes to subjective ambivalence. *Journal of Personality and Social Psychology*, 71(3), 431.
- Rescorla, R. A. (1980). *Pavlovian second-order conditioning: studies in associative learning*. Hillsdale, N.J.; New York: L. Erlbaum Associates ; Distributed by Halsted Press.
- Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (Vol. 2, pp. 64–99). New York, NY: Appleton Century Crofts.
- Roth, W. T., Dawson, M. E., & Fillion, D. L. (2012). Publication recommendations for electrodermal measurements. *Psychophysiology*, 49(8), 1017–1034.
- Schulz, E., Zherdin, A., Tiemann, L., Plant, C., & Ploner, M. (2012). Decoding an individual's

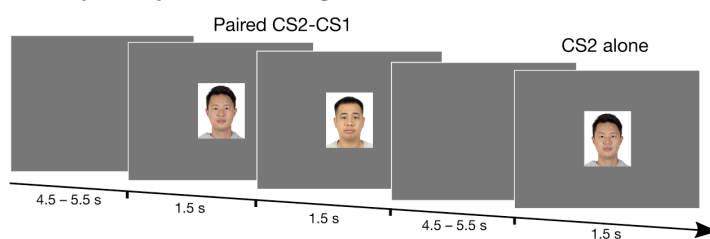
- sensitivity to pain from the multivariate analysis of EEG data. *Cerebral Cortex*, 22(5), 1118–1123.
- Thiruchselvam, R., Blechert, J., Sheppes, G., Rydstrom, A., & Gross, J. J. (2011). The temporal dynamics of emotion regulation: An EEG study of distraction and reappraisal. *Biological Psychology*, 87(1), 84–92.
- Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology*, 66(1), 519–545.
- Tortosa, M. I., Strizhko, T., Capizzi, M., & Ruz, M. (2013). Interpersonal effects of emotion in a multi-round trust game. *Psicologica: International Journal of Methodology and Experimental Psychology*, 34(2), 179–198.
- van 't Wout, M., & Sanfey, A. G. (2008). Friend or foe: The effect of implicit trustworthiness judgments in social decision-making. *Cognition*, 108(3), 796–803.
<https://doi.org/10.1016/j.cognition.2008.07.002>
- Veropoulos, K., Campbell, C., & Cristianini, N. (1999). Controlling the sensitivity of support vector machines. In *Proceedings of the international joint conference on AI* (pp. 55–60).
- Walther, E. (2002). Guilty by mere association: Evaluative conditioning and the spreading attitude effect. *Journal of Personality and Social Psychology*, 82(6), 919–934.
- Yang, D., Qi, S., Ding, C., & Song, Y. (2011). An ERP study on the time course of facial trustworthiness appraisal. *Neuroscience Letters*, 496(3), 147–151.

Figures

Day 1, initial conditioning: trust game



Day 2, secondary delay conditioning



Post-conditioning tests

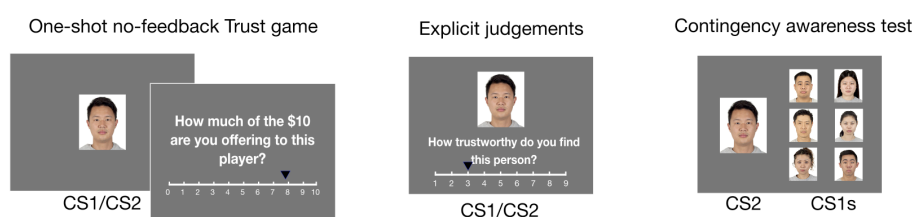


Figure 1. Experiment timeline. On Day 1, only explicit judgments were recorded. On Day 2 all three tests were presented.

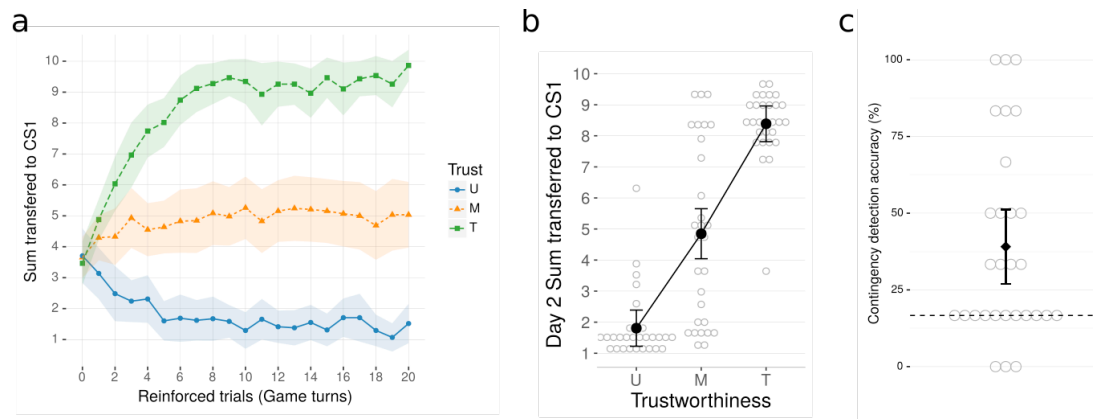


Figure 2. Behavioral responses to CS1. (a) Sum transferred to CS1 faces in Day 1 trust game as a function of Trustworthiness condition and the number of experienced reinforced trials. (b) Sum transferred to CS1 faces on Day 2 single-shot trust game. (c) CS1-CS2 contingency detection accuracy on Day 2. Dotted line indicates chance performance level. Error bars represent 95% CIs. Dot plot layer displays raw ratings for each item and participant. U: Untrustworthy condition, M: Moderately untrustworthy, T: Trustworthy.

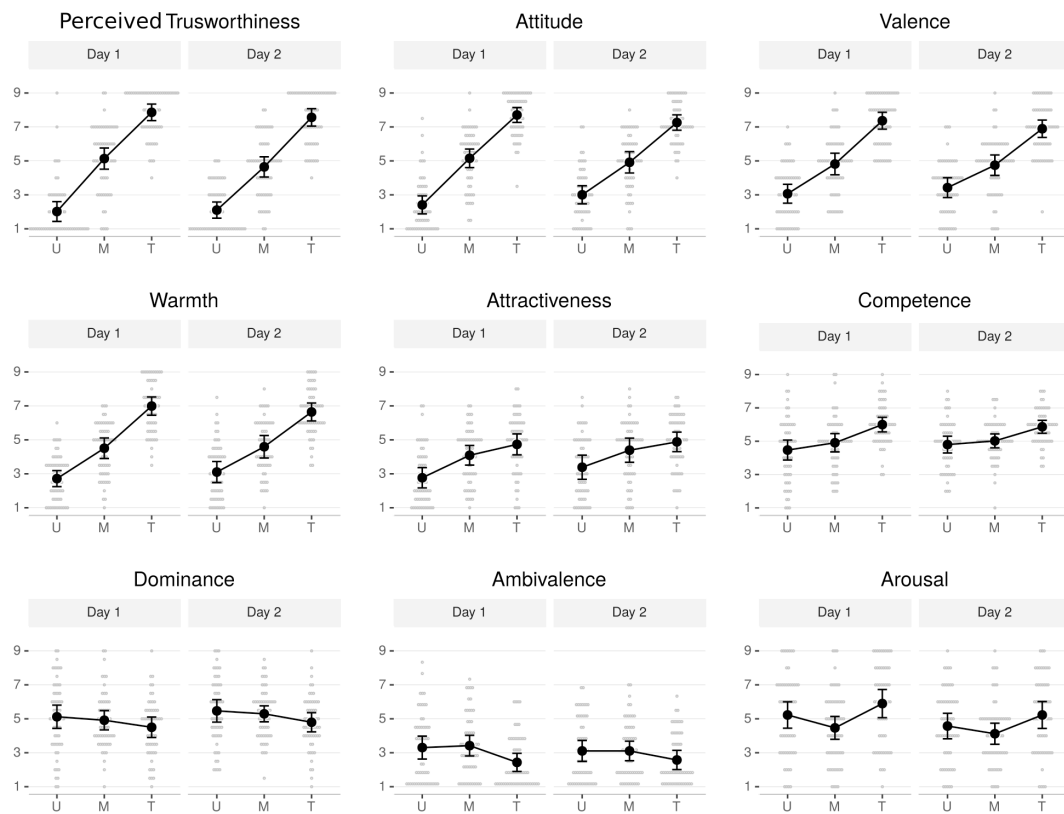


Figure 3. Mean behavioral judgments of CS1 faces on Day 1 and Day 2 after the trust game. Error bars indicate the 95% CIs derived from LMM model estimates. Dot plot layer displays raw ratings for each item and participant. U: Untrustworthy condition, M: Moderately Untrustworthy, T: Trustworthy.

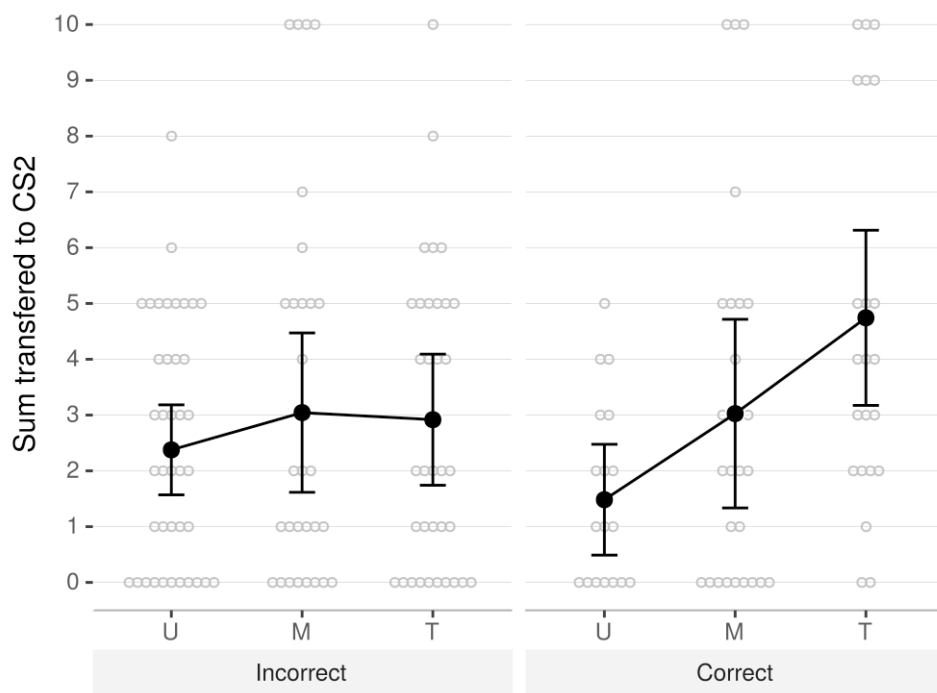


Figure 4. Sum transferred to CS2 after second-order conditioning, as a function of associated CS1 trustworthiness and CS1-CS2 contingency awareness. Error bars indicate the 95% CIs derived from LMM model estimates. Dot plot layer displays raw response for each item and participant. U: Untrustworthy condition, M: Moderately Untrustworthy, T: Trustworthy.

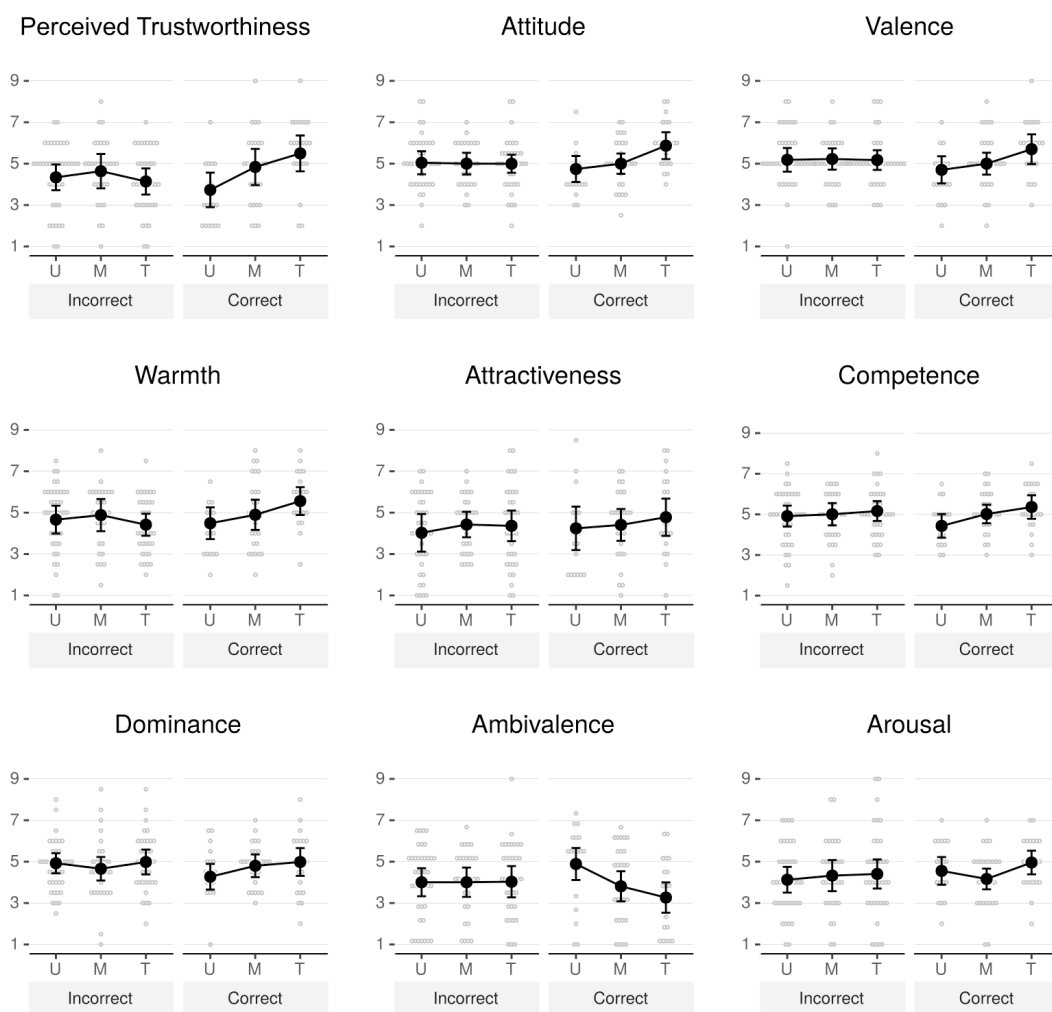


Figure 5. Mean behavioral ratings of CS2 faces after second-order conditioning as a function of associated CS1 trustworthiness, and CS1-CS2 contingency awareness. Error bars indicate the 95% CIs derived from LMM model estimates. Dot plot layer displays raw ratings for each item and participant. U: untrustworthy CS1, M: moderately untrustworthy CS1, T: trustworthy CS1.

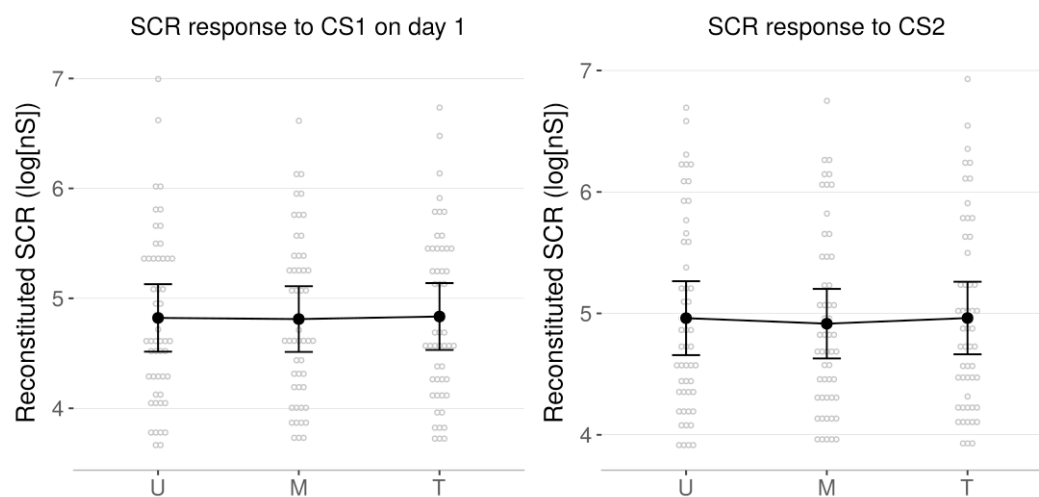


Figure 6. Skin conductance response to CS1 and CS2 faces after second-order conditioning as a function of associated CS1 trustworthiness. Error bars indicate the 95% CIs derived from LMM model estimates. Dot plot layer displays raw average response for each item and participant. U: Untrustworthy CS1, M: Moderately Untrustworthy CS1, T: Trustworthy CS1.

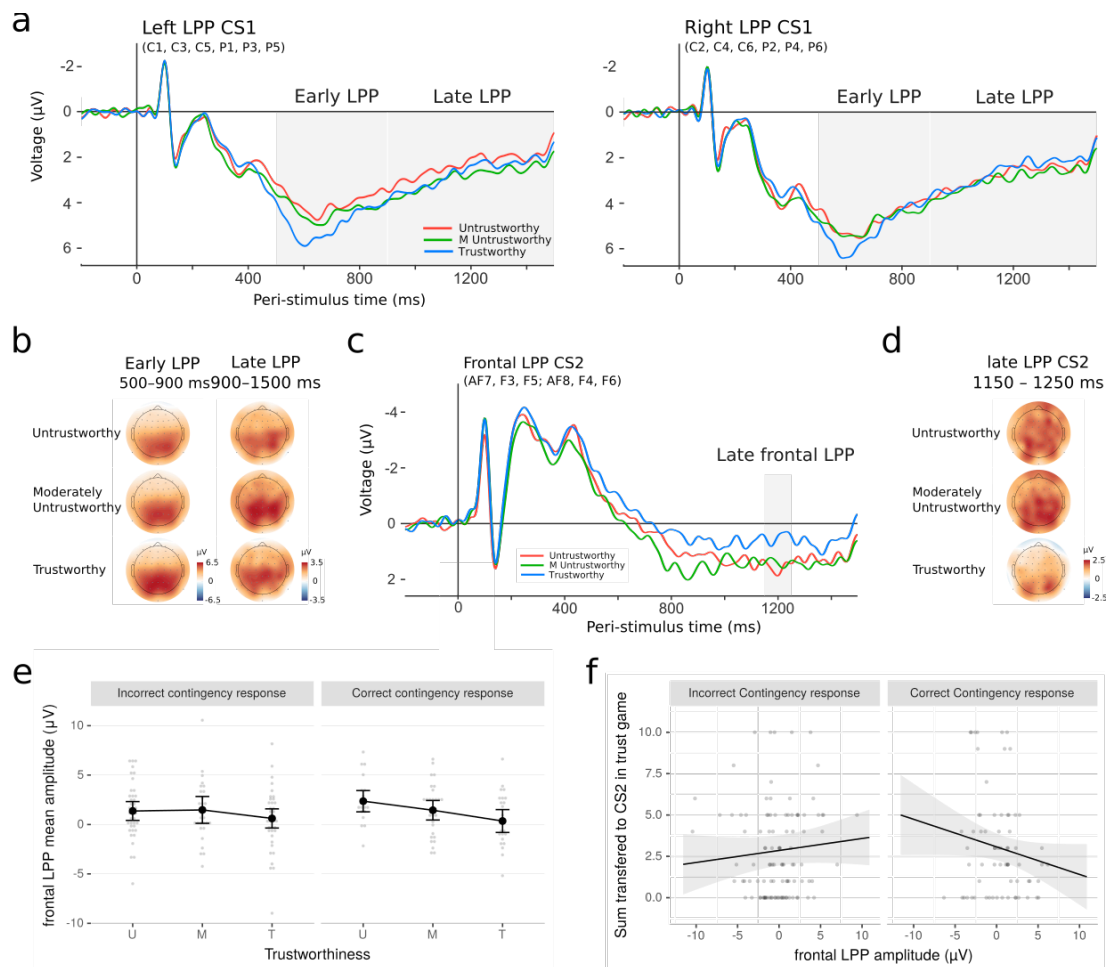


Figure 7. Event-related potential response to CS1 and CS2 faces. (a) LPP response to CS1 faces during Day 1 conditioning (repeated trust game) as a function of CS1 learned Trustworthiness. (b) Scalp topography of early and late LPP response to CS1 faces during Day 1 conditioning. (c) Frontal LPP to CS2 faces during Day 2 second-order conditioning as a function of associated CS1 trustworthiness. (d) Scalp topography of frontal LPP during second-order conditioning. (e) LPP response to CS2 as function of associated CS1 trustworthiness and CS1-CS2 contingency awareness. (f) Sum transferred to CS2 in the one-shot trust game as a function of frontal LPP amplitude and CS1-CS2 Contingency awareness. Error bars indicate the 95% CIs derived from LMM model estimates. Dot plot layer displays raw average response for each item and participant. U: Untrustworthy CS1, M: Moderately untrustworthy CS1, T: Trustworthy CS1.

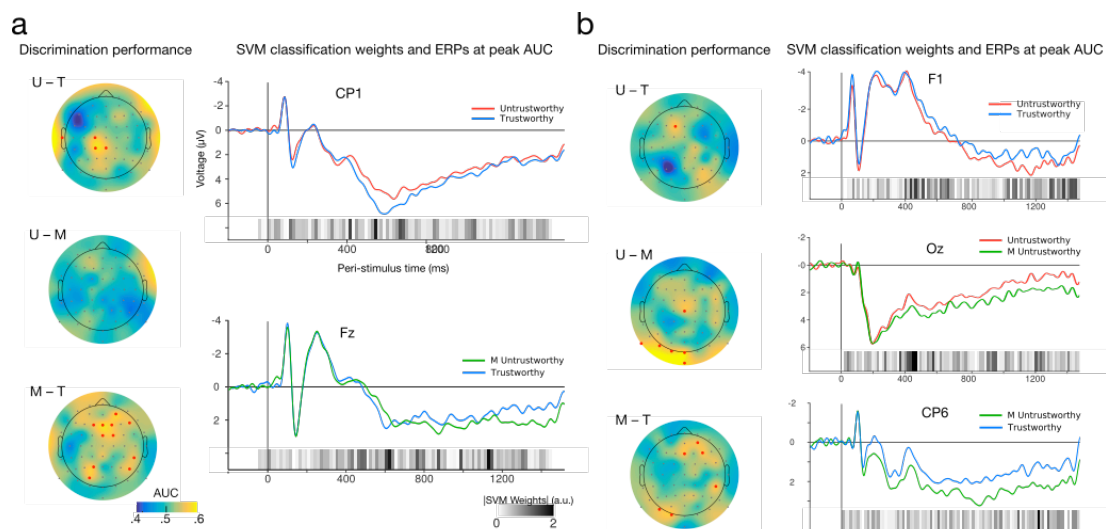


Figure 8. ERP multivariate analysis of the Exploratory dataset. Linear SVM classifiers were trained at each channel using the time course of the ERP as discrimination features. (a) Day 1 Trust game, scalp map of cross-validated discrimination performance for each condition pair (left), and feature weight plotted against corresponding average ERPs (right). (b) Day 2 secondary conditioning, scalp map of discrimination performance (left), and feature weight plotted against corresponding average ERPs (right). Red markers indicate channels with above-chance performance.

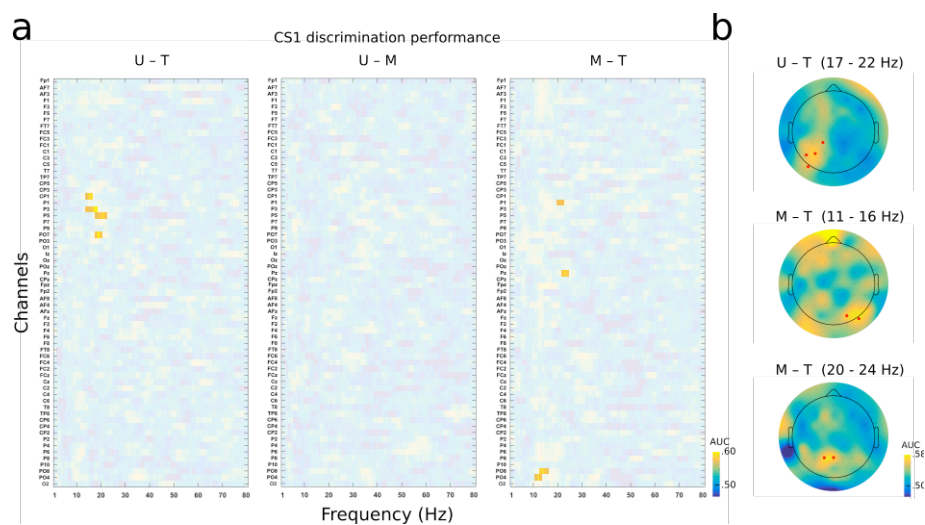


Figure 9. EEG multivariate analysis results of Exploratory dataset. Linear SVM classifiers were trained at each channel and frequency using time course of EEG power as discrimination features. (a) Cross-validated discrimination performance for each condition pair at each channel and frequency pair. Above-mask clusters mark above-chance discrimination performance that passed a cluster threshold of 3 frequency bins and 2 neighboring channels. (b) Scalp topography of discrimination performance for each condition pair and above-chance cluster. At each channel, performance was averaged over the frequency range spanned by each cluster. Red markers indicate channels with above-chance performance.

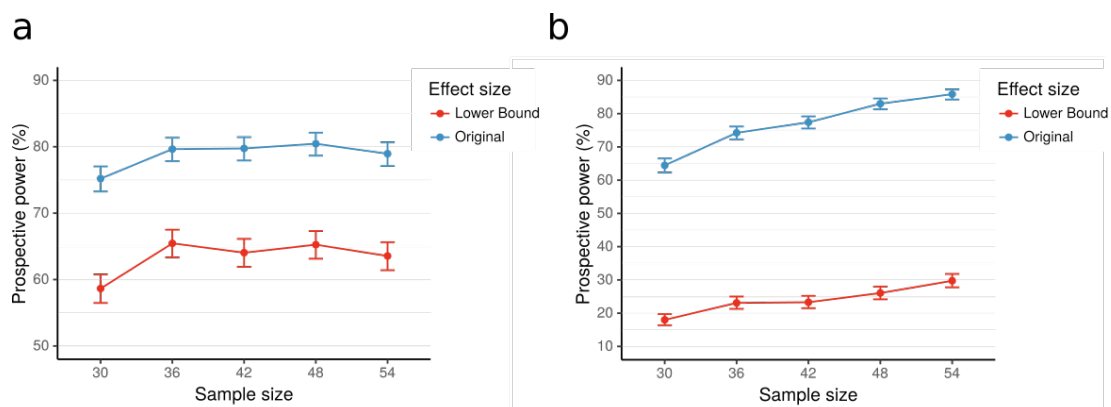


Figure 10. Behavior and ERP simulations of prospective power for tests at $\alpha = 0.05$ of original and lower bound effects sizes. (a) Behavior effect of interest was the difference in sum transferred between trustworthy and untrustworthy CS2s for which contingency was correctly remembered. Lower bound effect was derived from the 95% CI lower bound. (b) ERP effect of interest was the LPP prediction of sum transferred on Day 2, when contingency was remembered. For lower bound the lower effect of theoretical interest was used.

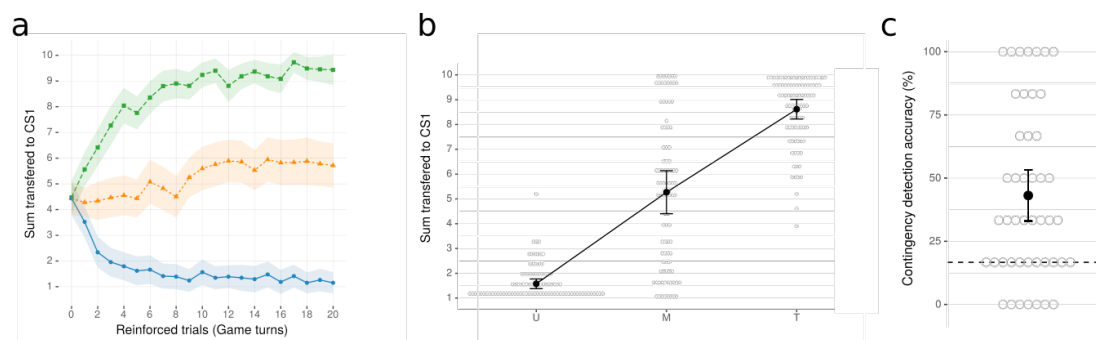


Figure 11. Confirmatory dataset analysis of behavioral responses to CS1. (a) Sum transferred to CS1 faces in Day 1 trust game as a function of Trustworthiness and the number of experienced reinforced trials. (b) Sum transferred to CS1 faces on Day 2 single-shot trust game. (c) CS1-CS2 contingency detection accuracy on Day 2. Dotted line indicates chance performance level. Error bars represent 95% CIs. Dot plot layer displays raw ratings for each item and participant. U: Untrustworthy, M: Moderately Untrustworthy, T: Trustworthy.

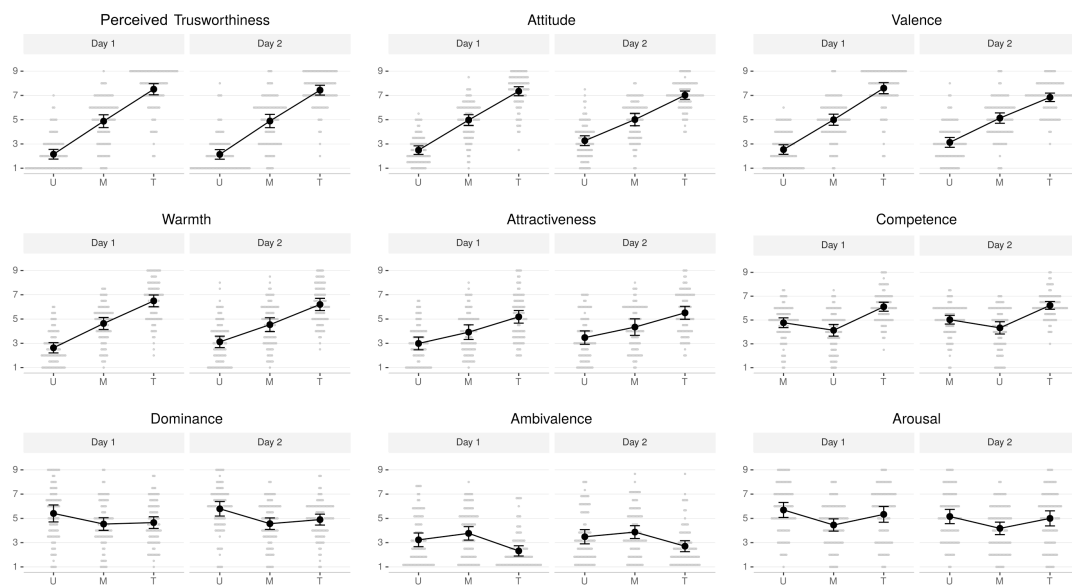


Figure 12. Mean validation dataset behavioral ratings of CS1 faces after trust game conditioning on Day 1 and Day 2. Error bars indicate the 95% CIs derived from LMM model estimates. Dot plot layer displays raw ratings for each item and participant. U: Untrustworthy, M: Moderately Untrustworthy, T: Trustworthy.

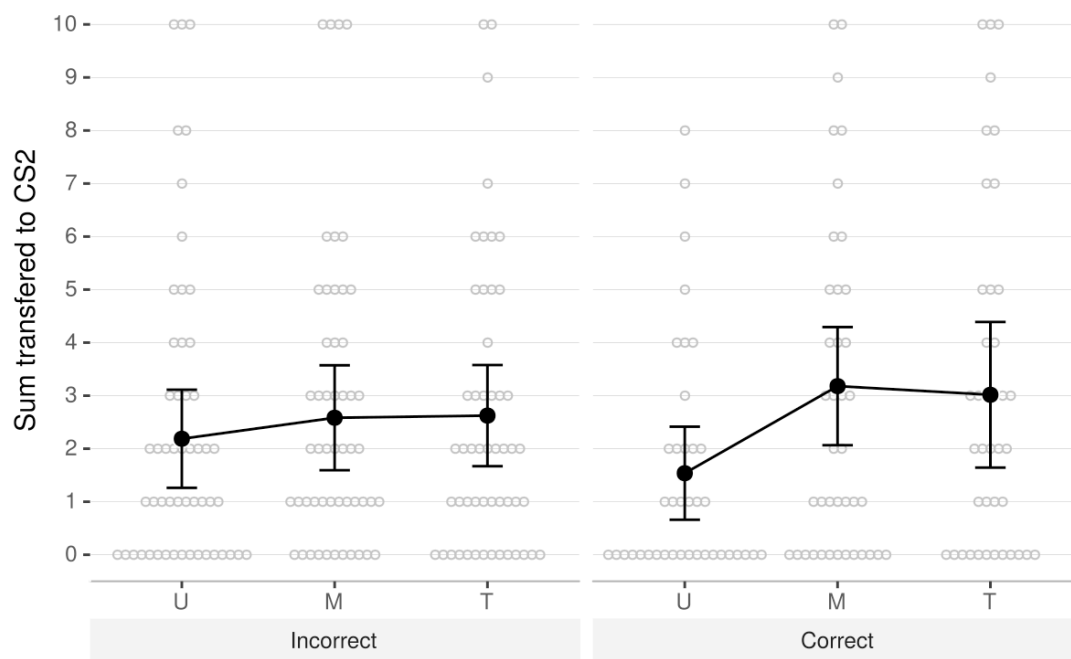


Figure 13. Sum transferred to CS2 after second-order conditioning, as a function of associated CS1 trustworthiness and CS1-CS2 contingency awareness. Error bars indicate the 95% CIs derived from LMM model estimates. Dot plot layer displays raw response for each item and participant. U: Untrustworthy, M: Moderately Untrustworthy, T: Trustworthy.

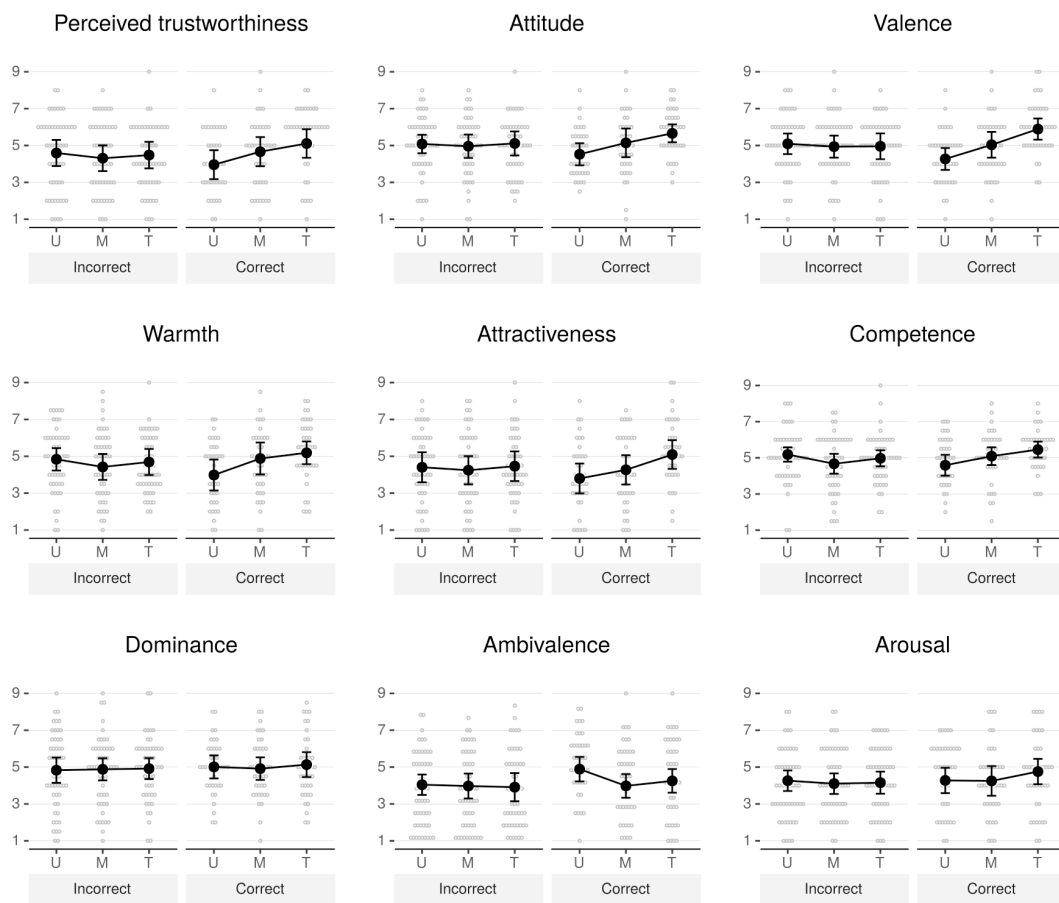


Figure 14. Mean Confirmatory dataset behavioral ratings of CS2 faces after second-order conditioning as a function of associated CS1 trustworthiness, and CS1-CS2 contingency awareness. Error bars indicate the 95% CIs derived from LMM model estimates. Dot plot layer displays raw ratings for each item and participants. U: Untrustworthy CS1, M: Moderately Untrustworthy CS1, T: Trustworthy CS1.

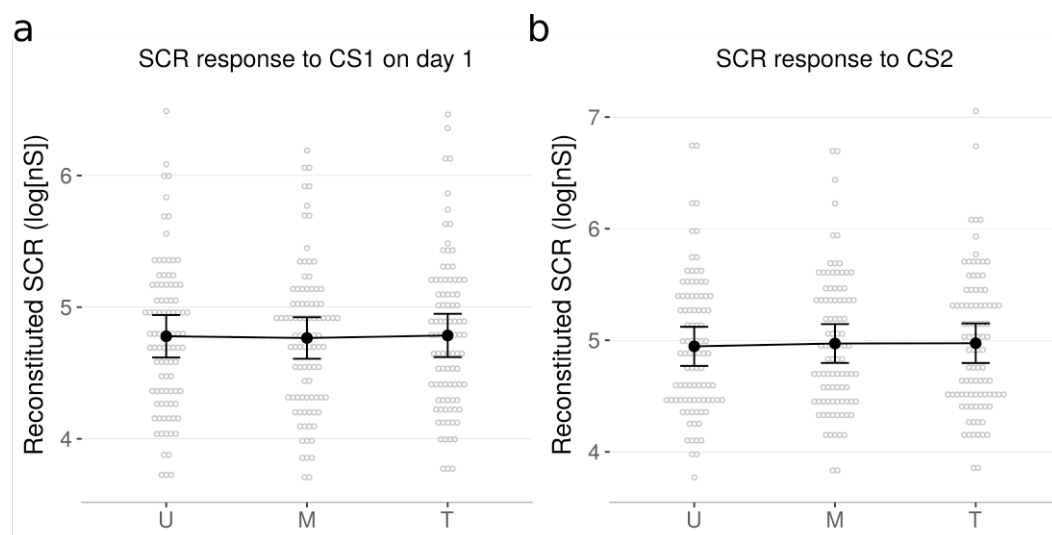


Figure 15. Confirmatory dataset skin conductance response to CS1 and CS2 faces after second-order conditioning as a function of associated CS1 trustworthiness. Error bars indicate the 95% CIs derived from LMM model estimates. Dot plot layer displays raw average response for each item and participant. U: Untrustworthy CS1, M: Moderately Untrustworthy CS1, T: Trustworthy CS1.

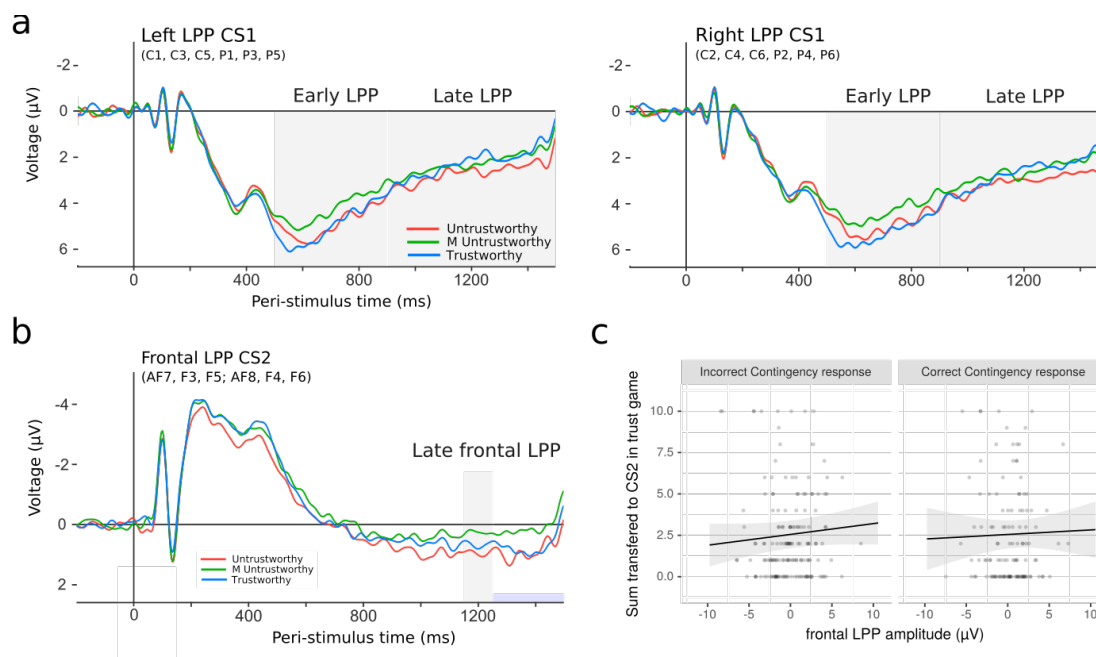


Figure 16. Event-related potential response to CS1 and CS2 faces in the Confirmatory dataset. (a) LPP response to CS1 faces during Day 1 conditioning (repeated trust game) as a function of CS1 learned trustworthiness. (b) Frontal LPP to CS2 faces during Day 2 second-order conditioning as a function of associated CS1 trustworthiness. Blue shading indicates the window used for unplanned analysis. (c) Sum transferred to CS2 in post-conditioning one-shot trust game as a function of frontal LPP amplitude and CS1-CS2 Contingency awareness. Error bars indicate the 95% CIs derived from LMM model estimates. Dot plot layer displays raw average response for each item and participant. U: Untrustworthy CS1, M: Moderately Untrustworthy CS1, T: Trustworthy CS1.

List of Publications and Significant Collaborations that resulted from your AOARD supported project:

Escoffier, N., Yen, S.C., Schirmer, A., & Penney, T.B. (manuscript in preparation). Using Brain-State Information to Facilitate Conditioned Attitude Formation.