Predictably angry-facial cues provide a credible signal of destructive behavior

van Leeuwen, B.; Noussair, C.N.; Offerman, T.; Suetens, S.; van Veelen, M.; van de Ven, J.

DOI
10.1287/mnsc.2017.2727

Publication date
2018

Document Version
Submitted manuscript

Published in
Management Science

Predictably Angry

Facial cues provide a credible signal of destructive behavior

Boris van Leeuwen[1], Charles N. Noussair[2], Theo Offerman[3],
Sigrid Suetens[4], Matthijs van Veelen[5], and Jeroen van de Ven[6]

November 2016

Abstract: Evolutionary explanations of anger as a commitment device hinge on two key assumptions. The first is that it is predictable, ex-ante, whether someone will get angry when feeling that they have been badly treated. The second is that anger is associated with destructive behavior. We test the validity of these two assumptions. We collected photos of responders in an ultimatum game *before* they were informed about the game that they would be playing, and filmed responders with webcams during play. We then showed pairs of photos, consisting of one responder who rejected, and one responder who accepted, a low offer, to an independent group of observers. We find that observers are better than chance at detecting who rejected the low offer; they do 10% better than random guessing would. We also find that anger at receiving a low offer is associated with rejection.

Keywords: anger, commitment, facial cues, ultimatum game, laboratory experiment.

Acknowledgements: We thank the Associate Editor and three anonymous referees of this journal for valuable comments and suggestions. For helpful comments and suggestions, we like to thank participants of seminars at U. Antwerp, U. Autònoma de Barcelona, NHH Bergen, U. Bonn, Chapman U., U. Edinburgh, U. East Anglia, U. Essex, Florence U., U. Frankfurt, U. Lyon, Maastricht U., U. Nijmegen, RWTH Aachen U., UC San Diego, UC Santa Barbara, U Vienna, at internal seminars at U Amsterdam, Tilburg U, Tinbergen Institute Amsterdam, and U. Toulouse, and at the following conferences or workshops: 2016 Discrimination at Work workshop at Erasmus U. Rotterdam, 2015 ASFEE conference in Paris, 2015 IMEBESS conference in Toulouse, 2015 Morality, incentives and unethical behavior conference at the UCSD Rady School of Management, 2014 Christmas meeting of Belgian economists at U Louvain, 2014 M-BEES Symposium at Maastricht U, the 2013 ABEE Symposium in Amsterdam, the 2013 TIBER Symposium in Tilburg, the 2013 ESA world meetings in Zurich. Financial support from the Research Priority Area Behavioral Economics of the University of Amsterdam, the ANR – Labex IAST and from the Netherlands Organization for Scientific Research (NWO) through the VIDI program (project number 452-11-012) is gratefully acknowledged.

[1] Department of Economics and CentER, Tilburg University, b.vanleeuwen@uvt.nl
[2] Department of Economics, University of Arizona, cnoussair@email.arizona.edu
[3] CREED and Tinbergen Institute, University of Amsterdam, t.j.s.offerman@uva.nl
[4] Department of Economics and CentER, Tilburg University, s.suetens@uvt.nl
[5] CREED and Tinbergen Institute, University of Amsterdam, and Program for Evolutionary Dynamics, Harvard University, c.m.vanveelen@uva.nl
[6] ACLE and Tinbergen Institute, University of Amsterdam, j.vandeven@uva.nl

1. Introduction

Anger is costly. Angry people are sometimes willing to pay a price to reduce the payoff of those whom they are angry with. But what benefits does anger bring? Frank (1987, 1988) suggests that anger in particular, and emotions in general, serve as a commitment device (see also Nesse, 2001). He rationalizes the existence of emotions by the effect that they have on the behavior of others. People might think twice before taking advantage of a person whom they believe could get angry. Getting angry, and being thought of as someone who becomes angry, can thus be beneficial, and lead to greater evolutionary success.[7] We refer to individuals, who are likely to engage in destructive behavior out of anger, as individuals with an 'angry button.'

The 'angry button hypothesis' states that anger serves as a commitment device to engage in costly punishment in case of a conflict. It rests on two crucial assumptions. The first assumption is that angry buttons are observable. Others should be able to recognize in some manner, for instance by using facial cues, whether someone will get angry when feeling badly treated. The second assumption is that anger leads to destructive behavior. Individuals with an angry button should be willing to pay a price to lower a competitor's payoff. Clearly, these are strong assumptions. It is far from obvious that people have the ability to identify who is prone to get angry. One difficulty is that anger is generated spontaneously in response to the action of another party, rather than in advance. Individuals must thus be able to correctly judge who will get angry before the actual anger is experienced and expressed. Moreover, it is not evident that observable correlates of angry buttons exist at all. If perceptible cues that correlate with a tendency to get angry do exist, and yield benefits to those that have them, then evolutionary pressure to mimic the signals of an angry person, but not the actual anger, must exist as well. This pressure could erode the credibility of angry buttons (Samuelson, 2001). In the current paper we rigorously test both components – observability and credibility – of the angry button hypothesis.

We report results from a laboratory experiment in which participants play an ultimatum game, hereafter UG (Güth, Schmittberger and Schwarze, 1982). In our version of the game, each pair of participants was given an endowment of €9. Participants in the role of proposer could offer either a (7,2) or a (4,5) split with the responder. Responders had the choice between accepting the proposal, and thereby dividing the money as proposed, or rejecting, and thereby destroying the entire endowment. As expected, all responders accepted the (4,5) offer, and a substantial number of responders (29 percent) rejected the (7,2) offer. This game provides an appropriate environment to test our hypothesis. A prominent feature of the game is the commitment problem that the responder has. In terms of purely monetary payoffs, responders have no reason to reject the (7,2) offer once it is on the table. The possibility that a responder might react with anger to a low offer, however, may help her commit to

[7] Güth and Yaari (1992) and Güth (1995) introduced the "indirect evolutionary approach" to formally demonstrate that the evolution of preferences can result in a stable equilibrium, in which there are agents with preferences over actions that differ from those that maximize their fitness. Dekel, Ely, and Yilankaya (2007) extend the results to a broader class of preferences.

rejecting the unfavorable offer. This makes it in the proposer's interest to offer (4,5) and have it accepted, rather than to offer (7,2) and have it rejected.

Several studies have suggested that anger is important in the decision to reject low offers in the UG. For instance, rejection rates are reduced if responders have more time to think or 'cool off' (Grimm and Mengel, 2011; Oechssler et al., 2008; Sutter et al., 2003), or if they can express their emotions (Xiao and Houser, 2005). Yamagishi et al. (2009) find that receivers reject low offers even if it only reduces their own payoff, and not that of the proposer, and they attribute this behavior to feelings of anger. Also, rejection of unfair offers has been associated with heightened activity in brain regions related to emotional decision-making (Gospic et al., 2011; Sanfey et al., 2003) and with greater skin conductance (van 't Wout et al., 2006).[8] To our knowledge, the ability to anticipate anger and rejection in others – a crucial part of evolutionary explanations of anger – has not yet been tested.

Our strategy to test whether people are able to recognize who rejects is the following. Before we explained the game to participants, we took photos and videos of participants in the role of responders. We also videotaped some of the responders during the game to verify whether responders who receive a low offer get angry. We then formed random pairs of responders who received the (7,2) offer, with one accepter and one rejecter each, and showed these to an independent set of observers. The observers were asked to identify the rejecter within the pair.

The observers do not have a stake in the game itself. This allows us to isolate the ability to predict rejections, and avoids potential confounds such as social preferences or risk aversion. Also, in order to exclude the possibility that responders were expressing certain emotions strategically, we emphasized that the photos and videos were taken before the responders were provided with the instructions for the UG. Andrade and Ho (2009) show that responders do strategically misrepresent their self-reported anger when they have the possibility to influence the proposers before they choose their offer. Gneezy and Imas (2014) show that people anticipate that anger may influence decision making and performance in strategic interaction and that people choose to strategically anger others if this in their best interest.

Our results provide support for both components – observability and credibility – of the angry button hypothesis: angry buttons exist and can be detected. We find that observers do 10% better than chance in identifying the rejecter when they see photos of responders. The effect is statistically significant, and is not driven by a few responders who are easy to judge: 71 percent of the responders were judged correctly more than half of the time. We also find that responders who reject a low offer express more anger when they see the low offer than those who accept it. Some rejecters have a 'larger' angry button, that is, they get angrier than others after receiving a low offer. Observers do particularly

[8] Note that feelings of unfairness and anger are not necessarily the same. One may well perceive a low offer in the UG as unfair without getting angry. Furthermore, a responder might get angry at a low offer without rejecting it.

well in identifying rejecters when the rejecter has a large angry button; they do about 25% better than chance in such instances.

If observers never directly see which responders got angry after receiving a low offer, on what basis can they identify the rejecting responder? As they must rely entirely on facial cues, we measured a range of facial features that have been related to aggression or (anti-)social behavior. These include facial width-to-height-ratio (fWHR), facial asymmetry (fA), masculinity, and attractiveness. We investigated whether they correlate with rejection of low offers and/or observers' judgments of rejection. We find that fA is strongly (positively) correlated with the decision to reject a low offer. To illustrate, observers would have identified the rejecter correctly in 69% of the cases if they had always selected the more asymmetric person of the pair as the rejecter. Observers correctly perceive fA as a correlate of rejection, though they underestimate the magnitude of the true correlation.

Our paper is related to a literature that studies whether people can identify a cooperative attitude in others. Several studies suggest that people do better than chance in detecting who cooperates in social dilemmas, such as a prisoner's dilemma or public good game (Belot, Bhaskar, and van de Ven 2012; Brosig 2002; Dawes, McTavish, and Shaklee 1977; Frank, Gilovich, and Regan 1993; Kovács-Bálint, Bereczkei, and Hernádi 2013; Tognetti et al. 2013; Verplaetse, Vanneste, and Braeckman 2007; Vogt, Efferson, and Fehr 2013; Yamagishi 2003), who reciprocates trust in a trust game (Bonnefon, Hopfensitz, and De Neys, 2013; Centorrino, Djemai, Hopfensitz, Milinski and Seabright, 2011; De Neys, Hopfensitz and Bonnefon, 2013; Efferson and Vogt, 2013; Stirrat and Perrett, 2010), who tries to exploit private information in a bargaining game (Ockenfels and Selten, 2000), who gives positive amounts in a dictator game (Fetchenhauer, Groothuis and Pradel, 2010), and who offers high amounts in an UG (Jaschke, Primes and Koppensteiner, 2013). The reported accuracy rates are typically modest but significant. In trust games, Eckel and Petrie (2011) find that people are willing to pay to see a photo of the other player. Moreover, first movers are willing to pay more than second movers, which suggests that people believe that they can predict the behavior of others from their photos.

Our study differs from the above-mentioned papers in several respects. First, we focus on the role of anger as a commitment device. More than any of the other games listed above, the UG is well-suited to address this issue. It captures the essence of interpersonal commitment problems, and, because of its sequential nature, there is no uncertainty for the responder about the other player's strategy. The experienced emotions are also likely to be different in the other games, which mostly focus on cooperative behavior. Identifying people who are prone to getting angry is likely to be a quite different task than identifying cooperative people.[9] Furthermore, we directly measure anger expressed by responders who receive a low offer, and relate this to their behavior. Finally, we measure and identify facial correlates of anger that are *ex ante* visible to observers, that is, before the actual decision task is

[9] Studies looking at within-subject correlation across different games find that responders who reject in an UG are not necessarily the same people as those who are conditionally cooperative in a sequential prisoner's dilemma or trust game (Blanco, Engelmann and Normann, 2011; Yamagishi and Horita, 2012).

even known to the individuals being evaluated. To register these facial correlates, we make use of face-reading software that, in contrast to other protocols for measuring emotions, such as the elicitation of self-reports, yields more objective physiological measures of emotional states in real time.

There are two other papers that we are aware of that employ photos of responders to study behavior in an UG. Reed et al. (2014) study whether proposers in an UG adjust their offers depending on a responder's demand made prior to the game and associated emotional expression. They find that high demands that are accompanied by an angry expression lead to increased offers by proposers. However, the responders in this study were actors, and never actually made a real decision. Thus, while the study shows that looking angry can be effective in getting higher offers, it does not establish that anger is a *credible* signal of rejection, nor whether people can distinguish credible from non-credible signals. Jaschke et al. (2013) study whether showing photos of responders whose minimal acceptable offer (MAO) is elicited has an effect on the amount that proposers offer. They find that trustworthy-looking responders receive higher offers, although perceived trustworthiness is not correlated with the MAO's. They also find that dominant-looking responders have lower MAO's, but this is not 'recognized' by the proposers, as perceived dominance does not have a significant effect on the offered amounts. Neither perceived trustworthiness nor perceived dominance thus seem to be credible and observable at the same time.

The remainder of the paper is organized in the following manner. Section 2 describes the experimental design and procedures. Section 3 presents the results and Section 4 describes additional treatments that we conducted. Section 5 provides a concluding discussion.

2. Experimental design and procedures

The experiment consisted of two phases, conducted at different universities and on different dates. In the first phase, conducted at the CentERlab at Tilburg University, responders were photographed before being paired with proposers to play an UG. In the second phase, conducted at the CREED lab at the University of Amsterdam, we showed photos of responders who received a low offer in the UG played in the first phase to an independent set of observers, and asked them to identify which of two responders rejected the offer. Conducting the two phases at different universities in different cities made it very unlikely that any observer knew any of the responders whom they were evaluating. Details on the procedures and instructions are included in Appendix A1.

Responders were also videotaped just after being photographed, and these videos were similarly shown to different observers tasked with identifying the rejecter. In the current section and Section 3 we focus on the photo treatments. Section 4.1 discusses the video treatments.

2.1 Photos and UG at Tilburg University

We invited proposers and responders to come to the laboratory separately. In order to reduce the variability in facial appearances, we recruited responders from the pool of native-language-speaking subjects. Proposers were recruited from the general subject pool at Tilburg University, which has a substantial share of foreign students. The total number of participants was 262 (131 proposers and 131 responders). The experiment was programmed in z-Tree (Fischbacher, 2007).

Responders arrived 30 minutes before proposers and at a different entrance. *Before* the instructions for the UG were handed out, responders were photographed. Hence, they were not aware that they would participate in an UG. Responders were asked to maintain a neutral expression and to look straight into the camera. We took care that each responder was photographed in exactly the same way. Responders received €4 for participating in the photo shoots.

The decision-making part of the experiment had two parts, and subjects were informed that one of the parts would be randomly drawn for payment. In the first part, each proposer was matched with a responder and they played the following 'hot' UG. Proposers were asked to choose one of the two following allocations of €9: (A) €7 for herself and €2 for the responder, or (B) €4 for herself and €5 for the responder.[10] The proposal was communicated to the matched responder, who was aware that the responder could only select one of these two possible allocations. If the responder accepted the proposal, the money was divided as proposed. If the responder rejected the proposal, then neither player earned any money. At the end of the first part, each proposer was informed about whether her offer was accepted or rejected.

In order to track how angry responders become when they receive the low offer in the hot UG, we videotaped some of them during their decision-making, using webcams. We videotaped 57 out of the 131 responders. Taping a subset of responders enabled us to test whether the videotaping *per se* influences the rejection decision.[11]

In the second part of this phase of the experiment, subjects kept their roles, but were rematched with another subject. They played a 'cold' UG, using the strategy method. Proposers were asked to propose a division of €9 between themselves and a responder in any multiple of €0.5. We elicited the minimum acceptable offers (MAO's) from responders by asking them to indicate, for each possible offer, whether they would accept it. If the actual proposed offer was smaller than the MAO elicited from the matched responder, none of the players earned money. Otherwise, the money was divided as proposed.

There were 14 sessions in total. Most had 20 participants, but due to variable show-up rates there were two sessions with 18 participants, one session with 16, and one session with 10 participants. The experimental sessions ended with a questionnaire, included in order to collect some background

[10] In order to find a pair of payoff vectors that would result in relatively many low offers and rejections, so that our sample of rejecters in the main study would be sufficiently large, we ran a pilot study at the University of Amsterdam. In this pilot, we varied the two options available to proposers in the UG systematically.

[11] Videotaping responders did not significantly affect the decision to reject a low offer. 35 percent of those videotaped rejected the low offer, against 22 percent for those that were not videotaped ($\chi^2(1) = 1.296$, $p = 0.255$, $N = 69$). It also did not affect MAO's in the cold UG (mean MAOs are 2.4 for those videotaped and 2.2 for those not videotaped, $p = 0.455$, two-sided MWU test, $N = 131$).

information about participants. Sessions lasted about 60 minutes for responders and 30 minutes for proposers. Participants earned between €3 and €12.

2.2 Behavior in the UG

We first briefly report the behavioral results of the first phase of the experiment, since this will help describing the second phase of the experiment. In the hot UG, 62 out of the 131 proposers made a high offer (4,5) and 69 proposers made a low offer (7,2). The high offer was accepted by all responders. The low offer was accepted by 49 of the responders, and rejected by the other 20.

2.3 Observers at the University of Amsterdam

Observers were shown pairs of photos of responders who later received the (7,2) proposal in the hot UG played in Tilburg. Each of the pairs consisted of one responder who rejected and another who accepted the proposal. Observers were asked to identify which of the two responders rejected their offer.[12] Since there were 20 responders in our sample who rejected the (7,2) proposal, observers were shown 20 randomly formed pairs of accepters and rejecters. No responder was displayed more than once to any observer. In total, 128 observers participated in the two photo tasks. This part of the experiment was programmed using php and MySQL.

We implemented two different treatments. In one condition (photo 1s), observers saw photos of the responders with a neutral expression for one second only. In the second condition (photo 5s), the responders' photos were displayed for five seconds.[13] In the photo 5s task, observers were asked to state the confidence they had in the correctness of their prediction by moving a slider between 'completely unsure' and 'completely sure'.

Observers did not receive any feedback about their performance or earnings at any time during the experiment. At the end of the experiment, they were informed about their final earnings. Subjects knew that two trials were going to be randomly selected for payment, and that for each rejecter in the selected trials they identified correctly, they would receive €5.

Besides the guessing task, observers completed three other tasks, including guessing the behavior of responders in the cold UG, a 'reading the mind in the eyes' test (Baron-Cohen et al., 2001), and playing an UG themselves (see Appendix A1 for details). They finished with a post-experimental questionnaire that gathered some demographic data after which they received their payments. Sessions lasted about 60 minutes and participants earned between €5 and €32 with an average of €18.

[12] An advantage of using pairs consisting of one accepter and one rejecter is that observers know the base rate of rejection (50 percent), so that we have no confound of subjects having different priors on the base rate and differences in how they update their beliefs. This procedure is similar as the one that Todorov et al. (2005) use to predict the outcomes of U.S. congressional outcomes on the basis of observers' judgments of the competence of paired political candidates.

[13] In fact, we implemented more treatments. Section 4 discusses the other treatments.

2.4 Facial measures

We measured the emotional content of facial expressions using the Noldus FaceReader 6 software package. The software has been trained to classify expressions in photos and videos on the basis of their conformity to happiness, sadness, anger, surprise, fear, disgust and neutrality. The classification is done by training an artificial neural network with over 10,000 annotated images. It uses distances between 538 points on the face as well as the texture (registering muscle tightness) at each of the points as input. Overall, the software has an accuracy level of 90 percent when it rates the intended expressions of trained test persons (Bijlstra and Dotsch, 2011). It classifies human expressions as well as trained human observers do (Kuderna-Iulian et al., 2009; Terzis, Moridis and Economides, 2010), correlating highly with self-reported emotions (Den Uyl and Van Kuilenburg, 2005) and those described by observers. D'Arcey (2013) validates the software by comparing it to facial electromyography, which measures the activity of specific muscles.

We also measure other facial features that have been shown to be correlated with aggression or (anti-)social behavior. Doing so enables us to explore the correlates of rejection behavior of responders, as well as whether observers' predictions reflect these relationships. The facial measures we use are (1) facial width-to-height ratio (fWHR), (2) facial masculinity (fM), (3) and facial asymmetry (fA). In human males, a small but significant relationship exists between fWHR and aggressive tendencies (see Geniole et al., 2015; Haselhuhn, Ormison and Wong, 2015 for meta-studies).[14] Lefevre et al. (2013) suggest a potential mechanism driving this effect, which is that men with high fWHR have high (baseline and reactive) testosterone levels. Burnham (2007) shows that high-testosterone men reject unfair offers in UG's more frequently than low-testosterone men. Furthermore, high fM has also been associated with high (baseline) testosterone levels (Penton-Voak and Chen, 2004). Finally, fA is a proxy for fluctuating asymmetry, which is generally accepted as a marker of developmental instability.[15] High fA is associated with poorer health, problems at birth, psychological maladaptation, lower reproductive success, and lower attractiveness (Van Dongen and Gangestad, 2011). It has been found to be correlated with reacting aggressively under provocation (Lalumière et al., 2001; Benderlioglu et al., 2004). However, asymmetric people also give more in UG's (Zaatari and Trivers, 2007), cooperate more frequently in prisoner's dilemmas (Sanchez-Pages and Turiegano, 2010), and perceive themselves as less aggressive and more pro-social than symmetric people (Furlow et al., 1998; Holtzman et al., 2011).

The three facial measures are constructed by marking 19 different points on each responder's face. We use software that computes the distances between these points. A research assistant, who was unfamiliar with the purpose of the study, measured all of the faces. In Appendix A3, we describe how

[14] See Carré, McCormick and Mondloch (2009) and Carré and McCormick (2008) for studies using data from ice hockey players to show that fWHR is associated with aggression. In capuchin monkeys, fWHR is correlated with assertiveness (Wilson et al., 2014).

[15] Fluctuating asymmetry refers to features that are on average symmetric in the entire population (Van Dongen and Gangestad, 2011), rather than, e.g. left- or right-handedness which is not symmetrically distributed over the population.

we constructed these measures in more detail. All three measures are standardized to have a mean of zero and a standard deviation equal to one. fM is standardized within gender.

We also gathered data on other, more subjective, facial cues. These cues are perceived attractiveness, perceived intelligence, perceived weight, and perceived masculinity. We included attractiveness and intelligence because some of the observers spontaneously mentioned these in the post-experimental questionnaires as cues that they use.[16,17] We included perceived weight to use as a control for fWHR, since fWHR is positively correlated with weight. We included perceived masculinity in order to obtain a more subjective measure of facial masculinity, since it may be the subjective perception rather than the objective measure that influences observers' beliefs about whether an individual is likely to reject an offer.

To obtain these measures, an independent cohort of 32 subjects rated all 131 responders of phase 1 in the laboratory at the University of Amsterdam on one of the four subjective measures. These subjects were highly unlikely to know any of the individuals whose photos they were rating and they are recruited from the same population of participants that participated in the observer tasks of phase 2. Four men and four women rated each responder on a 7-point Likert-scale on one of the four dimensions. The pictures were presented to them in random order, but they evaluated either all of the men, or all of the women, first. Sessions lasted around 30 minutes and subjects received a fixed payment of €10. For all four measures, we took the mean rating and standardized it to a mean of 0 and a standard deviation of 1.[18] Perceived masculinity is standardized within gender. Table A1 in the Appendix summarizes the main (non-standardized) background characteristics of the proposers, responders, and observers.

3. Results

3.1 Accuracy of observers' judgments

Our main measure of accuracy is the percentage of times that the observers correctly identified the rejecter within a pair of responders. If observers are unable to recognize rejecters, this measure equals 50 percent (the percentage that can be expected from random guessing). Each responder is observed multiple times by different observers. We take the mean for each responder as the independent unit of observation.[19]

[16] Arguments in both directions are made in the questionnaire. Other cues frequently reported by observers are anger, other emotions, and gender (typically guessing that men are more likely to reject). Moreover, many observers indicate that they follow their gut feeling or state that the task is impossible.

[17] Solnick and Schweitzer (1999) find that attractive people reject at the same rate as others in ultimatum games.

[18] The inter-rater consistency (Cronbachs' alpha) is $\alpha = 0.76$ for attractiveness, $\alpha = 0.74$ for intelligence, $\alpha = 0.78$ for masculinity and $\alpha = 0.92$ for weight. A value of α above 0.7 is usually taken as an acceptable degree of consistency.

[19] Taking each observer as the unit of observation gives similar results.

Table 1: Accuracy of judgments

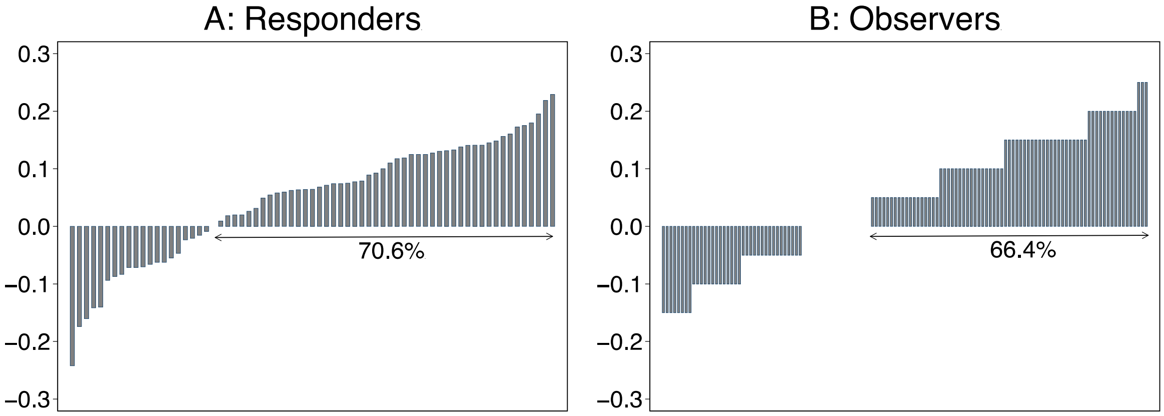
	(1) Accuracy in % (*p*-values)		(2) Responders (%) correctly judged over 50% of instances	(3) Observers (%) with over 50% correct judgments	(4) *N* observers
5s Photos	54.8	(0.001)	70.8	69.7	74
1s Photos	54.9	(0.002)	61.8	61.4	54
Pooled	54.8	(0.000)	70.6	66.4	128

Notes: Accuracy rates in (1) and *p*-values from two-sided Wilcoxon signed-ranks test (in parentheses) are based on the mean accuracy of each responder as the unit of observation (N = 69). Percentages in (2) and (3) are computed by dividing the number of responders or observers with more than 50 percent correct judgments by the total number of responders or observers with strictly more or less correct judgments than 50 percent.

The achieved accuracy rates for the different tasks are reported in column (1) of Table 1. The accuracy is nearly 55 percent for both photo tasks, meaning that observers do about 10% better than chance. In both cases, we reject the hypothesis that observers are guessing randomly ($p < 0.01$ for both tasks, two-sided Wilcoxon signed-ranks test, $N = 69$). This result is not driven by a few responders that are easy to identify, nor by a few observers that are particularly good at identifying rejecters. To illustrate this, we report, in column (2) of Table 1, the percentage of responders that are judged correctly by more than 50 percent of the observers.[20] 70.8 percent of the responders are judged correctly by a majority of observers in the 5-second photo task. For the 1-second photo task, the corresponding number is 61.8 percent. This can also be seen in Panel A of Figure 1, that shows the percentage of correct judgments for each responder in terms of deviations from 50 percent chance levels (both photo tasks pooled). Column (3) of Table 1 reports the percentage of observers that judge the majority of responders correctly. 69.7 percent of the observers correctly judge the majority of responders in the 5-second photo task, and 61.4 percent do so in the 1-second photo task. Panel B of Figure 1 shows the percentage of correct judgments by each observer.

[20] To be precise, the reported percentage is the percentage of all responders that are judged correctly by a majority of observers (excluding responders that were judged correctly exactly 50 percent of the time).

Figure 1: Percentage of responders judged correctly and percentage of observers making correct judgments as compared to random guessing



Notes: The figure shows deviations from 50 percent correct judgments for each of the responders (panel A) and observers (panel B). Each bar represents one responder (panel A) or one observer (panel B).

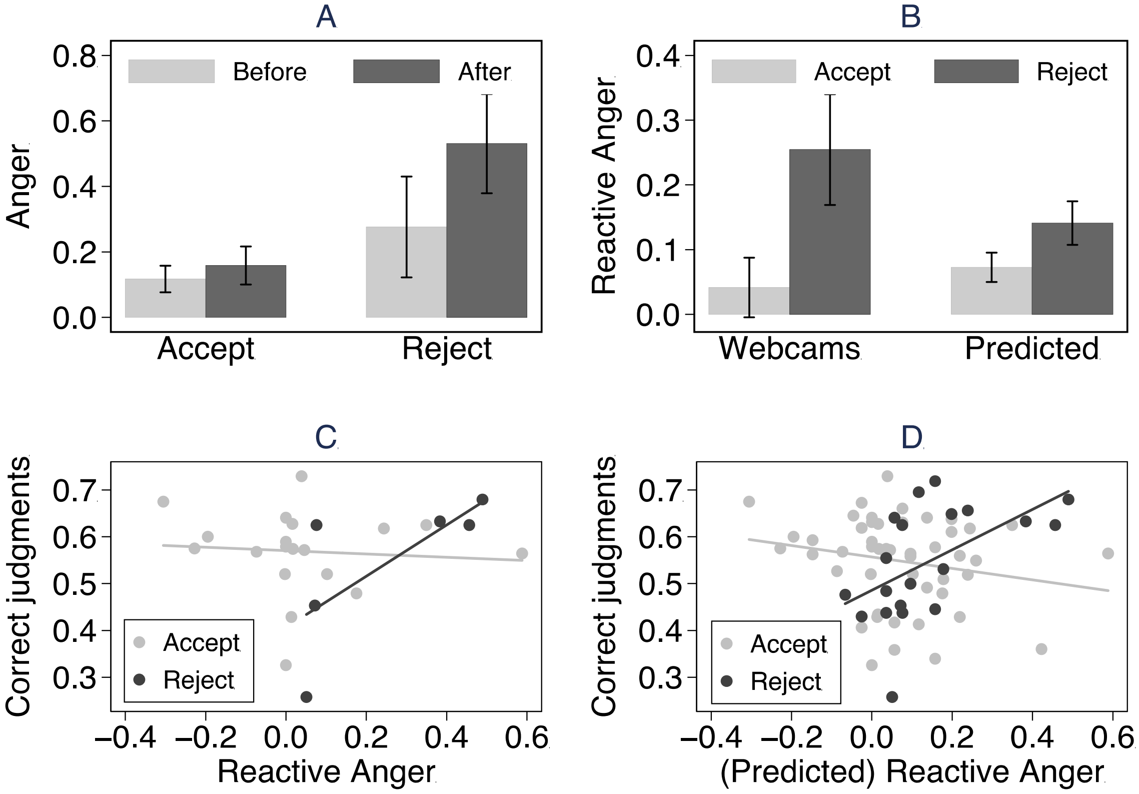
3.2 Identification of the angry button

3.2.1 Reactive anger

So far we presented evidence that observers are able to identify rejecters at above chance levels. The next step is to find out if observers are able to do this by recognizing who has an 'angry button,' i.e., recognizing responders who get angry and reject. If so, higher accuracy rates should be expected for pairs in which the rejecter becomes angrier after receiving the low offer. Observers may of course take other dimensions into account as well; maybe they can infer less rage-driven fairness ideals that lead responders to reject. Because we have a clear anger-based hypothesis, we focus on the anger pathway, but this is not to say that other pathways do not play a role.

We first need to determine which responders get angry at the time that they receive a low offer. For the subsample of responders who were videotaped during the experiment, we can measure this 'reactive anger' directly using the face-reading software. For this measure, we compare the mean anger expressed in the 5 seconds before and after observing the low offer. Some participants looked away from the camera, but the software successfully captured the facial expressions of 25 of the responders. Panel A of Figure 2 shows the mean anger expressed by accepters and rejecters right before and after observing the low offer. Rejecters express more anger than accepters before and after observing the low offer. Anger expressed before observing the offer is not significantly higher for rejecters than accepters, but the difference in anger after observing the low offer is significant (Z = 1.020, p = 0.308 and Z = 2.356, p = 0.019 respectively, two-sided MWU tests). Moreover, the increase in anger is significant for rejecters but not for accepters (Z = 2.201, p = 0.028 and Z = 1.049, p = 0.294 respectively, two-sided Wilcoxon signed-ranks tests). Finally, and most importantly, as illustrated in Panel B of Figure 2 (see 'webcams' bar), the increase in anger is larger for rejecters than for accepters (Z = 2.483, p = 0.013, two-sided MWU test). We will refer to this increase in anger as 'reactive anger'.

Figure 2: Reactive anger and accuracy of judgments



Notes: Panel A shows the mean anger in the 5 seconds before and after observing the low offer. Panel B shows the mean reactive anger (webcams) and predicted reactive anger for accepters and rejecters. (Predicted) reactive anger is defined as the difference in (predicted) anger before and after observing the low offer (see Panel A). Error bars indicate standard errors of the mean. Panel C illustrates the correlation between reactive anger and the accuracy of judgments, both for accepters (light grey) and rejecters (dark grey). The fraction of correct judgments is taken over all trials, across both photo tasks. Straight lines come from OLS regressions of the fraction of correct judgments on reactive anger and a constant. Along the same lines, Panel D illustrates the correlation between (predicted) reactive anger and the accuracy of judgments. Panels A and C, and the 'webcams' data of Panel B, are based on the data of the responders for whom we have direct measurements of reactive anger. Panel D, and the 'predicted' data of Panel B, are based on all responders; for those responders for whom we do not have direct measurements we used the model specified in Table 2 to predict reactive anger.

We then study whether the observers' judgments are more accurate for rejecters with a higher level of reactive anger. The relative frequency of the observers' correct judgments as a function of the level of reactive anger can be seen in Panel C of Figure 2. The figure shows that observers are better able to identify rejecters who got angry at the time they received a low offer. In contrast, observers are not able to detect the behavior of accepters. Reactive anger and correct judgments are strongly and positively correlated for rejecters (Spearman rank correlation, ρ = 0.899, p = 0.015, N = 6), but not for accepters (Spearman rank correlation, ρ = -0.081, p = 0.742, N = 19).

Table 2: Signals associated with reactive anger

Dep. var.: Reactive anger		
Perceived intelligence	-.118	(.049)**
Constant	.095	(.040)**
Observations	25	
R^2	.199	
Adjusted R^2	.165	

Notes: The table reports results from an OLS regression. Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

These results indicate that observers are able to identify who is likely to reject a low offer out of anger. Recall that observers never saw the actual reactive anger: they are able to recognize who gets angry based on images that were recorded before the emotion was triggered. The better-than-chance accuracy thus seems to be driven by a particular subset of responders: those who have an angry button. The propensity to anger is therefore a relevant dimension to determine who will accept or reject.[21]

3.2.2 Predicted reactive anger and angry buttons

Ultimately, we would like to determine whether every responder, not only those who were recorded on the webcam, has an angry button or not. To do so, we first examine which facial traits are associated with their reactive anger. The facial traits that we consider are fWHR, fM, fA, perceived attractiveness, perceived intelligence, perceived weight, perceived masculinity, and gender. We also include a measure of 'baseline anger', which is the subject's expressed level of anger on the photo. We use a parsimonious approach to explore which facial cues predict reactive anger. We included cues one by one on the basis of which variable has the lowest *p*-value, and we continued doing so until none of the variables would enter with a *p*-value below 0.10 (e.g., Heij et al., 2004). The resulting specification is shown in Table 2. The only variable that turned out to be significantly related to reactive anger is perceived intelligence: those with lower perceived intelligence anger more after receiving a low offer.

[21] Instead of predicting who would reject a low offer, one could also ask observers who would get angrier after receiving a low offer. We did not use this task in our experiment, because the main question was firstly whether people could predict who would reject a low offer, and secondly whether rejections are associated with anger. We focus on the anger pathway by relying on face-reading software, see Sections 3.2 and 3.3 for detailed results.

Table 3: Accuracy of judgments for different responder types (in %)

Rejecters →	No angry button	Angry button	
↓ Accepters		All	Big
	(N = 9)	(N = 11)	(N = 6)
Not angry (N = 26)	50.4	59.5***	63.5***
Angry accepters (N = 23)	48.1	60.0***	61.5***
All accepters (N=49)	49.5	59.8***	62.6***

Notes: The table shows accuracy rates for pairs, depending on the types of the matched responders. (Big) angry buttons refer to rejecters who have a level of (predicted) reactive anger in the top 50 percent (25 percent), angry accepters are accepters who have a level of RA in the top 50 percent, other accepters and rejecters are responders who have a level of (predicted) reactive anger in the bottom 50 percent. Stars indicate significance levels from two-sided Wilcoxon signed-ranks tests taking the mean accuracy of each responder as the unit of observation: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

We then use these regression results to construct a measure of reactive anger for those 44 out of our 69 responders for whom we don't have direct measurements. We label this measure 'predicted reactive anger.' Panel D of Figure 2 shows the relative frequency of the observers' correct judgments as a function of (predicted) reactive anger. In line with the above-reported results on reactive anger, (predicted) reactive anger and correct judgments are strongly and positively correlated for rejecters (Spearman rank correlation, ρ = 0.620, p = 0.004), but not for accepters (Spearman rank correlation, ρ = -0.210, p = 0.147). In addition, as can be seen in Panel B of Figure 2, rejecters' overall (predicted) reactive anger is higher than that of accepters, but the difference is not significant (Z = 1.542, p = 0.123, two-sided MWU test).

We further illustrate the finding that responders with an angry button are frequently judged correctly in Table 3, where we present accuracy levels conditional on the composition of the pairs. Based on their (predicted) reactive anger, we classify responders as having an angry button if they have an above median level of (predicted) reactive anger and reject the low offer, and to have a 'big' angry button if they reject and have a level of (predicted) reactive anger in the top quartile. If responders have an above median level of (predicted) reactive anger but nevertheless accept the low offer, we label them as 'angry accepters.'

Table 3 confirms that rejecters with an angry button are detected particularly well. Pairs that involve a responder with a big angry button are judged accurately in around 62.6 percent of the cases. Overall, observers thus do about 25% better than chance judging pairs where a rejecter with a big angry button is present. From a different angle, 83.3 percent of the responders are judged correctly by a majority of observers when we only consider pairs where a rejecter with a big angry button is present (see Table A2 in the Appendix).

Table 3 highlights some other interesting findings. A first one is that angry accepters and rejecters can be distinguished quite well: the accuracy rate for pairs where the rejecter has an angry button and the accepter has above-the-median (predicted) reactive anger is 60.0 percent (significantly higher than 50 percent). Second, Table 3 also shows that when rejecters do not have an angry button, the accuracy rates drop to around 50 percent.

3.3 Do observers use cues associated with rejections?

We examine in detail which cues predict the responders' decision to reject, and whether observers use those cues in making their judgments. To do so, we first regress the responders' decision to reject on a range of facial cues. These cues include the anger expressed on the photos, fWHR, fM, and fA, and perceived attractiveness, intelligence, weight, and masculinity.[22] Column (1) of Table 4 reports the results. Responders with a more asymmetric face are more likely to reject. A one standard deviation increase in fA increases the likelihood of rejection by 19.6 percentage points. Responders who belong to the top half in terms of fA reject 18.2 percentage points more often than responders in the bottom half. This finding seems to be robust. In another sample we used for the pilot study, the difference in rejection rates between the 14 responders in the top half of fA and the 14 responders in the bottom half of that sample equals 20.7 percentage points.[23] In addition, fA is positively correlated with reactive anger (Spearman rank correlation, $\rho = 0.329$, $p = 0.109$, $N = 25$). Most of the other coefficients are small and insignificant.

Do observers use the relevant cues? Column (2) of Panel A in Table 4 shows the results of regressing the proportion of times that a responder is judged as a rejecter on the above-mentioned facial cues. In addition to the explanatory variables used in column (1), we include the dummy variable 'rejects,' indicating whether or not the responder who was identified as the rejecter actually did reject. This variable is supposed to pick up all cues, unobservable to us, that observers used to make correct judgments. Specifically, a positive sign of the associated coefficient would imply that observers have used cues associated with rejection behavior that are not included in our regression. In Panel B of the table, we also report the coefficient of the reject dummy in regressions that have no other explanatory variables. This coefficient captures the extent to which observers have done better than chance.

Overall, observers appear to correctly ignore most of the cues that are uncorrelated with the actual decision to reject (column (2)). They do use perceived masculinity as a cue, even though we find no evidence that perceived masculinity is statistically predictive of rejecting. Most notably, observers rely on fA as a cue, which is indeed the best predictor of rejecting.

[22] We do not consider ethnicity as a cue, since all responders are Dutch speaking, and only 4 out of 69 responders appear to be from ethnic minorities upon visual inspection. Moreover, we do not find that domestic observers are better at making predictions than foreign observers.

[23] Zaatari and Trivers (2007) do not find a significant correlation between fA and rejection in the UG but they have a sample of only 15 rejecters.

Table 4: Cues associated with rejection

	Panel A	
	(1)	(2)
Dep. var.:	Responder rejects	Responder identified as rejecter by observers
Baseline anger	-.014 (.060)	-.009 (.013)
fA (facial asymmetry)	.196 (.062)***	.040 (.015)***
fWHR (width-to-height ratio)	-.052 (.068)	-.005 (.015)
fM (facial masculinity)	.047 (.057)	-.016 (.012)
Perceived attractiveness	.067 (.065)	-.000 (.014)
Perceived intelligence	-.116 (.073)	-.029 (.016)*
Perceived weight	-.122 (.078)	-.001 (.017)
Perceived masculinity	.053 (.063)	.028 (.014)**
Female	.068 (.123)	.006 (.027)
Horizontal head orient.	-.069 (.048)	-.018 (.011)*
Vertical head orient.	.008 (.014)	-.002 (.003)
Rejects		.059 (.029)**
Constant	.359 (.106)***	.476 (.025)***
Observations	69	69
Adj. R^2	.069	.256
R^2	.219	.387
	Panel B	
Rejects		.095 (.027)***
Observations		69
Adj. R^2		.140
R^2		.152

Notes: The table reports results from OLS regressions. The dependent variable in (1) is the responder's choice to reject the unfair offer in the hot UG, and in (2) the proportion of times that a responder was identified as the rejecter by observers. All independent variables are normalized to have mean zero and a standard deviation of one, except for the dummy variables and the head orientation controls. Standard errors in parentheses.*** p < 0.01, ** p < 0.05, * p < 0.1

It is interesting to examine the coefficients of the reject dummy. Without any further controls (Panel B), it is significant: observers are 9.5 percentage points more likely to select a responder if that responder actually rejected the low offer. This is consistent with our earlier analysis of accuracy. After adding controls (Panel A), the difference gets smaller (0.059) for the photo tasks and is significant. It seems therefore that observers use some additional cues that can explain the accuracy of their guesses. Of course, it is conceivable that observers used the same cues in a different manner than modeled in our specification. Instead of looking at absolute levels, they may make their judgments based on the *differences* between the paired respondents, as they see them side-by-side. In Table A3 of the Appendix we present estimations from a model specification based on differences (column (1)). While some of the explanatory variables become significant, our main conclusions are unaffected. With that specification, there is no evidence either that observers use any additional cues beyond any of our control variables.

It could be that observers use different cues for females and males. In Table A4 of the Appendix we report regressions by gender. Facial asymmetry is used for both females and males, but the effect is only significant (at the 10% level) for males. Facial asymmetry is also associated with rejection for both females and males, but again the effect is only significant for males. For females, fWHR turns out to be negatively related to rejection, and perceived intelligence is negatively associated with rejection. Observers do not use fWHR as a cue for females but do use perceived intelligence, although none of the coefficients are significant. For males, we don’t find a relation between fWHR and rejection, despite the fact that fWHR has been associated with threat behavior for men and not for women (Geniole et al. 2015).

4. Other treatments

We also ran treatments where observers judged videos, rather than photos, of the responders, as well as treatments where observers predicted responder behavior in the cold UG that was played after the hot UG at Tilburg University. We briefly discuss the main findings in these data sets in this section. Section A4 of the Appendix provides detailed information about the procedures and results of the video treatments. Section A5 does so for the treatments related to the cold UG.

4.1 Video treatments

We have videos of responders in a neutral state as well as videos of responders with an angry facial expression. For both types of videos it was difficult for observers to detect responders who reject low offers in the hot UG. The overall accuracy rate for the video tasks is 51.7 percent, and not significantly different from 50 percent (p = 0.221, N = 156). As in the photos tasks, however, the accuracy rate increases significantly above chance in pairs where the rejecter gets relatively angry; for responders with an angry button, the accuracy rate is 55.5 percent (p = 0.001, N = 156) When we pool the data from both the photo and the video tasks, the overall accuracy is 53.1 percent, and significantly better than chance (p = 0.002, N = 284).

One reason why it is more difficult for observers to detect rejecters in videos than in photos may be that videos contain ‘too much’ information. For example, Bonnefon et al. (2013) find that pictures that include hair and clothing impair trustworthiness detection, compared to pictures including only facial features. More information, or longer exposure to it, may spur people to replace intuitive decision-making by conscious decision-making, which may impede judgments. Another, potentially more plausible, reason may be that the videos were too unnatural to responders, and therefore contain a substantial amount of noise that distracted observers. In the neutral-state videos, responders read an unrelated text. In the angry-state videos responders were asked to imitate anger (and other emotions), and it could be that acting out an emotion as it would be naturally experienced is just too difficult for typical subjects. We conclude that whether additional information improves or impedes judgments is an interesting question for future research.

4.2 Cold UG treatments

The accuracy rate for the cold UG is 50.2 percent for the photo tasks (p = 0.900, N = 128) and 50.9 percent for the video tasks (p = 0.309, N = 132). Thus, observers do not perform better than they would if they guessed randomly. Combining results based on the hot and the cold UG suggests that it is the emotional response that is correctly anticipated by observers. Arguably, it is likely that emotions have much less of an influence on rejections under cold than under hot decision-making. For example, Jaschke et al. (2013) find that behavior of responders in a strategy method UG is not anticipated by proposers. However, there are other differences between the two games in how behavior was elicited than merely the 'hotness' of the response. In particular, in the cold UG we elicited the minimal acceptable offer by asking responders to indicate for each possible offer from €0.5 to €9 whether they would accept it, whereas the hot UG was a binary UG. Therefore, we leave it for future research to establish whether hotness in strategic interactions on its own has an influence on predictability of people.

5. Concluding discussion

In our experiment, we find considerable support for the hypothesis by Frank (1987, 1988) that anger serves as a credible commitment device. In particular, we find that observers who are shown pairs of photos of a rejecter and an accepter of a low offer taken before they played an ultimatum game (UG) detect the rejecter 10% more frequently than would be the outcome under random guessing. We also find that responders who reject a low offer express more anger when they see the low offer than those who accept it. These results provide support for both observability and credibility, which are the two components of Frank's hypothesis.

The ability to observe the preferences of others can have large consequences for the equilibrium configuration of preferences, even if the level of observability is low (Dekel, Ely and Yilankaya, 2007; Güth and Yaari, 1992). For example, Nowak et al. (2000) illustrate, in an evolutionary model, that a modest degree of observability can suffice to sustain fair behavior in a population. The model shows that if the proposer has some information about what the responder will accept in an UG, generous proposer offers will evolve in the population. While the model assumes that proposers acquire this information through access to some of the history of offers accepted by the responder, a similar mechanism is at work if the information consists of physical cues that are correlated with responder behavior.

Moreover, in a world where both accepters and rejecters of low offers coexist, it would be natural that (at least some) people have developed an *imperfect* ability to detect rejecters. On the one hand, if it were impossible to detect rejecters, both types would receive the same distribution of offers. Accepters would be able to mimic rejecters perfectly. The rejecters, who would receive a lower payoff than the accepters, would eventually be extinguished from the population. On the other hand, if proposers could predict perfectly beforehand who accepts and who rejects, accepters would get lower offers and thus lower payoffs than rejecters and their proportion would decline.

Observers are particularly able to identify rejecters when the rejecters are likely to become angry. Observers of pairs where the rejecter has an angry button, that is, where the rejecter angers more than the median responder after receiving a low offer, achieve an accuracy rate of 59.8 percent, which is 20% better than chance. For pairs that include a person with a 'big' angry button, who is among the 25 percent of angriest responders, the accuracy rate increases further to 62.6 percent, 25% better than chance. Even if a rejecter with an angry button is matched with an accepter who 'looks like' she has an angry button, the accuracy rate is equal to 60.0 percent.

In both our main and our pilot experiment, facial asymmetry is a cue that is surprisingly strongly correlated with the rejection of unfair offers. If observers had used the cue perfectly, by always selecting the more asymmetric person of the pair as the rejecter, they would have been correct 69 percent of the time. Although our observers do not use this cue perfectly, they do correctly sense that asymmetric responders tend to anger and reject more easily. This begs the question of why facial asymmetry plays the role that it does. Facial asymmetry is a proxy for fluctuating asymmetry, and it has been shown that fluctuating asymmetry can serve as a marker of developmental instability (Van Dongen and Gangestad, 2011). One possible pathway between asymmetry and rejection in the UG is that asymmetry signals developmental instability, that developmental instability makes people less able to control emotional impulses, and that this triggers hot rejections of stingy offers in the UG. In line with this pathway, Benderlioglu et al. (2004) find that asymmetric people respond more aggressively when being angered. In a similar vein, Lalumière et al. (2001) report that those who commit serious offences score higher on asymmetry measures than non-offenders. Interestingly, studies that rely on cold self-reported measures of anger find a *negative* association between anger and fluctuating asymmetry (Furlow et al., 1998; Manning and Wood, 1998). These results agree with our finding that asymmetry is positively associated with rejection of low proposals in the hot UG but not in the cold UG (see Section A5 in the Appendix for details). Although this evidence is suggestive, we think that we are still at an early phase of understanding the exact underlying biological mechanism linking asymmetry and behavior.

Because our experiment was designed to study the two above-mentioned components of the angry button hypothesis in the cleanest possible way, we made sure that decision-makers (observers) had no stake in the game itself. Showing proposers photos of responders and studying how their offers depend on the photos would not serve our purpose. The reason is that offers by proposers depend on their risk and social preferences, which may on their own be affected by certain characteristics of the responder. Our design rules out these influences. A relevant question, though, particularly for managerial applications, is whether having an angry button translates into getting higher offers in social interactions that are not anonymous. Possibly, being able to assess the intentions and reactions of others during the bargaining process can help closing deals and avoid negotiation failures, which may be one of the reasons why businesspeople find it important to meet prospective trading partners face-to-face (see Forbes, 2009). Ultimately, the question of how our findings translate into bargaining behavior is an empirical question. The answer depends on risk and social preferences of the bargainers, in particular on preferences of proposers towards responders. If proposers' risk or social preferences depend on certain (facial) features of responders that are at least to some extent correlated with their rejection behavior, the effect may be diminished or magnified. If not, there is no reason to expect that effects would be different in social interactions. Moreover, teams might strategically select whom to send to a bargaining interaction. If the person with the clearest 'angry button' is selected to do the bargaining, the effect on proposers' offers is likely to be magnified.

References

Andrade, E.B, & Ho, T.H. (2009). Gaming Emotions in Social Interactions. *Journal of Consumer Research, 36* (4), 539–552.

Apicella, C., Dreber, A., Campbell, B., Gray, P., Hoffman, M., & Little, A. (2008). Testosterone and financial risk preferences. *Evolution and Human Behavior*, *29*(6), 384–390.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test Revised Version: A Study with Normal Adults, and Adults with Asperger Syndrome or High-functioning Autism. *Journal of Child Psychology and Psychiatry*, *42*(2), 241–251.

Belot, M., Bhaskar, V., & van de Ven, J. (2012). Can observers predict trustworthiness? *The Review of Economics and Statistics*, *94*, 246–259.

Benderlioglu, Z., Sciulli, P. W., & Nelson, R. J. (2004). Fluctuating asymmetry predicts human reactive aggression. *American Journal of Human Biology*, *16*(4), 458–69.

Bijlstra, G., & Dotsch, R. (2011). FaceReader 4 emotion classification performance on images from the Radboud Faces Database. *Mimeo*.

Blanco, M., Engelmann, D., & Normann, H. T. (2011). A within-subject analysis of other-regarding preferences. *Games and Economic Behavior*, *72*(2), 321–338.

Bonnefon, J.-F., Hopfensitz, A., & De Neys, W. (2013). The modular nature of trustworthiness detection. *Journal of Experimental Psychology*, *142*(1), 143–50.

Brandts, J., & Charness, G. (2011). The strategy versus the direct-response method: a first survey of experimental comparisons. *Experimental Economics*, *14*(3), 375–398.

Brosig, J. (2002). Identifying cooperative behavior: some experimental results in a prisoner's dilemma game. *Journal of Economic Behavior & Organization Organization*, *47*, 275–290.

Burnham, T. C. (2007). High-testosterone men reject low ultimatum game offers. *Proceedings. Biological Sciences / The Royal Society*, *274*(1623), 2327–30.

Carré, J. M., & McCormick, C. M. (2008). In your face: facial metrics predict aggressive behaviour in the laboratory and in varsity and professional hockey players. *Proceedings. Biological Sciences / The Royal Society*, *275*(1651), 2651–6.

Carré, J. M., McCormick, C. M., & Mondloch, C. J. (2009). Facial structure is a reliable cue of aggressive behavior. *Psychological Science*, *20*(10), 1194–8.

Centorrino, S., Djemai, E., Hopfensitz, A., Milinski, M., & Seabright, P. (2011). Smiling is a Costly Signal of Cooperation Opportunities: Experimental Evidence from a Trust Game. *Mimeo*.

D'Arcey, J. T. (2013). Assessing the validity of FaceReader using facial EMG. *Mimeo*.

Dawes, R. M., McTavish, J., & Shaklee, H. (1977). Behavior, communication, and assumptions about other people's behavior in a commons dilemma situation. *Journal of Personality and Social Psychology*, *35*(1), 1–11.

De Neys, W., Hopfensitz, A., & Bonnefon, J.-F. (2013). Low second-to-fourth digit ratio predicts indiscriminate social suspicion, not improved trustworthiness detection. *Biology Letters*, *9*(2), 20130037.

Dekel, E., Ely, J. C., & Yilankaya, O. (2007). Evolution of Preferences. *The Review of Economic Studies*, *74*(3), 685–704.

Den Uyl, M. J., & Van Kuilenburg, H. (2005). The FaceReader: Online facial expression recognition. *Proceedings of Measuring Behavior*, *30*.

Efferson, C., & Vogt, S. (2013). Viewing men's faces does not lead to accurate predictions of trustworthiness. *Scientific Reports*, *3*, 1047.

Fetchenhauer, D., Groothuis, T., & Pradel, J. (2010). Not only states but traits — Humans can identify permanent altruistic dispositions in 20 s. *Evolution and Human Behavior*, *31*(2), 80–86.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*(2), 171–178.

Forbes. (2009). Business Meetings: The Case for Face-to-Face. *Forbes Insight*.

Frank, R. H. (1987). If Homo Economicus Could Choose His Own Utility Function, Would He Want One with a Conscience? *American Economic Review*, *77*(4), 593–604.

Frank, R. H. (1988). *Passions within reason: The strategic role of the emotions.* New York: W. W. Norton & Co.

Frank, R. H., Gilovich, T., & Regan, D. T. (1993). The evolution of one-shot cooperation: An experiment. *Ethology and Sociobiology*, *14*(4), 247–256.

Furlow, B., Gangestad, S. W., & Armijo-Prewitt, T. (1998). Developmental stability and human violence. *Proceedings. Biological Sciences / The Royal Society*, *265*(1390), 1–6.

Geniole, S. N., Denson, T. F., Dixson, B. J., Carré, J. M., McCormick, C. M. (2015). Evidence from meta-analyses of the facial width-to-height ratio as an evolved cue of threat. *PLoS ONE*, *10*(7).

Gospic, K., Mohlin, E., Fransson, P., Petrovic, P., Johannesson, M., & Ingvar, M. (2011). Limbic justice--amygdala involvement in immediate rejection in the Ultimatum Game. *PLoS Biology*, *9*(5), e1001054.

Gneezy, U. & Imas, A. (2014). Materazzi effect and the strategic use of anger in competitive interactions. *Proceedings of the National Academy of Sciences, 111*(4), 1334-1337.

Grimm, V., & Mengel, F. (2011). Let me sleep on it: Delay reduces rejection rates in ultimatum games. *Economics Letters*, *111*(2), 113–115.

Güth, W. (1995). An evolutionary approach to explaining cooperative behavior by reciprocal incentives. *International Journal of Game Theory*, *24*(4), 323–344.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, *3*(4), 367–388.

Güth, W., & Yaari, M. (1992). An evolutionary approach to explain reciprocal behavior in a simple strategic game. In U. Witt (Ed.), *Explaining Process and Change–Approaches to Evolutionary Economics* (pp. 23–34).

Haselhuhn, M.P., Ormiston, M. E, & Wong, E. M. (2015). Men's facial width-to-height ratio predicts aggression: a meta-analysis. *PLoS ONE, 10*(4).

Heij, C., De Boer, P., Franses, P. H., Kloek, T., & Van Dijk, H. K. (2004). *Econometric methods with applications in business and economics*. Oxford: Oxford University Press.

Holtzman, N. S., Augustine, A. a, & Senne, A. L. (2011). Are pro-social or socially aversive people more physically symmetrical? Symmetry in relation to over 200 personality variables. *Journal of Research in Personality*, *45*(6), 687–691.

Jaschke, J., Primes, G., & Koppensteiner, M. (2013). What you see is what you get? How facial dominance and trustworthiness affect expectation formation and decision-making in ultimatum bargaining. *Mimeo*.

Kovács-Bálint, Z., Bereczkei, T., & Hernádi, I. (2013). The telltale face: possible mechanisms behind defector and cooperator recognition revealed by emotional facial expression metrics. *British Journal of Psychology*, *104*(4), 563–76.

Kuderna-Iulian, B., van Kuilenburg, H., Eligio Xolocotzin, U., Den Uyl, M., Cremene, M., Hoszu, A., & Octavian, C. (2009). Evaluation of a System for Real-Time Valence Assessment of Spontaneous Facial Expressions. *Distributed Environments Adaptability, Semantics and Security Issues International Romanian-French Workshop, Cluj-Napoca*.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H.J., Hawk, S.T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. *Cognition & Emotion*, 24(8), 1377—1388.

Lalumière, M. L., Harris, G. T., & Rice, M. E. (2001). Psychopathy and developmental instability. *Evolution and Human Behavior*, *22*(2), 75–92.

Lefevre, C. E., Lewis, G. J., Perrett, D. I., & Penke, L. (2013). Telling facial metrics: facial width is associated with testosterone levels in men. *Evolution and Human Behavior*, *34*(4), 273–279.

Little, A. C., Jones, B. C., Waitt, C., Tiddeman, B. P., Feinberg, D. R., Perrett, D. I., Apicella, C. L., &

Marlowe, F. W. (2008). Symmetry is related to sexual dimorphism in faces: data across culture and species. *PLoS ONE*, *3*(5), e2106.

Lowyck, Luyten, Vandeneede, Verhaest, Vermote, & Peuskens. (2007). Adult Eyes Test. *Mimeo*.

Manning, J., & Wood, D. (1998). Fluctuating asymmetry and aggression in boys. *Human Nature*, *9*(1), 53–65.

Nesse, R. M. (ed) (2001). Evolution and the Capacity for Commitment New York: Russell Sage Foundation.

Nowak, M., Page, K., & Sigmund, K. (2000). Fairness Versus Reason in the Ultimatum Game. *Science*, *289*(5485), 1773–1775.

Ockenfels, A., & Selten, R. (2000). An Experiment on the Hypothesis of Involuntary Truth-Signalling in Bargaining. *Games and Economic Behavior*, *33*(1), 90–116.

Oechssler, J., Roider, A., & Schmitz, P. W. (2008). Cooling-Off in Negotiations - Does It Work? *Mimeo*.

Oosterbeek, H., Sloof, R., & van de Kuilen, G. (2004). Cultural Differences in Ultimatum Game Experiments: Evidence from a Meta-Analysis. *Experimental Economics*, *7*(2), 171–188.

Penton-Voak, I. S., & Chen, J. Y. (2004). High salivary testosterone is linked to masculine male facial appearance in humans. *Evolution and Human Behavior, 25*(4), 229–241.

Reed, L. I., DeScioli, P., & Pinker, S. A. (2014). The Commitment Function of Angry Facial Expressions. *Psychological Science*, 0956797614531027–. doi:10.1177/0956797614531027

Samuelson, L. (2001). Introduction to the Evolution of Preferences. *Journal of Economic Theory*, *97*(2), 225–230.

Sanchez-Pages, S., & Turiegano, E. (2010). Testosterone, facial symmetry and cooperation in the prisoners' dilemma. *Physiology & Behavior, 99*(3), 355–61.

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

Solnick, S., & Schweitzer, M. (1999). The Influence of Physical Attractiveness and Gender on Ultimatum Game Decisions. *Organizational Behavior and Human Decision Processes*, *79*(3), 199–215.

Stirrat, M., & Perrett, D. I. (2010). Valid facial cues to cooperation and trust: male facial width and trustworthiness. *Psychological Science, 21*(3), 349–54.

Sutter, M., Kocher, M., & Strauß, S. (2003). Bargaining under time pressure in an experimental ultimatum game. *Economics Letters*, *81*(3), 341–347.

Terzis, V., Moridis, C. N., & Economides, A. A. (2010). Measuring instant emotions during a self-assessment test. In *Proceedings of the 7th International Conference on Methods and Techniques in Behavioral Research - MB '10* (pp. 1–4). New York, New York, USA: ACM Press.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science*, *308*(5728), 1623–6.

Tognetti, A., Berticat, C., Raymond, M., & Faurie, C. (2013). Is cooperativeness readable in static facial features? An inter-cultural approach. *Evolution and Human Behavior*, *34*(6), 427–432.

Van Dongen, S., & Gangestad, S. W. (2011). Human fluctuating asymmetry in relation to health and quality: a meta-analysis. *Evolution and Human Behavior*, *32*(6), 380–398.

Verplaetse, J., Vanneste, S., & Braeckman, J. (2007). You can judge a book by its cover: the sequel.A kernel of truth in predictive cheating detection. *Evolution and Human Behavior*, *28*(4), 260–271.

Vogt, S., Efferson, C., & Fehr, E. (2013). Can we see inside? Predicting strategic behavior given limited information. *Evolution and Human Behavior*, *34*(4), 258–264.

Wilson, V., Lefevre, C. E., Morton, F. B., Brosnan, S. F., Paukner, A., & Bates, T. C. (2014). Personality and facial morphology: Links to assertiveness and neuroticism in capuchins. *Personality and Individual Differences*, *58*, 89–94.

Wout, M. van 't, Kahn, R. S., Sanfey, A. G., & Aleman, A. (2006). Affective state and decision-making in the Ultimatum Game. *Experimental Brain Research*, *169*(4), 564–8.

Xiao, E., & Houser, D. (2005). Emotion expression in human punishment behavior. *Proceedings of the National Academy of Sciences*, *102*(20), 7398–401.

Yamagishi, T. (2003). You can judge a book by its cover Evidence that cheaters may look different from cooperators. *Evolution and Human Behavior*, *24*(4), 290–301.

Yamagishi, T., & Horita, Y. (2012). Rejection of unfair offers in the ultimatum game is no evidence of strong reciprocity. *Proceedings of the National Academy of Sciences*, *109*(50), 20364–8.

Yamagishi, T., Horita, Y., Takagishi, H., Shinada, M., Tanida, S., & Cook, K. S. (2009). The private rejection of unfair offers. *Proceedings of the National Academy of Sciences*, *106*(28), 11520–11523.

Zaatari, D., & Trivers, R. (2007). Fluctuating asymmetry and behavior in the ultimatum game in Jamaica. *Evolution and Human Behavior*, *28*(4), 223–227.

Appendix (Online)

A1 Procedures and instructions

A1.1 Detailed procedures Tilburg University phase

The laboratory at Tilburg University has three different rooms arranged in a row. Each room has a separate external entrance and they are connected to each other directly by two doors. Responders and proposers were allocated to the two outer rooms, and the middle room was used to photograph the responders. The physical separation of responders from proposers prevented them from seeing each other during the photographing phase.

Upon arrival at the lab, responders were seated randomly. They were given a consent form stating that photos and videos of them would be taken, that these would be used exclusively for research purposes, and that these would not be shown to anyone who could be reasonably expected to know them (see Appendix A1.3). Participants were told that they could leave the experiment if they objected to any of the procedures, but no participant did so. Participants remained seated in their cubicles, and were called one by one to the middle room where an experimenter was waiting to take photos and videos. Responders received €4 for participating in this initial part of the experiment.

We first took a photo in which the participant showed his computer number. The purpose was to allow matching of the photo and videos with their decisions in the game that they played subsequently. We then took a close-up photo of each responder's face. This was the photo that we used in the photo tasks. After the photos had been taken, we videotaped the responders (see Section A4 in the Appendix for detailed procedures).

While photos and videos of responders were being taken, proposers arrived at the other room of the lab and were randomly seated. The door between the rooms where the proposers sat and where the shoots were taken was closed so that the proposers were unaware of the photography and filming that was occurring. After all responders were photographed and videotaped, instructions for the hot UG were handed out to all participants, and the doors connecting the three rooms in the lab were opened. This ensured that all subjects knew that they played the game with a human opponent, who was sitting in the other room. Instructions for the cold UG were handed out only after they had finished with the first part.

A1.2 Detailed procedures University of Amsterdam phase

Upon arrival at the lab, observers were seated randomly and handed out the instructions (both on-screen and printed). As soon as the photos appeared on the screen, observers could make a decision by clicking on one of the photos. After either 1 or 5 seconds (depending on the task) blue silhouettes of an unknown person replaced the photos. Observers could then still make their decision. Only after they made the decision could they proceed to the next trial. At the start of each trial, a three-second countdown animation was shown so that participants knew when the photos would appear.

For both tasks, rejecters were paired randomly with accepters in such way that a given observer never saw the same person twice. For each experimental session, 10 different random sequences of 20 pairs were generated from the 20 rejecters and 49 accepters in the sample. The positions of rejecters and accepters on the screen (left and right) were also randomly determined for each sequence.

A1.3 Pre-experimental consent form

One of the things that we shall do in this experiment is to take a photograph and a video of you.
The photograph and video will be used only for research purposes.
Your photo and video will **not** be used today and will not be shown to anyone else participating today.
We do ask your permission to use the photos and videos in other experiments. You will remain anonymous in these experiments (your photo and video will not be shown to anyone that we have reason to believe might know you). We will not share your background information with participants in other experiments.
You get 4 EUR for having your photo and video taken and agreeing with the procedure described above.
If you object to having your photo and video taken, you may leave the experiment now.

A1.4 Script for making photos and videos

The script below is translated from Dutch. The transcript in Dutch is available upon request.

Sequence of events:

1) Photo of the participant is taken where he/she visibly shows her table id
2) Neutral photo is taken from the participant’s face
3) A video is taken where the participant reads the instructions of how to replace a cartridge aloud
4) A video is taken where the participant is asked to express a sequence of emotions

Script:

“I will make some photos and videos of you. Could you sit on this chair?”

Now adjust the height of the tripods to ensure that the participant is clearly visible on both cameras.

“Could you hold-up your table number?”

Make a photo with the table number and face visible. Afterwards, zoom in on the participants face.

"I will now make a photo of your face. Could you look into this camera with a neutral facial expression?"

Make the neutral photo. Only repeat this if the participant's eyes are closed.

"I will now make a video of your face. Could read this text aloud after I said 'ok'?"

Say 'ok' and make the cartridge video.

"I will make another video of your face. While making this video, you will be shown a number of emotions on this screen. You should express these emotions until the next emotion appears. You should look into this camera and you should not talk while making the video. The emotions that you will be shown are: neutral, anger, fear, joy, disgust, sadness and surprise. We start when the first emotion is displayed on the screen."

Show the powerpoint. The sequence is: neutral, anger, fear, joy, disgust, sadness, surprise. Do not repeat the video if someone starts giggling or something.

"Thank you for your cooperation. You can now return to your table."

A1.5 Instructions ultimatum game (Tilburg University phase)

General Instructions

Welcome to the experiment.

The experiment takes place in both parts of the lab. All participants received the same instructions. Please read them carefully.

Do not communicate with any of the other participants during the entire experiment. If you have any questions, raise your hand or knock on your door, and wait until the experimenter comes to you to answer your question in private.

The amount of money you will earn depends on the decisions made by you and other participants in the other lab.

The experiment has two parts. For both parts, you will get a new set of instructions. The instructions for Part 1 are below. The instructions for Part 2 will appear on your computer screen at the point Part 2 starts. At the end of the experiment, we will randomly select one of the two parts and pay you the amount of money you earned in that part. You receive a fee of €3 for participating in these two parts of the experiment.

You will remain anonymous to the other participants. We will not reveal your identity, and pay you the next day by a bank transfer.

Instructions Part 1

There are two types of players: player 1 and player 2. Each player 1 will be randomly matched to a player 2.

The task of player 1 is to propose how to divide €9 between player 1 (so him- or herself) and player 2. Player 1 can choose between two options (A and B).

In option A, player 1 gets €7 and player 2 gets €2. In option B, player 1 gets €4 and player 2 gets €5.

After seeing the proposal by player 1, player 2 chooses to accept or refuse it. If player 2 accepts the proposal, the money is divided as specified in the proposal. If player 2 refuses, none of the players earns money.

Instructions Part 2 (shown on computer screen after Part 2 had finished)

In this part there are again two types of players, player 1 and player 2. You will keep the same role as in the previous part, and you will be randomly rematched with another player.

The task of player 1 is again to propose how to divide 9 EUR between player 1 and player 2.

This time, player 1 can choose to offer any amount between 0 and 9 EUR (in multiples of 0.5 EUR) to player 2 and keep the rest of the 9 EUR to him- or herself.

Player 2 indicates for each possible amount that player 1 may offer, whether he/she accepts or refuses the proposal.

Once both players have made their choices, the actual proposal of player 1 is shown to player 2. If player 2 has indicated he/she would accept that particular proposal, the 9 EUR is divided as specified in the proposal. If player 2 has indicated he/she would reject that particular proposal, none of the players earn money.

A1.6 Instructions observer tasks (University of Amsterdam Phase)

The instructions are those for the Photo 5s task in the hot ultimatum game. Instructions for the other tasks are very similar and available upon request.

Welcome

Welcome to this experiment. During the experiment you are **not allowed to communicate**. If you have any questions at any time, please raise your hand. An experimenter will assist you privately. You will make your decisions **privately and anonymously**. Your name will never be linked to your decisions and other participants will never be able to link you with your personal decisions or earnings from the experiment.

Today's experiment consists of **4 parts**. At the beginning of each part, you will receive new instructions. Your earnings depend on your decisions and the decisions of other participants. Your earnings will be **paid to you privately** at the end of today's session.

Instructions part 1

You will be shown **photos of participants in another experiment**. They played the following game.

The game

There were two types of players: **player 1 and player 2**. Each player 1 was randomly and anonymously matched to a player 2. The task of player 1 was to propose how to **divide €9** between player 1 (so him- or herself) and player 2.

Player 1 could choose between **two proposals (A and B)**

Proposal A: player 1 gets €4 and player 2 gets €5.

Proposal B: player 1 gets €7 and player 2 gets €2.

After seeing the proposal by player 1, player 2 chose to **accept or refuse** it. If player 2 accepted the proposal, the money was divided as specified in the proposal. If player 2 refused, none of the players earned money.

Your task

You will be shown photos of participants in the game which is described above. Each time, you will be shown a pair of **photos of participants that were player 2** in the game described above.

Both participants on the photos **received proposal B**. In each pair that you will be shown, one participant accepted the proposal and one refused the proposal. It is your task to **select the participant that refused** the proposal.

In this part, you will be shown 20 pairs. From these 20 pairs, 2 will be randomly selected for payment. For these 2 pairs, you will receive €5 for each pair that you judged correctly. If your judgment was incorrect, you will receive nothing. At the end of today's experiment, you will be informed how many of the selected 2 pairs you judged correctly.

On the next page, you will see an example of the task.

Example

The participants on the photos were asked to show a **neutral expression**. The photos were taken before the participants knew anything about the game that they would play.

In the experiment and the example below, the photos will be visible for 5 seconds. After this, they will disappear but you can still make a choice. **You make a choice by clicking on the photo** of your choice.

Remember, it is your task to select the participant that **refused** the proposal.

After making your decision, you will be asked how sure you are about your last decision. You can indicate this by moving the slider closer to 'completely unsure' or 'completely sure'. You can try this in the example below. Your payment does not depend on where you put the slider: **only your choice between the two participants matters**.

[Example]

You can still go back to the instructions by clicking on 'back to instructions' below. If you understand the task, click on 'continue' to make your decisions. If you need help, please raise your hand.

A1.7 Instructions eyes-test task (University of Amsterdam Phase) [1]

For each set of eyes, choose and select which word best describes what the person in the picture is thinking or feeling. You may feel that more than one word is applicable but please choose just one word, the word which you consider to be most suitable. Before making your choice, make sure that you have read all 4 words. You should try to do the task as quickly as possible but you will not be timed. If you really don't know what a word means you can look it up by moving your mouse over the question mark. By doing so, you will also see the Dutch translation.

For you participation in this part, you will receive €5. Your payment does not depend on your choices in this part.

Below is an example. You can proceed by selecting a word and clicking on 'OK'.

A1.8 Instructions ultimatum game for observers (University of Amsterdam Phase)

In this part, you will play **the following game.**

There are two types of players: **player 1 and player 2**. Each player 1 is randomly matched to a player 2. The task of player 1 is to propose how to **divide €9** between player 1 (so him- or herself) and player 2. Player 1 can choose between **two proposals (A and B).**

Proposal A: player 1 gets €4 and player 2 gets €5.
Proposal B: player 1 gets €7 and player 2 gets €2.

Player 2 chooses to **accept or refuse** it. If player 2 accepts the proposal, the money is divided as specified in the proposal. If player 2 refuses, none of the players earns money.

On the right, you **indicate what you would do** in this game, both as player 1 and as player 2.

[1] In this task, observers are presented 36 photos of pairs of eyes. They are asked to choose which of four possible words describes best what the person on the photo is feeling. We used the original instructions (Baron-Cohen et al., 2001). Participants had on-screen access to a short description of the words (in English) as well as the Dutch translations of the words (taken from Lowyck et al., 2007).

After everyone in the lab has made a decision, the computer will **randomly match you** with another participant. The computer will randomly determine who will be player 1 and player 2. Then, your decisions will be implemented and the earnings of this game will be **added to your payment**.

If you understand the instructions above, click on 'continue' to make your decisions. If you need help, please raise your hand.

A1.9 Instructions attractiveness, weight, intelligence and masculinity ratings

The text in italics was specific for the different rating tasks. The tasks are denoted by (i) attractiveness, (ii) weight, (iii) intelligence and (iv) masculinity.

Thank you for participating in this experiment. This is a short experiment and you will earn a flat fee of 10 euro for your participation. Please remain silent throughout the experiment.
You will see pictures of people who participated in an experiment in Tilburg.

(i) *Your task is to rate the attractiveness of each person. You can do this on a seven point scale, where 1 indicates "very unattractive", 4 indicates "average-looking", and 7 indicates "very attractive".*

(ii) *Based on the picture, we ask you to give your best estimate of the person's weight. You can do this on a seven point scale, where 1 indicates "heavily underweight", 4 indicates "average weight", and 7 indicates "heavily overweight".*

(iii) *Based on the picture, we ask you to give your best estimate of the person's intelligence. You can do this on a seven point scale, where 1 indicates "very unintelligent", 4 indicates "average intelligence", and 7 indicates "very intelligent".*

(iv) *Your task is to rate the masculinity or femininity of each person. For pictures of men, you can do this on a seven point scale, where 1 indicates "not masculine at all", 4 indicates "average masculinity", and 7 indicates "very masculine". For pictures of women, 1 indicates "not feminine at all", 4 indicates "average femininity", and 7 indicates "very feminine".*

In total, you will see 131 pictures. Please take your time to make careful evaluations and remain seated once you have finished. The experiment is over when all participants have rated all pictures.
If you have any questions, please raise your hand and we will come to you to answer your question in private.
Press "ready" when you are ready to start.

A2 Supplementary tables

Table A1: Descriptive statistics

	N	Mean	St. dev.	Min.	Max.
Proposers (Tilburg)					
Makes low offer (7,2) in hot UG	131	0.53			
Offer in cold UG	131	3.65	0.93	0	6
Female	131	0.55			
Age	131	23.4	2.68	18	36
Responders (Tilburg)					
Accept low offer (7,2) in hot UG	69	0.71			
Accept high offer (4,5) in hot UG	62	1.00			
Minimum accepted offer (MAO) cold UG	131	2.27	1.47	0	6
Female	131	0.47			
Age	131	21.3	2.25	18	32
Facial asymmetry (fA)	131	0.00	3.07	-4.95	10.01
Facial width-to-height-ratio (fWHR)	131	2.07	0.16	1.77	2.70
Facial masculinity (fM)	131	0.00	2.41	-7.03	4.85
Attractiveness (mean of peer ratings)	131	3.44	0.64	1.50	5.00
Intelligence (mean of peer ratings)	131	4.01	0.73	1.75	5.50
Weight (mean of peer ratings)	131	4.06	0.79	2.50	6.63
Masculinity (mean of peer ratings)	131	3.78	0.72	2.00	5.38
Observers (Amsterdam)					
Female	304	0.50			
Age	304	22.6	3.10	17	41

Table A2: Accuracy of judgments of rejecters with an angry button

	(1)		(2)	(3)	(4)
	Accuracy in % (*p*-values)		Responders (%) judged correctly over 50% of the time	Observers (%) with over 50% correct judgments	*N* observers
Angry buttons	59.8	(0.000)	87.7	75.8	128
Big angry buttons	62.6	(0.000)	83.3	78.3	128

Notes: Accuracy rates in (1) and *p*-values from two-sided Wilcoxon signed-ranks test (in parentheses) are based on the mean accuracy of each responder as the unit of observation. Percentages in (2) and (3) are computed by dividing the number of responders or observers with more than 50 percent correct judgments by the total number of responders or observers with strictly more or less correct judgments than 50 percent.

Table A3: Cues associated with rejection (mixed effects model)

Dep. var.: Identified as rejecter by observers	
Δ Baseline anger	-.015 (.011)
Δ fA	.032 (.011)***
Δ fWHR	.005 (.012)
Δ fM	-.022 (.010)**
Δ Perceived attractiveness	-.010 (.011)
Δ Perceived intelligence	-.032 (.013)**
Δ Perceived weight	-.004 (.014)
Δ Perceived masculinity	.023 (.011)**
Female in mixed gender pair	-.002 (.039)
Mixed gender pair	.001 (.028)
Horizontal head orient.	-.014 (.011)
Vertical head orient.	-.003 (.003)
Rejects	.029 (.037)
Constant	.509 (.034)***
Observations	2560
Groups	69

Notes: The table reports results from a mixed effects model with random effects for the chosen responder and the paired responder. The dependent variable is a dummy indicating whether the responder was chosen as rejecter by observers. Variables with prefix Δ refer to differences between the responder and the paired responder. All independent variables are normalized to have mean zero and a standard deviation of one, except for the gender dummy variables and the head orientation controls. Standard errors in parentheses. *** p < 0.01, ** p < 0.05, * p < 0.1

Table A4: Cues associated with rejection by gender

	Panel A			
	(1)	(2)	(3)	(4)
Dep. var.:	Females Responder rejects	Females Responder identified as rejecter	Males Responder rejects	Males Responder identified as rejecter
Baseline anger	.012 (.066)	-.007 (.017)	-.444 (.800)	.082 (.164)
fA (facial asymmetry)	.088 (.119)	.049 (.031)	.192 (.090)**	.035 (.020)*
fWHR(width-to-height ratio)	-.341 (.149)**	-.017 (.042)	.037 (.091)	-.007 (.019)
fM (facial masculinity)	.094 (.094)	-.022 (.024)	-.028 (.092)	.009 (.019)
Perceived attractiveness	.055 (.115)	-.022 (.029)	.178 (.109)	.008 (.023)
Perceived intelligence	-.341 (.167)*	-.046 (.046)	-.109 (.086)	-.019 (.018)
Perceived weight	-.211 (.124)	-.026 (.033)	-.114 (.116)	.031 (.024)
Perceived masculinity	-.023 (.116)	.004 (.029)	.006 (.097)	.027 (.020)
Horizontal head orient.	-.024 (.061)	-.017 (.015)	-.131 (.096)	-.019 (.020)
Vertical head orient.	.029 (.022)	-.004 (.006)	.005 (.025)	-.004 (.005)
Rejects		.034 (.055)		.080 (.040)*
Constant	.376 (.151)**	.485 (.044)***	.346 (.248)	.520 (.052)***
Observations	32	32	37	37
Adj. R^2	.170	.131	.011	.282
R^2	.437	.439	.285	.501

	Panel B			
Rejects		.085 (.042)*		.105 (.036)***
Observations		32		37
Adj. R^2		.089		.171
R^2		.118		.194

Notes: The table reports results from OLS regressions. The dependent variable in (1) and (3) is the responder's choice to reject the unfair offer in the hot UG, and in (2) and (4) the proportion of times that a responder was chosen by observers in the photo tasks. All independent variables are normalized to have mean zero and a standard deviation of one, except for the dummy variables and the head orientation controls. Standard errors in parentheses.*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

A3 Construction of facial measures

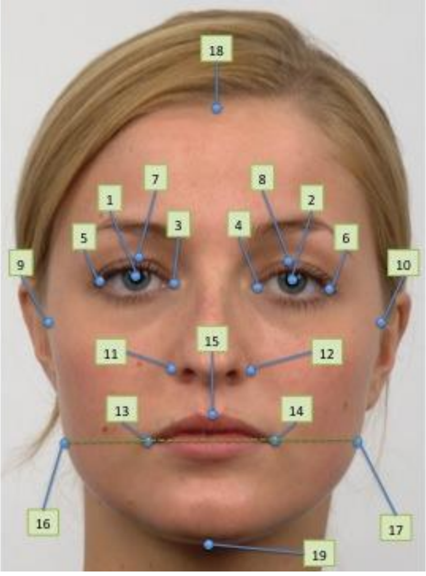
We marked 19 different points on each responder's face, using the Image J application.[2] Figure A1 shows the location of these points. Before measuring the points, the picture was rotated so that the pupils fell on the same y-coordinate. Based on the 19 points, we computed 11 different distances. The measure we use for the facial width-to-height ratio is based on the measure used by Lefevre et al. (2013). It takes the ratio between the bizygomatic width (X10-X9) and the upper face height (the distance between highest point of the eyelids and the top of the mouth: Y15-(Y7+Y8)/2).

The measure for facial asymmetry is based on the measure used by Little et al. (2008). To calculate facial asymmetry, we compute the absolute differences between the left and right distance from a 'midline' on 6 different points. The x-coordinate of the midline is computed by the midpoint of the distance between the pupils (M=X1+(X2-X1)/2). We compute the absolute differences for the inner eye corners (|(X4-M)-(M-X3)|), outer eye corners (|(X6-M)-(M-X5)|), cheekbones (|(X10-M)-(M-X9)|), nose (|(X12-M)-(M-X11)|), mouth (|(X14-M)-(M-X13)|) and the jaw (|(X17-M)-(M-X16)|). To account for possible differences in distance from the camera, each of the absolute differences is normalized by dividing it by the inter-pupillary distance (X2-X1). Each of the absolute differences were converted to a *z*-score and summed up to one asymmetry score.

Facial masculinity was computed using the measure by Apicella et al. (2008). It consists of four different ratios that have found to be sexually dimorphic. These four ratios are 'cheekbone prominence', which takes the ratio between the facial width at the cheekbones and at the jaws (ChP=(X10-X9)/(X17-X16)), the ratio between the 'jaw height' and the 'lower face height' (JH/LFH=(Y19-(Y16+Y17)/2)/(Y19-(Y1+Y2)/2)), the ratio between the 'lower face height' and the 'face height' (LFH/FH=(Y19-(Y1+Y2)/2)/(Y19-Y18)), and the ratio between facial width at the cheekbones and 'lower face height' (FW/LFH=(X10-X9)/(Y19-(Y1+Y2)/2)). Each of the four ratios is converted to a *z*-score and these *z*-scores are summed to one score: (JH/LFH+LFH/FH)-(ChP+FW/LFH).

[2] This can be downloaded from http://rsbweb.nih.gov/ij/download.html

Figure A1: The 19 points marked



Notes: The person on the photo did not take part in this study. The photo merely serves as an example. This photo is part of the Radboud Faces Database (Langner et al. 2010).

A4 Video treatments

We discuss treatments in which observers predicted responder behavior using videos instead of photos.

A4.1 Procedures and motivation

A4.1.1 Tilburg phase: video recordings

We took two videos of each responder after the photos had been taken. We videotaped responders while they read instructions on how to replace a printer cartridge aloud. The aim of this task was to acquire a video of the responders in a neutral state. We also videotaped responders while they were asked to express the following sequence of seven emotions, each for a period of 10 seconds: neutrality, anger, fear, joy, disgust, sadness, and surprise. The purpose of this task was to have a video of the responders with an angry facial expression. To avoid revealing the purpose of the experiment, emotions other than anger were videotaped as well.

A4.1.2 Amsterdam phase: video tasks

In the video tasks, observers watched silent video clips of two responders, played side-by-side. As in the photo tasks, one of the responders shown rejects, and one accepts, an offer of (7,2).These recordings either contained responders (a) reading a neutral text on how to replace a cartridge in a printer (cartridge**)**, (b) expressing a neutral state followed by expressing anger (short emotions, this clip consisted of the neutrality and anger part of the emotion video described in Section A4.1.1**)**, or (c) expressing all seven videotaped emotions (long emotions, consisting of the entire emotion video described in Section A4.1.1**)**. The purpose of the video clips was to provide observers with more information about the responder than the (static) photos. On the one hand, reading a text out loud (as in the cartridge video) creates natural movement. On the other hand, letting subjects mimic emotions (as in the other videos) may provide cues on the extent to which they naturally express these emotions. The variety of tasks allowed us to consider whether more information helps observers to identify rejecters.

A4.2 Results and discussion

For the video tasks we cannot reject that the observers' predictions are at chance levels. The accuracy rates are between 50.8 and 52.4 percent (see Table A5). Across all photo and video tasks, the achieved accuracy rate is 53.1 percent, which is significantly better than chance. Overall, observers do 6.2% better than chance.

Interestingly, rejecters with an angry button are judged more accurately, just like in the photo tasks (see Table A5). To illustrate, pairs where the rejecter has an angry button (top 50% (predicted) reactive anger) are judged correctly in 55.5 percent of the cases. In pairs where the rejecter has a big angry (top 25% (predicted) reactive anger) button this increases to 57.5 percent of the cases. In both cases, accuracy is significantly better than chance.

Table A5: Accuracy of judgments in the video tasks

	(1) Accuracy in % (*p*-values)		(2) Responders (%) judged correctly over 50% of the time	(3) Observers (%) with over 50% correct judgments	(4) *N* observers
All responders					
Cartridge	50.8	(0.620)	55.2	52.1	60
Short emotions	52.4	(0.206)	56.3	57.8	52
Long emotions	51.6	(0.251)	60.6	65.7	44
All video tasks	51.7	(0.221)	57.4	57.8	156
Photo and video tasks	53.1	(0.002)	63.8	61.8	284
Angry buttons					
All video tasks	55.5	(0.001)	69.0	68.6	156
Photo and video tasks	57.5	(0.000)	82.1	71.8	284
Big angry buttons					
All video tasks	60.2	(0.000)	78.8	75.0	156
Photo and video tasks	61.3	(0.000)	90.4	76.4	284

Notes: Accuracy rates in (1) and *p*-values from two-sided Wilcoxon signed-ranks test (in parentheses) are based on the mean accuracy of each responder as the unit of observation (*N* = 69). Percentages in (2) and (3) are computed by dividing the number of responders or observers with more than 50 percent correct judgments by the total number of responders or observers with strictly more or less correct judgments than 50 percent.

That the accuracy rates are lower for the video tasks suggests that observers focus on different signals than in the photo tasks and use some irrelevant cues. Table A6 presents results from regressions A comparison of column (2) in Table 4 and column (2) in Table A6 shows that, for the most part, observers use the same cues in the photo and video tasks.[3] The main difference is that they rely less on fA in the video tasks, possibly because it is harder to assess fA when people are in motion. This difference can explain roughly half a percentage point of the 3 percentage points difference in accuracy between the tasks.

[3] Baseline anger on the videos is measured by taking the mean anger expressed on the cartridge video and the part of the emotions video in which subjects take on a neutral expression.

Table A6: Actual and perceived correlates of rejection in the video tasks

	Panel A	
	(1)	(2)
Dep. var.:	Responder rejects	Responder identified as rejecter
Baseline anger	-.043 (.059)	.009 (.012)
fA (facial asymmetry)	.193 (.062)***	.026 (.014)*
fWHR (width-to-height ratio)	-.062 (.067)	.021 (.014)
fM (facial masculinity)	.041 (.055)	-.005 (.012)
Perceived attractiveness	.072 (.065)	-.017 (.014)
Perceived intelligence	-.128 (.071)*	-.007 (.015)
Perceived weight	-.120 (.076)	-.006 (.016)
Perceived masculinity	.057 (.061)	.020 (.013)
Female	.033 (.120)	-.038 (.025)
Horizontal head orient.	-.063 (.047)	-.007 (.010)
Vertical head orient.	.009 (.014)	-.003 (.003)
Rejects		.008 (.028)
Constant	.354 (.104)***	.507 (.024)***
Observations	68	68
Adj. R^2	.085	.137
R^2	.235	.292
	Panel B	
Rejects		.033 (.026)
Observations		69
Adj. R^2		.010
R^2		.024

Notes: The table reports results from OLS regressions. The dependent variable in (1) is the responder's choice to reject the unfair offer in the hot UG, and in (2) the proportion of times that a responder was identified as the rejecter by observers. All independent variables are normalized to have mean zero and a standard deviation of one, except for the dummy variables and the head orientation controls. Standard errors in parentheses.*** $p <$ 0.01, ** p < 0.05, * p < 0.1

A5 Cold ultimatum game

In this appendix, we discuss the treatments conducted in which observers predicted responder behavior in the cold UG, which was played after the hot UG in the lab of Tilburg University.

A5.1 Procedures and motivation

In all observer sessions except the ones based on the long emotions videos, observers were asked to identify responders who reject low offers in the cold UG.[4] Doing so allowed us to study the conjecture that the emotional underpinnings of the decision to reject are not as strong with the strategy method as with the direct response method, and that observers recognize this difference. We elicited minimal acceptable offers (MAO) in the cold UG, and these are noisy proxies for how people decide when they are actually confronted with a low offer, and thus potentially harder to predict for observers than hot decisions. Previous evidence suggests that low offers are less frequently rejected with the strategy method than with the direct method (Brandts and Charness, 2011; Oosterbeek et al., 2004).

To create the sample used for this task, we took the 62 responders who received the high offer in the hot UG (which they all accepted) and split the sample at a cutoff MAO. This resulted in 34 responders who would have accepted a (7,2) split and 28 responders who would have rejected it. Every observer saw 28 pairs. In each trial observers were asked to guess which of the two responders shown would have rejected a (7,2) split.

A5.2 Results and discussion

We conjectured that under the strategy method, the emotional underpinnings of the decision to reject are not as strong as with the direct method, and thus potentially harder for observers to detect.

Table A7 shows the accuracy of observers' judgments for the cold UG. As can be seen in the table, the accuracy rate across all tasks is 50.5 percent, which is not significantly different from chance ($p = 0.441$, two-sided Wilcoxon signed-ranks test). With the exception of the cartridge video task, the accuracy is at chance levels for each task considered individually. The same is the case for the pooled data from the photo tasks and the pooled data from the video tasks.

We also compare the cues used in the hot and cold UG. Table A8 shows the results of regressions that establish which cues are predictive in the cold UG and which cues observers use. The asymmetry result that we find for the hot UG does not carry over to the cold game. Facial asymmetry is not significantly correlated with rejection in the cold UG. This suggests that our observers seem to sense that both types of rejections are different in nature, because they do not use facial asymmetry as a cue in the guessing task based on the sample drawn from the cold UG.

[4] For the long emotion videos, observers judged either decisions based on the hot or the cold UG.

Table A7: Accuracy of judgments for the cold UG

	(1) Accuracy in %	(*p*-values)	(2) Responders (%) judged correct over 50% of the time	(3) Observers (%) with over 50% correct judgments	(4) *N* observers
Photo tasks					
5s	51.1	(0.371)	53.6	55.4	74
1s	48.9	(0.455)	42.4	44.2	54
All photo tasks	50.2	(0.900)	48.4	50.9	128
Video tasks					
Cartridge	53.5	(0.009)	67.2	68.5	60
Short emotions	47.7	(0.184)	40.4	36.2	52
Long emotions	51.2	(0.414)	55.1	58.8	20
All video tasks	50.9	(0.309)	59.3	54.2	132
Photo and video tasks	50.5	(0.441)	56.5	52.7	260

Notes: Accuracy rates in (1) and the *p*-values (in parentheses) come from two-sided Wilcoxon signed-ranks test taking the mean accuracy of each responder as the unit of observation ($N = 62$). Fractions in (2) and (3) are computed by dividing the number of responders or observers with more than 50 percent correct judgments by the total number of responders or observers with strictly more or less correct judgments than 50 percent.

Table A8: Actual and perceived correlates of rejection in the cold UG

	Panel A		
	(1)	(2)	(3)
Dep. var.:	Responder rejects	Photo tasks Responder identified as rejecter	Video tasks Responder identified as rejecter
Baseline anger (photos)	-.028 (.072)	-.009 (.013)	
Baseline anger (videos)			-.016 (.014)
fA (facial asymmetry)	.060 (.072)	-.002 (.013)	-.001 (.014)
fWHR(width-to-height ratio)	.071 (.086)	.013 (.016)	-.008 (.016)
fM (facial masculinity)	-.021 (.089)	-.021 (.016)	-.015 (.017)
Perceived attractiveness	.079 (.101)	-.023 (.019)	-.004 (.019)
Perceived intelligence	-.068 (.087)	-.041 (.016)**	-.013 (.016)
Perceived weight	-.126 (.111)	-.020 (.020)	.022 (.021)
Perceived masculinity	.054 (.073)	.030 (.013)*	.023 (.014)*
Female	-.015 (.181)	.004 (.033)	-.030 (.033)
Horizontal head orient.	.033 (.059)	-.003 (.011)	-.014 (.011)
Vertical head orient.	-.006 (.019)	.001 (.003)	-.000 (.004)
Rejects		-.005 (.026)	.024 (.026)
Constant	.426 (.132)**	.520 (.027)***	.523 (.027)***
Observations	61	61	61
Adj. R^2	-.126	.246	.113
R^2	.081	.397	.291
	Panel B		
Rejects		.004 (.029)	.018 (.027)
Observations		61	61
Adj. R^2		-.016	-.009
R^2		.000	.007

Notes: The table reports results from OLS regressions. The dependent variable in (1) is the responder's choice to reject the unfair offer in the cold UG, and in (2) and (3) the proportion of times that a responder was chosen by observers. All independent variables are normalized to have mean zero and a standard deviation of one, except for the dummy variables and the head orientation controls. Standard errors in parentheses.*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$