SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
SOCIAL EVOLUTION
BIOLOGICAL ANTHROPOLOGY
PSYCHOLOGY
ANIMAL BEHAVIOUR

Received
14 August 2012

Accepted
3 December 2012

Published
10 January 2013

Correspondence and requests for materials should be addressed to C.E. (charles.efferson@econ.uzh.ch) or S.V. (sonja.vogt@econ.uzh.ch).

Viewing men's faces does not lead to accurate predictions of trustworthiness

Charles Efferson[1,2] & Sonja Vogt[1,2]

[1]Department of Economics, University of Zurich, Zurich, Switzerland, [2]Laboratory for Social and Neural Systems Research, University of Zurich, Zurich Switzerland.

The evolution of cooperation requires some mechanism that reduces the risk of exploitation for cooperative individuals. Recent studies have shown that men with wide faces are anti-social, and they are perceived that way by others. This suggests that people could use facial width to identify anti-social men and thus limit the risk of exploitation. To see if people can make accurate inferences like this, we conducted a two-part experiment. First, males played a sequential social dilemma, and we took photographs of their faces. Second, raters then viewed these photographs and guessed how second movers behaved. Raters achieved significant accuracy by guessing that second movers exhibited reciprocal behaviour. Raters were not able to use the photographs to further improve accuracy. Indeed, some raters used the photographs to their detriment; they could have potentially achieved greater accuracy and earned more money by ignoring the photographs and assuming all second movers reciprocate.

The evolution of cooperation is a difficult and controversial puzzle because a cooperative individual pays a personal cost to produce benefits for others[1–4]. This renders the cooperator vulnerable to exploitation by selfish types who take advantage of these benefits without cooperating themselves. The result is a potential selective disadvantage for those who cooperate. In spite of this general theoretical point, humans often do cooperate. In particular, we do so with unrelated individuals in one-shot interactions, and this especially strong form of altruistic cooperation seems to distinguish us from closely related species[3–9].

A number of theories can explain the evolution of cooperation[3,10], including the altruistic variants of cooperation that characterize human social behavior[4]. These theories differ enormously in terms of the precise mechanisms involved, and disagreements over which mechanisms are important lie at the root of the persistent controversy about the evolution of human pro-social behavior[3,10,11]. Regardless, candidate mechanisms share the feature that they somehow produce positive assortment, which we can broadly define as an outcome in which defectors are at least partially excluded from the benefits produced by cooperation[3,4]. In general, to the extent that some mechanism directs the benefits produced by cooperation toward others who also cooperate, the mechanism limits the risk of exploitation. It thus attenuates or even reverses the selective disadvantage associated with cooperating.

Generating assortment would be straightforward if individuals could quickly identify each other by type. A mechanism of this kind would be particularly useful in interactions with unfamiliar individuals where information about one's partner is limited. One version of such a mechanism is the green-beard effect[1–3]. As originally formulated, green beards work in the following way. A single allele produces both an observable trait, namely the metaphorical green beard, and a behavioural tendency to recognize and cooperate with others having the same observable trait. The green beard, whatever it is in practice, renders the presence of a shared allele observable, and so individuals with the allele can find each other and enjoy the benefits of mutual cooperation while limiting the risk of exploitation.

Green beards represent an interesting theoretical possibility because they aptly illustrate the importance of assortment for the evolution of cooperation. As a specific mechanism, however, they have a serious problem[2,3]. To work effectively, green beards require a strong and stable association between the observable trait and the unobservable presence of a specific allele. If we admit the possibility of a rare mutant who has the observable trait but not the allele of interest, the system will unravel because cooperating given the presence of a green-bearded partner makes one vulnerable to green-bearded defectors. We should thus not be surprised that, in the nearly half century since the green-beard effect was first suggested, only a few possible examples involving ants[12], amoebas[13], and yeast[14] have been found.

Intriguingly, however, a number of recent studies have produced results suggesting that men could have observable traits analogous to green beards. Specifically, these studies show that men with wide faces, adjusted

for height, are more aggressive[15,16], more prone to unethical behavior[17], and less trustworthy in social dilemmas[18] than men with narrow faces. A relationship exists, in other words, between an observable measure of facial structure and the tendency to behave pro-socially or anti-socially. Facial width is analogous to a green beard because it is an observable marker that reveals what is unobservable ex ante but nonetheless critically important, namely the tendency to behave in a certain way in a specific type of social interaction.

If one key relationship is the link, discussed immediately above, between facial structure and *actual* behavioural tendencies, a second key relationship is the link between facial structure and *perceived* behavioural tendencies. At least two recent studies show that men with wide faces are not only more aggressive and more selfish than men with narrow faces, they are apparently perceived that way by others[16,18]. Additional studies also show that people can use facial characteristics to make accurate judgements about specific personality traits in others[19,20]. Facial width, given these results, is not only a marker of behavioural tendencies that are important but unobservable ex ante, it might also be used by others as a cue of such behavioural tendencies. Again, facial width would be analogous to green beards in this case. The original discussion of green beards posited simply that green-bearded individuals cooperate with other green-bearded individuals. More generally, however, we can think of green beards as arbitrary conspicuous markers that, if reliable, can be used by others as cues to reduce the statistical risk of exploitation in social dilemmas via some kind of conditional behaviour. The critical consideration is the reliability of the marker and, by extension, the accuracy of those who use the marker as a cue.

Reliability matters because without it the unobservable trait of primary interest, a trait like the tendency to cooperate in social dilemmas, is not really observable. Using the marker to generate assortment with respect to this primary trait is thus impossible. Even if the marker is reliable, however, it must also be appropriately used as a cue by others. Indeed, assortment can only occur when individuals use the observable trait to draw accurate inferences about others and condition their behaviour accordingly. This is true if accuracy is a rather figurative concept in which green-bearded cooperators follow a simple algorithm by cooperating categorically with other green-bearded individuals. This is the original green-beard effect, and accuracy in this case refers to the rate at which the green-beard algorithm leads to mutual cooperation rather than exploitation. The importance of accuracy also holds more generally if an individual uses an observable trait to consciously estimate whether she is facing another cooperator, and she then cooperates if she concludes that she is. We will call this latter procedure "green-bearded typecasting", and the inferential accuracy at the root of green-bearded typecasting is the focus of this paper.

We focus on green-bearded typecasting instead of the original green-beard mechanism for three reasons. First, in terms of actual behaviour, it is the more general of the two ideas; the original green-beard effect is behaviourally equivalent to a special case of typecasting in which potential partners with a green beard are estimated to be cooperators with probability 1. Second, green-bearded typecasting is of special interest in humans because in many domains humans are prone to typecast others about whom they know little[21,22]. Finally, in terms of observable behaviour, the original green-beard algorithm conflates preferences over outcomes with beliefs about the likely behaviour of one's partner. Assume, for example, that we observe a focal individual with a green beard cooperating with another green-bearded individual. On the one hand, the focal individual might cooperate because she has preferences that make her unconditionally generous toward green-bearded partners. This explanation depends exclusively on the focal individual's preferences regarding people with green beards. On the other hand, the focal individual might want to cooperate with *any* person she believes is also willing to cooperate, and the presence of a green beard simply affects her beliefs about this all-important question. Both mechanisms are interesting and important possibilities, but for the sake of analytical clarity we focus on the beliefs-based component of the latter possibility. To isolate effects associated with beliefs, we asked independent raters, in effect, to typecast but not to interact with others for whom we had behavioural data from a social dilemma game. The critical task is to determine if typecasting is accurate. Recent empirical results[15–20] suggest it could be, while the theoretical vulnerabilities of green beards[2,3] suggest it should not be.

To address the question of accuracy, we conducted a study involving two tasks (see Methods and Supplementary Information, SI). The first task was a behavioural experiment conducted in Munich, Germany, with male participants. The game played in this experiment was a sequential social dilemma, specifically a modified trust game[23]. In this game, first movers could either transfer their entire endowment of nine Euros to second movers, with associated efficiency gains, or they could transfer nothing at all. Consequently, each first mover faced a binary choice; he could either trust his partner or not. After learning which of these choices a first mover made, the second mover could back transfer any amount, in one-Euro increments, between zero and his endowment. Back transfers also brought efficiency gains. The choices of second movers provided us with a behavioural measure of their individual tendencies to exploit others.

The second task was implemented in Konstanz, Germany. In Konstanz, independent raters viewed photographs of the second movers in Munich. For each second mover, in addition to viewing the photograph, raters also knew whether the associated first mover trusted the second mover in the photograph. Given both the photograph of a second mover's face and the transfer decision of the first mover, raters made guesses about second mover back transfers. The accuracy of these guesses is our principal but not exclusive concern. Analyses discussed below also make use of the facial width-to-height ratios and the attractiveness of second movers (SI).

Results

Raters viewed photos and guessed the choices of 54 second movers. Of these 54 second movers, a total of 41 were trusted by their partners. Given 28 raters, we have a total of 1512 observations to evaluate accuracy. In some analyses below, we restrict attention to the 41 second movers who were trusted by their partners. These analyses hold first-mover behaviour constant, and in that sense they isolate the informational content of the photographs themselves. Given 41 second movers who were trusted, we have 1148 observations for these analyses. We explain below as needed how we account statistically for the fact that we have multiple observations per rater.

We first address the relationships between the back transfers in Euros of second movers and their facial characteristics. We focus on the 41 second movers who were trusted. These second movers are especially important because they were the players in an explicit position to exploit or reciprocate their partners' trust. We use ordered probit models for these analyses. We do so because second mover back transfers were strongly bimodal, with many second movers back transferring everything or nothing (Fig. 1). Ordered probit models require that responses are ordered, but responses do not need to be normally distributed. Moreover, the ordered probit model is most appropriate when modelling, as in our case, discrete behaviours that involve more than two options with an ordinal though not necessarily cardinal relation to an underlying set of preferences[25].

For the 41 second movers who were in a position to exploit their partners' trust, we found no relationship between facial structure and trustworthiness or between attractiveness and trustworthiness (Fig. 1). In particular, using models with single independent variables, the estimated relationship between back transfers and facial width-to-height ratios is not significant (ordered probit; estimate for

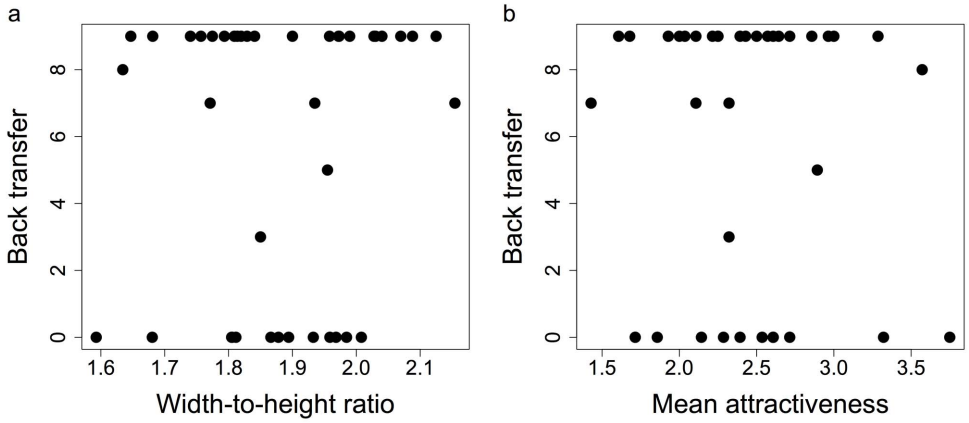

Figure 1 | Back transfers for the 41 second movers who were trusted. Back transfers are shown as a function of the width-to-height ratios of second mover faces (a) and as a function of the mean attractiveness ratings for second movers (b). Attractiveness levels range from 1 for "very unattractive" to 5 for "very attractive", and mean attractiveness levels shown here are averages over 28 independent raters of attractiveness (SI). Ordered probit regressions (Tables 1 and 2) provide no evidence for a relationship between back transfers and the facial structure or attractiveness of second movers.

width/height is 0.897, $P = 0.472$); nor is the relationship between back transfers and attractiveness levels (ordered probit; estimate for mean attractiveness is -0.126, $P = 0.706$). These results mean that neither the facial width nor the attractiveness levels of second movers could have revealed anything about their tendencies to exploit or reciprocate another's trust.

The same conclusions follow from analysing the choices of all 54 second movers using a multivariate approach. Specifically, we conducted two model selection exercises in which we modelled second mover back transfers as a function of several independent variables in various combinations (SI). The independent variables include the second mover's width-to-height ratio, the second mover's attractiveness, and a dummy variable indicating if the second mover was trusted by his partner. We repeated this model selection exercise using two different approaches to the dependent variable. Specifically, we used (i) ordered probit regressions in which we modelled second mover back transfers in Euros, and we used (ii) simple probit regressions that dichotomized second mover back transfers as zero or positive. We focus on the ordered probit results because, as discussed above, the ordered probit model provides an appropriate and thorough treatment of second mover back transfers. The simple probit treatment, however, is an important robustness check because it collapses all second movers who back transferred a positive amount into a single category. It thus provides a treatment of second mover choices that maximises our ability to identify any systematic relationship between our independent variables and pro-social choices, broadly defined, by second movers. Although we present ordered probit results, we would like to emphasize that our results and conclusions are entirely robust across both treatments of the dependent variable.

Table 1 presents the set of ordered probit models and the results of model selection based on an information theoretic criterion[24]. The weight of evidence (Table 1, $\sum_{i\in\{1,3\}} w_i = 0.999$) shows that the key independent variable is the dummy indicating whether a second mover was trusted. The coefficient on this dummy is always, if included in a specific regression (e.g. Table 2), positive and highly significant (Table 2, estimate is 1.730, $P < 0.001$), whereas the coefficients on width-to-height ratios (e.g. Table 2, $P = 0.680$) and attractiveness levels (e.g. Table 2, $P = 0.826$) are never significant at any conventional level. This latter point is true if we control for first mover behaviour by including it as an independent variable, or if we restrict attention to the 41 second movers who were trusted (ordered probit; estimate for width/height is 0.880, $P = 0.690$; estimate for attractiveness is -0.119, $P = 0.720$). In sum, second movers who were trusted reliably back transferred more than those who were not trusted. Second mover back transfers, however, bore no significant relation to facial width or attractiveness.

We next conducted a model selection exercise in which we modelled rater guesses, in Euros, about second mover back transfers (SI). The independent variables include the width-to-height ratio of the second mover's face, the attractiveness of the second mover, a dummy indicating if the second mover was trusted, and the actual back transfer of the second mover. The first three variables allow us to identify information the raters may have used to make inferences

Table 1 | Model selection, ordered probit, back transfers of all 54 second movers. The independent variables include (i) the width-to-height ratios of second mover faces, (ii) the attractiveness levels for second movers, and (iii) a dummy indicating which second movers were trusted. The final columns show the number of parameters estimated, the AIC_c values, and the Akaike weights (w_i). AIC_c is an improved form of Akaike's criterion[24,36], and Akaike weights rescale AIC_c values to show the proportional weight of evidence for each model. In this case, because the total Akaike weight over models 1 and 3 is 0.999, the exercise clearly shows that the trust of the second mover's partner is the critical independent variable

Model	W/H	Att.	Trusted	Parameters	AIC_c	w_i
1	✓	✓	✓	9	138.719	0.061
2	✓	✓		8	151.791	<0.001
3			✓	7	133.268	0.938

Table 2 | Ordered probit results for model 1 from Table 1. The intercepts reflect the back transfers that actually occurred. Although model 1 is not the best model, it is the full model, and conclusions are robust to model specification. For this reason, we show model 1

Parameter	Estimate	Robust std. error	z	P
W/H	0.516	1.252	0.41	0.680
Att.	−0.070	0.314	−0.22	0.826
Trusted	1.730	0.393	4.40	<0.001
Intercept 0/1	1.981	2.803		
Intercept 1/3	2.103	2.788		
Intercept 3/5	2.167	2.792		
Intercept 5/7	2.231	2.790		
Intercept 7/8	2.414	2.788		
Intercept 8/9	2.474	2.788		

about second mover behaviour. The fourth variable, namely the actual back transfers of second movers, is included to capture the accuracy of rater inferences. Specifically, if the model selection exercise tells us that the actual back transfer is an important independent variable, and if the estimated coefficient on this variable is positive and significant, then rater guesses about back transfers were accurate in the sense that they are significantly and positively related to actual back transfers. Importantly, however, this kind of significant relationship could disappear when adding a variable that controls for the information used by raters to make their accurate guesses. By including both the kinds of information raters might use and the actual back transfers of second movers in our model selection exercise, we can identify both any accuracy that exists and potentially the information being used to generate it.

In addition, we repeated the same model selection exercise by modelling rater guesses using simple probit regressions that treat both back transfers and rater guesses as binary (i.e. zero or positive). As above, this is an important robustness check because it collapses second mover choices and rater guesses about these choices into two simple categories. This allows us to test for any systematic relationship between actual pro-social tendencies, broadly defined, and perceived pro-social tendencies, also broadly defined. Our results and conclusions are entirely robust under both types of model, and we focus on the ordered probit treatment.

Raters apparently used first mover behaviour as well as the facial width and attractiveness of second movers as cues. To see this, Table 3 presents the set of ordered probit models fit to the data and the associated results of model selection. Model selection shows that the width-to-height ratios of second mover faces, the attractiveness levels of second movers, and first mover behaviours are all important independent variables (Table 3, $\sum_{i\in\{1,5\}} w_i = 0.905$). Raters guessed, with marginal significance, that men with wider faces back transferred less than men with more narrow faces (Table 4, estimate is -0.302, $P = 0.070$). They also guessed that more attractive second movers back transferred more than less attractive second movers (Table 4, estimate is 0.156, $P = 0.001$), and they guessed that second movers who were trusted were significantly more likely to back transfer large amounts than second movers who were not trusted (Table 4, estimate is 1.438, $P < 0.001$). Importantly, actual back transfers are significantly and positively related to guesses about back transfers under some model specifications, but the model selection results together with results from specific regressions clearly show that first mover behaviour mediates this effect.

For example, model 2 from Table 3 includes actual back transfers as an independent variable, but it does not include the dummy indicating if a second mover was trusted. The model selection criterion clearly indicates that model 2 is a poorly fitting model relative to other models under consideration (Table 3, Model 2, $w_2 < 0.001$). Nonetheless, the results from model 2 produce a highly significant relation between actual back transfers and rater guesses about back transfers (ordered probit; estimate for actual back transfer is 0.066, $P < 0.001$). Model 1 is identical except that it adds the behaviour of the first mover as a control. Because the difference in AIC_c values between these two models is 229.091 (Table 3), model 1 represents a truly enormous improvement[24] in terms of model selection. Moreover, model 1 results show a significant positive relation between rater guesses and the trust of first movers (Table 4, estimate is 1.438, $P < 0.001$). Importantly, however, under model 1 the relationship between rater guesses and actual back transfers is *not* significant (Table 4, $P = 0.231$), and this shows that it is specifically information about first mover behaviour that is responsible for the rater accuracy we identify here.

Altogether, these results indicate the following. We know from our analyses above that second movers who were trusted back transferred more than those who were not trusted. This is reciprocity, a force that commonly affects behaviour in social interactions[26,27]. If raters knew that reciprocity would influence second movers, they could have achieved some degree of accuracy by simply assuming that second movers who were trusted would back transfer more than those who were not. This reciprocity heuristic would have generated accuracy that appears, when first mover behaviour is *not* included in the regression, as a significant relationship between actual back transfers and rater guesses. When controlling for first mover behaviour, however, the effect associated with actual back transfers should disappear if raters could not or did not use any information other than first mover behaviour to improve accuracy. In this case, the dummy for first mover trust will pick up all the information used by raters to effectively generate a significant degree of accuracy. This is exactly what we find when we compare models 1 and 2 (Tables 3 and 4). Moreover, although we do not present detailed and largely redundant regression results, an analogous conclusion holds when we compare models 3 and 4 (Table 3). These findings indicate that raters achieved some degree of accuracy over all 54 second movers by assuming that at least some second movers reciprocated trust. Raters were not, however, able to achieve any additional degree of accuracy

Table 3 | Model selection, ordered probit, rater guesses about back transfers for all 54 second movers. The total number of observations is 1512. Independent variables include (i) the width-to-height ratios of second mover faces, (ii) the attractiveness levels for second movers, (iii) a dummy indicating which second movers were trusted, and (iv) the actual back transfers of second movers. The final columns show the number of parameters estimated, the AIC_c values, and the Akaike weights (w_i). Because models 1 and 5 constitute over 90% of the total Akaike weight, model selection clearly shows that width-to-height ratios, attractiveness levels, and first mover behaviour are all important predictors of rater inferences

Model	W/H	Att.	Trusted	BT	Parameters	AIC_c	w_i
1	✓	✓	✓	✓	13	4785.265	0.287
2	✓	✓		✓	12	5014.356	<0.001
3			✓	✓	11	4789.968	0.027
4				✓	10	5022.513	<0.001
5	✓	✓	✓		12	4783.730	0.618
6	✓	✓			11	5105.160	<0.001
7			✓		10	4788.163	0.067

Table 4 | Ordered probit results for model 1 from Table 3. The intercepts reflect the rater guesses that actually occurred. Although model 1 is not the best model, it is the full model, and conclusions are robust to model specification. For this reason, we show model 1. To account for the fact that we have multiple guesses per rater, we calculated robust standard errors by clustering on rater[25]

Parameter	Estimate	Robust std. error	*z*	*P*
W/H	−0.302	0.166	−1.81	0.070
Att.	0.156	0.047	3.31	0.001
Trusted	1.438	0.202	7.11	<0.001
BT	0.006	0.005	1.20	0.231
Intercept 0/1	0.944	0.401		
Intercept 1/2	1.028	0.394		
Intercept 2/3	1.154	0.383		
Intercept 3/4	1.291	0.376		
Intercept 4/5	1.448	0.370		
Intercept 5/6	1.664	0.371		
Intercept 6/7	1.774	0.372		
Intercept 7/8	1.919	0.374		
Intercept 8/9	1.987	0.377		

by using the photographs of second movers. The significant coefficients for facial width and attractiveness reveal that raters did respond to information in the photographs of second movers; they just could not use the information to improve the accuracy of their inferences. More generally, the lack of accuracy associated with the 41 second movers who were trusted shows that raters could not use the information in the photographs to identify the second movers who exploited their partners.

These results are based on regressions that model individual rater guesses and correct for multiple guesses per rater by calculating robust standard errors clustered on rater[25]. To verify the robustness of our conclusions, we also analysed rater accuracy directly by using a different approach. The results in this case confirm the lack of accuracy identified above, and they also suggest that some of the raters may have actually used the photographs to their detriment.

For each second mover, we categorized his back transfer as either zero or positive. We also categorized each rater's guess about a back transfer as zero or positive. We then calculated a simple binary variable that measures the accuracy of each guess. A guess was accurate if the back transfer and the guess were both positive or if both were zero. Otherwise, the guess was inaccurate. Given this binary variable, we tested accuracy at the individual level using binomial tests by rater. We then corrected for multiple tests with a procedure[28] that maximises power. This is a generous definition of accuracy that ignores the magnitudes of second mover back transfers and rater guesses and thus maximises the potential to identify raters who accurately identified second movers who made positive transfers of any kind. By this definition, a single rater had an accuracy rate above chance (i.e. a null of 0.5) when we restrict attention to the 41 second movers who were trusted (SI, Table S1). Over all 54 second movers, eight raters had accuracy rates above chance (SI, Table S2). Interestingly, however, 10 raters had an accuracy rate significantly lower than what they would have achieved by simply assuming that second movers reciprocated transfers of zero with back transfers of zero and positive transfers with positive back transfers. With this simple reciprocity heuristic, the number of correct binary guesses over all 54 second movers would have been 39. 10 raters were significantly below this number, and none of them were significantly above (SI, Table S2).

These results support our earlier analyses. They show that several raters were able to use explicit information about first mover behaviour to achieve some significant degree of accuracy when drawing inferences about second mover behaviour. A number of additional raters, however, were significantly less accurate than they could have been had they simply restricted attention to first mover behaviour and assumed that second movers reciprocate. This reduced level of accuracy presumably occurred because the raters in question were paying attention to information in the photographs that they could not use effectively. Importantly, we paid raters for accurate guesses (see Methods and SI). Although our incentive scheme was not based on the binary measure of accuracy we have derived here, raters were paid more on average for accurate guesses. Because so many raters had a binary accuracy rate below that allowed by a simple reciprocity heuristic, simply ignoring the photos and adhering to the heuristic might have allowed some raters to improve their performance and earn more money.

Discussion

Our results show that some raters made accurate inferences about the choices of second movers in a sequential social dilemma. They did so by using information about first mover behaviour. Raters could not on balance make accurate inferences by viewing photographs of second mover faces. Although some raters apparently used both the facial structure and attractiveness of second movers as cues, our data suggest they did so, if anything, to their detriment. Inferential accuracy has clear limits. The limits we found are entirely consistent with the dynamical instability of green-bearded typecasting, a dynamical instability that should militate against reliable markers of otherwise unobservable behavioural tendencies in social dilemmas.

Though not immediately obvious, the limits to inferential accuracy we found are potentially consistent with recent empirical findings on facial width in men. Recent findings have shown an association between wide faces, aggression, and dishonesty[15–17]. The social interactions in these studies were not experimental social dilemmas in the game theoretic sense[3,4]. We, in contrast, used an experimental game that is a social dilemma in this sense. Selection for reliable markers of behavioural tendencies can vary across strategic domains, with selection yielding reliable markers in some domains[29,30] but not others[2,3,31]. As a consequence, the link between aggression, dishonesty, and wide faces found in some studies could be compatible with the absence of a link between defection and wide faces in our study; the behavioural domains of the studies are different. Furthermore, the association between wide faces and aggression apparently does not hold in all populations[32]. Regardless, our specific results on facial width and second mover behaviour are at odds with one recent study showing that men with wide faces are relatively untrustworthy in a trust game[18]. For the moment this latter inconsistency remains a puzzle. Importantly, however, this study did not analyse the accuracy of rater inferences, and in this sense it addressed a question different from our main concern here.

Our primary interest concerns the accuracy of inferences about others in social dilemmas. We found accuracy associated with first mover behaviour but not with second mover faces. Other studies, in apparent contrast, have uncovered accurate inferences arising from brief exposure to the mannerisms, expressions, and faces of others[16,33,34]. We believe these differences with respect to our study are easily explained. For example, one study[16] identified an ability to accurately infer how aggressive others are, but again aggression is not the same as defection in a social dilemma. In domains more closely related to our own, researchers have found accurate inferences with respect to economic games similar to our trust game[33,34]. In one case[33], raters could accurately infer the choices of others in a prisoner's dilemma by viewing photographs taken at the moment a player made her decision. Photographs taken at other times, as in our study, did not lead to accurate inferences. Inferential accuracy based on markers observed at the moment behaviour is revealed is completely different from inferential accuracy based on markers observed before behaviour is revealed. In the former case, the markers are necessarily irrelevant in terms of the consequences that follow from the social interaction; in the latter case, they may or may not be relevant depending on what they reveal and how they are used by others.

In another study[34], raters achieved some degree of accuracy when guessing the choices of players in a dictator game, a game which provides a simple measure of altruism. Importantly, this study used 20-second videos of the dictators. These videos presumably provided much more information than the still photographs we used, and this additional information could be responsible for the accuracy of the raters. In any case, all of the studies on accuracy described here analysed inferences aggregated over raters in some way. The accuracy they identified, consequently, shows us how social groups can aggregate noisy information to yield accuracy that emerges at the group level. An evolutionary approach, however, requires an individual-level analysis insofar as selection in evolutionary systems is often strongest at or below the level of the individual organism[2]. Under individual selection, what matters with respect to selection is accuracy as experienced by individuals. Average inferences, in contrast, represent a group-level variable. Of particular importance, average inferences have a degree of accuracy that is at least as good as and typically better than the accuracy of individuals[35]. Thus, an analysis demonstrating the accuracy of inferences averaged over raters shows that guesses are systematic in some way, but it does not guarantee that accuracy at the individual level is sufficient to be statistically significant or evolutionarily meaningful.

We did not find any accuracy at the individual level associated with the use of photographs. To check any effects associated with aggregation, we also correlated the accuracy of average guesses about back transfers with actual back transfers. Over all 54 second movers, the correlation is large and highly significant (Pearson's product-moment correlation, 0.512, $P < 0.001$). Over the 41 second movers who were trusted, it is small and not significant (Pearson's product-moment correlation, 0.061, $P = 0.704$). Aggregation, in our case, does not generate accuracy that we cannot already detect at the individual level. At the individual level, photos did not lead to accurate inferences, while knowledge of first mover choices did. Although our data do not allow us to draw precise conclusions about why knowledge of first mover choices was important, we suspect it was because second movers and raters had some common understanding of reciprocity. Specifically, second movers who were trusted sent back more to their partners, and at least some of the raters made guesses reflecting the same pattern. Still photos did not lead to a further improvement in accuracy. Importantly, other forms of contact could. Short videos of a person[34], for example, represent more information than a still photo, but they still represent an intriguingly small amount of information. In the end, we should consider a continuum of information that one person can have about another. At one end of the continuum, we have only a brief glance at a person's face. At the other end of the continuum, we can imagine two people who have known each other intimately for many years. Somewhere along this continuum, inferences should become detectably accurate at the individual level. The ultimate task, in terms of assortment in social dilemmas, is to determine just how much information is required. Our results suggest that truly cursory contact is not enough. Because our methods, however, put us squarely at the low-information end of the continuum, our results still leave considerable room for accurate inferences based on relatively limited information.

Methods

We implemented a modified trust game with male participants in Munich. We modified the standard trust game[23] to maximise the potential to detect accurate typecasting (SI). In particular, both the first mover and the second mover had endowments of nine Euros. The first mover began the game by deciding whether to transfer his entire endowment. This means the first mover's choice was unambiguous; he either trusted the second mover or not, and so he was either in a position to be exploited by the second mover or not. If the first mover transferred his endowment, the transfer was doubled, leaving the second mover with 18 Euros plus his original endowment of nine Euros.

After learning the first mover's choice, the second mover could back transfer any amount between zero and nine Euros in one-Euro increments. Back transfers were also doubled. Importantly, because each second mover had an endowment, which is not typical of trust games[23], he could back transfer a positive amount even if not trusted by his partner. As a result, all second movers had real choices to make because their set of feasible actions did not depend on first mover behaviour. This feature allowed us to identify the accuracy of inferences about second mover behaviour for both second movers who were trusted and for the entire sample of second movers. Moreover, because second movers could back transfer any amount between zero and nine Euros, we were able to characterize accuracy in the different ways described in the Results section. This, in turn, allowed us to check the robustness of our conclusions by addressing accuracy under different operational definitions. Altogether, 67 pairs of men played the game. Raters, however, made guesses about the second movers from 54 of these pairs. In particular, we wanted to homogenise the sample of second movers in terms of age and ethnicity to avoid having to control for these variables with only a few observations responsible for the variation present. We did so in a strictly ex ante fashion before analysing any data from the trust game or conducting any sessions with raters (SI).

For the second task in Konstanz, 13 women and 15 men viewed photographs of the second movers in Munich and made guesses about their choices in the trust game. For each second mover shown, we also informed the raters if the second mover in the photo was trusted by his partner. The choices of both the players and the raters were fully incentivized. In particular, each player received a payment based on both his own choice and the choice of his partner. We paid raters for accurate guesses using a scheme that ensured they would earn more money on average for accurate guesses, but it also preserved the anonymity of choices made by individual players (SI). We also conducted an additional session in Konstanz in which 15 women and 13 men evaluated the attractiveness of the second movers (SI). We used these evaluations to see if the actual behaviour of second movers or rater guesses about second mover behaviour varied in relation to second mover attractiveness. Finally, we used the photographs of second movers to measure the width-to-height ratios of their faces (SI), and this allowed us to identify any relationships between facial width, actual behaviour, and perceived behavioural tendencies.

Further methodological details associated with both parts of the study are available in the Supplementary Information. Our study was approved by the Human Subjects Committee of the Faculty of Economics, Business Administration and Information Technology at the University of Zurich and by the Office of Data Protection and Privacy at the University of Munich (i.e. Datenschutzbeauftragte der Universität München).

1. Hamilton, W. D. The genetical evolution of social behavior. *J. of Theoretical Biology* **7**, 1–52 (1964).
2. Dawkins, R. *The Selfish Gene* (Oxford University Press, Oxford, UK, 1976).
3. Henrich, J. Cultural group selection, coevolutionary processes and large-scale cooperation. *J. of Economic Behavior and Organization* **53**, 3–35 (2004).
4. Bowles, S. & Gintis, H. *A Cooperative Species* (Princeton University Press, Princeton, NJ, 2011).
5. Fehr, E. & Fischbacher, U. The nature of human altruism. *Nature* **425**, 785–791 (2003).
6. Richerson, P. J. & Boyd, R. *Not By Genes Alone* (University of Chicago Press, Chicago, 2005).
7. Silk, J. B., Brosnan, S. F., Vonk, J., Henrich, J., Povinelli, D. J., Richardson, A. S., Lambeth, S. P., Mascaro, J. & Schapiro, S. J. Chimpanzees are indifferent to the welfare of unrelated group members. *Nature* **437**, 1357–1359 (2005).
8. Jensen, K., Call, J. & Tomasello, M. Chimpanzees are rational maximizers in an ultimatum game. *Science* **318**, 107–109 (2007).
9. Vonk, J., Brosnan, S. F., Silk, J. B., Henrich, J., Richardson, A. S., Lambeth, S. P., Schapiro, S. J. & Povinelli, D. J. Chimpanzees do not take advantage of very low cost opportunities to deliver food to unrelated group members. *Animal Behaviour* **75**, 1757–1770 (2008).
10. Nowak, M. A. Five rules for the evolution of cooperation. *Science* **314**, 1560–1563 (2006).
11. Pennisi, E. How did cooperative behavior evolve? *Science* **309**, 93 (2005).
12. Keller, L. & Ross, K. G. Selfish genes: a green beard in the red fire ant. *Nature* **394**, 573–575 (1998).
13. Queller, D. C., Ponte, E., Bozzaro, S. & Strassmann, J. E. Single-gene green-beard effects in the social amoeba *Dictyostelium discoideum*. *Science* **299**, 105–106 (2003).
14. Smukalla, S., Caldara, M., Pochet, N., Beauvais, A., Guadagnini, S., Yan, C., Vinces, M. D., Jansen, A., Prevost, M. C., Latgé, J.-P., Fink, G. R., Foster, K. R. & Verstrepen, K. J. *FLO1* is a variable green beard gene that drives biofilm-like cooperation in budding yeast.*Cell* **135**, 726–737 (2008).
15. Carré, J. M. & McCormick, C. M. In your face: facial metrics predict aggressive behaviour in the laboratory and in varsity and professional hockey players. *Proc. of the Royal Society B* **275**, 2651–2656 (2008).
16. Carré, J. M., McCormick, C. M. & Mondloch, C. J. Facial structure is a reliable cue of aggressive behavior. *Psychological Science* **20**, 1194–1198 (2009).
17. Haselhuhn, M. P. & Wong, E. M. Bad to the bone: facial structure predicts unethical behaviour. *Proc. of the Royal Society B* **279**, 571–576 (2012).
18. Stirrat, M. & Perrett, D. I. Valid facial cues to cooperation and trust. *Psychological Science* **21**, 349-354.
19. Penton-Voak, I. S., Pound, N., Little, A. C. & Perrett, D. I. Personality judgments from natural and composite facial images: more evidence for a "kernel of truth" in social perception. *Social Cognition* **24**, 607-640 (2006).
20. Little, A. C. & Perrett, D. I. Using composite images to assess accuracy in personality attribution to faces. *British Journal of Psychology* **98**, 111-126 (2007).
21. Dovidio, J. F., Glick, P. & Rudman, L. A., eds. *On the Nature of Prejudice* (Blackwell Publishing, Oxford, UK, 2005).
22. Willis, J. & Todorov, A. First impressions: making up your mind after a 100-Ms exposure to a face. *Psychological Science* **17**, 592–598 (2006).
23. Berg, J., Dickhaut, J. & McCabe, K. Trust, reciprocity, and social history. *Games and Economic Behavior* **10**, 122–142 (1995).
24. Burnham, K. P. & Anderson, D. R. *Model Selection and Multi-Model Inference.* (Springer-Verlag, New York, 2002), 2nd edn.
25. Verbeek, M. *A Guide to Modern Econometrics.* (John Wiley & Sons, West Sussex, UK, 2008), 3rd edn.
26. Fehr, E. & Fischbacher, U. Social norms and human cooperation. *Trends in Cognitive Sciences* **8**, 185–190 (2004).
27. Haley, K. J. & Fessler, D. M. T. Nobody's watching? Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior* **26**, 245–256 (2005).
28. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. of the Royal Statistical Society, Series B* **57**, 289–300 (1995).
29. McElreath, R., Boyd, R. & Richerson, P. J. Shared norms and the evolution of ethnic markers. *Current Anthropology* **44**, 122–129 (2003).
30. Efferson, C., Lalive, R. & Fehr, E. The coevolution of cultural groups and ingroup favoritism. *Science* **321**, 1844–1849 (2008).
31. Andrews, P. W. The influence of postreliance detection on the deceptive efficacy of dishonest signals of intent: understanding facial clues to deceit as the outcome of signaling tradeoffs. *Evolution and Human Behavior* **23**, 103–121 (2002).

32. Oezener, B. Facial width-to-height ratio in a Turkish population is not sexually dimorphic and is unrelated to aggressive behavior. *Evolution and Human Behavior* **33**, 169–173 (2012).
33. Verplaetse, J., Vanneste, S. & Braeckman, J. You can judge a book by its cover: the sequel. A kernel of truth in predictive cheating detection. *Evolution and Human Behavior* **28**, 260–271 (2008).
34. Fetchenhauer, D., Groothuis, T. & Pradel, J. Not only states but traits – humans can identify permanent altruistic dispositions in 20 s. *Evolution and Human Behavior* **31**, 80–86 (2010).
35. Page, S. E. *The Difference* (Princeton University Press, Princeton, NJ, 2007).
36. Akaike, H. Information theory as an extension of the maximum likelihood principle. In Petrov, B. N. & Csaki, F. (eds.) *Second International Symposium on Information Theory*, 267–281 (Akademiai Kiado, Budapest, 1973).

Acknowledgments

We would like to acknowledge the generous support of the Swiss National Science Foundation, Grant No. 100014-130127/1 entitled "The Social Dynamics of Normative Behavior – Population Fragmentation and Divergent Cultural Evolution". We also thank the Netherlands Organization for Scientific Research for a Rubicon Grant to S.V. (446-07-024) and the University of Zurich Research Priority Program, "Foundations of Human Social Behavior – Altruism versus Egoism". Finally, we extend our greetings and our thanks to all the folks at MELESSA in Munich and the Lakelab in Konstanz for letting us move in for a few days.

Author contributions

C.E. and S.V. designed and conducted the experiment. C.E. analysed the data. C.E. and S.V. wrote the paper.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Efferson, C. & Vogt, S. Viewing men's faces does not lead to accurate predictions of trustworthiness. *Sci. Rep.* **3**, 1047; DOI:10.1038/srep01047 (2013).