Title:

Reciprocity of social influence

Authors:

Ali Mahmoodi[1,2,*]

Bahador Bahrami [3, 4,5]

Carsten Mehring[1,2,5]

Affiliation:

[1] Bernstein Centre Freiburg, University of Freiburg, Hansastrasse 9a, 79104, Freiburg, Germany

[2] Faculty of Biology, University of Freiburg, Freiburg, Germany

[3] Institute of Cognitive Neuroscience, University College London, 17 Queen Square London, WC1N 3AR, United Kingdom

[4] Faculty of Psychology and Educational Sciences, Ludwig Maximilian University, Munich, Germany

[5] These authors contributed equally to this work

[*] Corresponding author

Abstract:

Humans seek advice, via social interaction, to improve their decisions. Social interaction is, very often, reciprocal. However, the role of reciprocity in social influence is unknown. Here we tested the hypothesis that our influence on others affects how much we are influenced by them. In our experiments, participants first made a visual perceptual estimate and then shared their estimate with an alleged partner. Then, in alternating trials, the participant either revised their decisions or observed how the partner revised theirs. We systematically manipulated the partner's susceptibility to influence from the participant. Our results show that people reciprocated influence with their partner by gravitating towards the susceptible (but not towards the insusceptible) partner's opinion. Then in separate experiments, we showed that reciprocity is (a) a dynamic process and (b) disappeared when people believed that they interacted with a computer. Reciprocal social influence, therefore, is a signaling medium for human-to-human communication that goes beyond aggregation of evidence for decision improvement.

Keywords:

Social influence, decision making, reciprocity

Introduction:

When we are uncertain we look for a second opinion and those opinions often change our decisions and preferences[1–5]. Moreover, social influence is not restricted to difficult or critical decisions: evaluations of the comments in the news media are affected by previous scores of the content[6] and risk preferences alter after observing other people's choices[7]. Social influence extends to perceptual judgment[8–10] and long term memory[11]. Social information can help improve decision accuracy[12], outcome value[13] and evaluative judgement [14–16]. On the other hand, social influence can also lead to catastrophic outcomes. Social influence causes information cascades[17] urging individuals to ignore their own accurate information in favor of the cascaded falsehoods. In some cases, groups are less biased when their individuals resist social influence[18]. Social influence can undermine group diversity[19] leading to disastrous phenomena such as market bubble[20], rich-get-richer dynamics[21], and zealotry[22].

Humans tend to reciprocate in social interaction[23–25]. We react to kindness with kindness[24], respect with respect and hostility with hostility[23]. Smiling staff get more tips[26]. Violators of trust in Trust Games[27,28], and free riders (i.e. those who do not contribute) in Public Good Games are punished[29]. Since social influence is, by definition, mediated via social interaction, one may wonder if reciprocity extends to social influence itself. However, despite a wealth of research on reciprocal behavior, to our knowledge, no study has examined the existence of reciprocity in how we receive and inflict influence on others[30,31].

Several studies in social decision making in humans have shown a sensible correspondence between the reliability of social information (e.g. advice) and the extent to which that advice is assimilated in decisions and preferences. For example human agents integrate their own choice with that of an advisor by optimally tracking the trustworthiness of the advisor[32] and are able to track the expertise of several agents concurrently[33]. Social information is integrated into value and confidence judgement based on its reliability[34] or credibility[35]. People can integrate information from themselves and others based on their confidence[9]. These studies offer strong evidence for what previous works in social psychology have called "informational conformity"[36] which is based on the assumption that others can have access to information that will help the agent achieve better accuracy. The prediction drawn from this informational account is that social influence should hinge on the reliability of the source and quality of the advice but not on conventions and norms such as reciprocity. If we could benefit from others' (critical) opinion and the accuracy of our opinion is our only concern, then we should welcome reliable advice irrespective of the advisor's attitude towards our opinion.

On the other hand, numerous studies indicate that people perform less than ideally when using social information[10,37–39]. When aggregating individual and social information about a perceptual decision, humans follow a simplifying heuristic dubbed equality bias: the tendency to allow everyone equal say in a collective decision irrespective of their differential accuracy or expertise[38,39]. The universal prevalence of the equality bias is important because it shows that even though humans do have the cognitive and computational capacity to track the trustworthiness[32], reliability[33] and credibility[35] of others, they still choose to employ a simple heuristic. Other studies have shown that in a different set of experimental conditions, people show an egocentric bias by relying on their own individual information more than they should[40,41]. These findings suggest that normative concerns such as equality and maintaining a good self-image may also play an important role in interactive decision-making. These factors are not consistent with the Bayesian theory of social information aggregation[32,34,35] which requires that the influence that people take from others should not depend on norms and conventions. An empirical observation of reciprocity in advice taking would be inconsistent with the Bayesian theory of social influence[32,34,35].

To summarize, aligning with other people's choices by taking their advice could be motivated informationally[42] to increase accuracy, or normatively[43] to affiliate with others and maintain positive self-esteem. Following other people's advice often leads to more accurate decisions[9,44,45]. Alignment can contribute to a positive self-image[46] and is used as a compensatory tool among minorities[47]. Being ignored in a virtual game can damage one's self-image[48]. Social exclusion has negative consequences on the excluded[49–51]. We hypothesised that people would reciprocate influence with others because reciprocity is a pervasive social norm[52]. If one breaks the norm of reciprocity, one should expect to be punished[53,54], for example by being ignored. Therefore, participants would reciprocate with a reciprocating partner in order to maintain influence over them and avoid the *negative experience of losing* influence[49–51]. On the other hand, for a non-reciprocating partner who violates the norm, participants may have a desire to punish the partner by ignoring their opinion. We tested these hypotheses by investigating if participants in a social decision making task take less advice from *insusceptible* partners, i.e. those who don't take advice from the participant. Conversely, we also tested if participants take more advice from *susceptible* partners, i.e. those who are influenced by the participant's suggestions.

We adopted and modified an experimental perceptual task inspired by recent work on aggregation of social and individual information[10]. Participants estimated the location of a visual target on a computer screen. Then they saw the estimate of their partner about the location of the same target. The participant did not know that this partner's opinion was, in reality, generated by sampling randomly from a

distribution centered on the correct answer. After the two initial estimates were disclosed, the participant or the partner was allowed to revise their estimate. An algorithm generated the partner's revised estimate simulating susceptible or insusceptible partners. Experiment 1 showed that, participants took more advice from the partner who took more advice from the participant. Experiment 2A and 2B investigated the dynamics of reciprocity and whether reciprocity depends critically on whether we believe it changes the partner's state of mind. Participants thought they worked with either a human or a computer partner and reciprocity disappeared when subjects believed they were working with a computer partner. Finally, our results also showed that reciprocity had a profound impact on participants' evaluation of their own performance which was lower when working with a non-reciprocating partner.

Results:

In experiment 1, 20 participants were recruited one at a time and told that they would cooperate with three partners who were participating in the same experiment simultaneously in other laboratory rooms connected via internet. In reality, each participant was coupled with a computer algorithm. The algorithm generated three distinct behavioral profiles, corresponding to three experimental conditions (see below) which were administered in a block design in counterbalanced order. Participants were not informed about this arrangement. In each trial, the participant made a perceptual estimate about the location of a target on the screen (Figure 1). After stating her initial estimate, the participant saw the opinion of the partner about the same stimulus. Next, the participant either revised her estimate or observed the partners revise theirs. Participants were required to put their second estimate between their own first estimate and that of their partner's. The acceptable range included staying on their first estimate or moving all the way to their partner's first estimate. Using this constraint, we assured that the amount of change made in the second stage is solely due to observing the partner's choice and not because of a change of mind[55]. The three partners differed in their susceptibility to taking influence from the participant. In the baseline condition, the participant always made the second estimate and the partner never contributed a second estimate. Hence, participants were not able to observe the susceptibility of the baseline partner. In the susceptible condition, the partner was influenced strongly by the participant and revised her initial estimate by conspicuously gravitating towards the participant's estimate. Vice versa, in the insusceptible condition, the partner more or less ignored the participant's opinion. The partner's initial estimate was generated identically in all three conditions by sampling randomly from a distribution centered on the correct answer.

We computed the influence that participants took from their partner as the ratio of the angular displacement (in radians) between their initial and final estimate toward their partner's estimate divided by their initial angular distance from their partner (Figure 1B). Overall, participants were influenced by their partners' opinions (mean influence ± std dev = .36±.11; Wilcoxon sign rank test vs zero, Z = 3.62 p = .0002). Consistent with previous studies[32,35,38], this indicates that our participants did use social information (the partners' choices) to improve their decisions (mean ± std dev error after first estimate 67±7 radians and after second estimate 64±6 radians, Wilcoxon sign rank test, Z = 3 p =.002). To test our main hypothesis, we asked whether revised opinions were more influenced by the susceptible than the insusceptible (Figure 2A) and baseline (Figure 2B) partners. The difference between the average influence in the susceptible and the insusceptible condition, which is a measure of reciprocity, was significantly larger than zero (Figure 2A-C, Wilcoxon sign rank test, Z = 3.33 p = .002 after Bonferroni correction). Similarly, the influence from the susceptible partner was larger than the influence in the baseline condition (Figure 2B-C, Wilcoxon sign rank test, Z = 2.34 p = .03 after Bonferroni correction). The difference between insusceptible and baseline was not significant (Wilcoxon sign rank test, Z = 1.11, p=.26).

The three different partners' initial estimates were produced from an identical generative process using exactly the same distribution and therefore ensuring that the partners' accuracies were perfectly controlled across conditions. However, one might argue that the observed result may be due to the difference in *perceived* accuracy of the partner. Indeed, we may think more highly of those who confirm our decisions more often and, subsequently, take our (misguided) assessment of their competence as grounds for integrating their estimate into our own revised opinion. To test this hypothesis directly, at the end of each experiment, we asked the participants to rate the precision of the partners they interacted with in the experiment, how much they liked different partners, and their own performance as well. A mixed effect model showed that perceived precision of the partners, the actual precision of the partners, the participants' actual precision in different conditions and the liking of the partner did not have any effect on reciprocity (see Supplementary Material). The performance ratings of the partner did not differ across conditions (Fig 2D, repeated measures ANOVA, F(2.49,39) = 1.07 p = .36). In each trial, after participants registered their estimates, they were required to report their confidence about their estimates using a scale from 1 to 6. Employing mixed effect models, we showed that condition (baseline, susceptible, or insusceptible) had a significant effect on influence even if confidence was included as a potential confound (see Supplementary Material).

At the end of the experiment we asked participants to rate how much they liked their three partners on a scale of 1 to 10. The mean score ± std dev was 7.64±.7 for the susceptible partner, 5.94±2.53 for the insusceptible partner, and 7.52±1.41 for the baseline partner. Our data shows that people liked the susceptible partner over the insusceptible one (Wilcoxon sign rank test Z = 2.62 P = .008). There was no difference between susceptible and baseline partners (Wilcoxon sign rank test Z = -.34 P = .72) and the p-value for the difference between the baseline and the insusceptible partner was only around the threshold (Wilcoxon sign rank test Z = -1.91 P = .056).

The results of our first experiment show that human participants were more influenced by partners which were reciprocally more influenced by the participants. We next asked whether human participants change their advice taking strategy in response to a change in the advice taking strategy of a partner over time. To answer this question, we carried out experiment 2A using the same paradigm as in experiment 1 but with an important modification. The participants were told that they are working with the *same* partner during the entire experiment. The partner's strategy changed across time: in one part of the experiment the partner was susceptible, in the other part she was not. Between the two conditions of the experiment, there was a smooth transition (from susceptible to insusceptible or vice versa) and the order of the conditions was counterbalanced across participants (see Figure S1).

We calculated participants' trial-by-trial influence in the two conditions. Replicating experiment 1, reciprocity, again defined as influence in the susceptible condition minus influence in the insusceptible condition, was significantly larger than zero (Figure 3A, Wilcoxon sign rank test, Z = 3.54 p = .0003). A mixed effect model showed that condition (susceptible or insusceptible) had a significant effect on influence even if confidence was included as a potential confound (see Supplementary Material). Again, using a mixed effect model, we showed that the change of influence across conditions cannot be explained by a change in perceived performance of self or partner (see Supplementary Material). An interesting question is whether reciprocity is affected by the condition (susceptible or insusceptible) with which the participant *started* the experiment. To answer this question, we compared the reciprocity for participants who were first exposed to the susceptible partner to participants who started with the insusceptible partner. Our result showed no difference in reciprocity between these two groups (mean±std dev for those who started with reciprocal partner .049±.1 and for those who started with non-reciprocal partner .08±.07 Wilcoxon rank sum test z = -1.07 p =.28).

The present results demonstrate that if a participant observes any change in the amount of influence she has over her partner, she will in return modify the amount of advice she takes from her partner. We,

therefore, hypothesized that our participants exploit reciprocity as a social signal to communicate with their partner. We predicted that participants would not show reciprocity when working with a computer. In other words, reciprocity depends critically on whether we believe it changes the partner's state of mind. To test this prediction, we conducted experiment 2B in which participants were told that they were working with a computer. All other aspects of the experiment remained as in experiment 2A. As predicted, we did not observe reciprocity when participants believed they were working with a computer (Figure 3B-D, Wilcoxon sign rank test, $Z = -1.57$ $p = .11$). In fact, a majority of participants showed the opposite pattern observed in experiment 2A (Figure 3B). A two-way ANOVA with factors condition (susceptible or insusceptible as within subjects factor) and type of partner (believed to be human or computer as between subjects factor) and influence as the dependent variable showed a significant interaction of type of partner and condition ($F(1,58) = 14.8$ $p = .00001$). The effect of condition alone was not significant ($F(1,58) = .49$ $p = .51$) but there was also a significant between-subject effect of type of partner (Figure 3C, $F(1,58) = 4.16$ $p = .04$). A post hoc analysis between reciprocity in human and computer condition confirmed a significant difference in reciprocity between these two conditions (Figure 3C-D, Wilcoxon ranksum test, $Z = 2.97$ $p = .003$). Taken together, these results demonstrate that participants were more influenced when they thought their partner is a computer (mean influence ± std dev 42±.16) compared to when they thought their partner is a human (mean influence ± std dev .34±.13).

We then investigated whether the distance between the participants' and the partners' initial estimates affected the influence that participants took from their partner or the strength of reciprocity (see Supplementary material for details). We did not find a significant effect of distance (mixed ANOVA, $F(2.6,151) = 1.17$ $p = .31$) nor significant interactions between distance and condition (mixed ANOVA, $F(2.6,151) = 1.47$ $p = .22$) or between distance, condition and experiment ($F(2.6,151) = 1$ $p = .38$). The interaction between distance and experiment was only around the significance threshold ($F(2.6,151) = 2.7$ $p = .05$). We also did not find a significant effect of distance on reciprocity in experiment 2A (repeated measures ANOVA, $F(2.61,75) = 1.63$ $p = .16$). Taken together, these results show that the distance between the initial estimates did not affect the influence that participants took from their partner nor the strength of reciprocity.

Finally, we asked how being in the susceptible and insusceptible conditions changed the participants' ratings of their own and their partner's performance. To answer this question, participants were asked to rate their own and their partners' performance on a 1 to 10 scale at the end of each condition, i.e. twice in each experiment 2A and 2B. As there wasn't any difference in performance rating between experiment

2A and 2B, neither for self nor partner (see Supplementary Material), we aggregated the data of the two experiments. Participants' rating of their own performance was significantly lower in the insusceptible than in the susceptible condition (Figure 3E, Wilcoxon sign rank test, $Z = 3.04$ $p = .002$) while participants' rating of their partners' performance remained unaffected by the partners' susceptibility (mean rating ± std dev 6.63±1.58 for susceptible condition, and 6.81±1.59 for insusceptible condition; Wilcoxon sign rank test, $Z = -.9$ $p = .36$).

Discussion:

An important question in human social interaction is how people weigh others' opinion[56]. Bayesian theories recommend that different opinions should be weighted by their reliability in order for the group to benefit from putting the opinions together[57,58]. Indeed, some empirical evidence has supported this view[9,32,34,35] while others have shown other decision aggregation strategies in human social interaction[10,38].

We developed an experimental paradigm inspired by previous work on social information aggregation[10]. We quantified how people weighted their peer's opinion in the context of a visual perceptual task. We tested if this weighting depends on the weight their peers assigned to the participants' opinion. In experiment 1, participants worked with three different alleged human partners in separate blocks. Our participants were more influenced by the susceptible partner compared to the baseline and insusceptible partners. In experiment 2A, the behaviour of a single partner changed dynamically within the same experiment from susceptible to insusceptible or vice versa. We replicated the result of experiment 1 by showing that participants were more influenced in the susceptible condition than in the insusceptible condition. In experiment 2B, we showed that participants did not reciprocate when their peer was a computer even though they took greater influence from the computer's advice.

When combining opinions in an optimal Bayesian way to maximize accuracy, each source of information should be weighted based on its reliability[59,31]. Consequently, reciprocity of social influence, i.e. weighting others' opinion by the weight they give to our opinion is not consistent with Bayesian reliability-based information aggregation. In our experiment, we systematically controlled the accuracy of the participants' partners such that they were identical across experimental conditions. Participants rated their own performance lower in the insusceptible condition than in the susceptible condition but did not distinguish between the partners' accuracies. With such judgement of their performance and that of their partners, a hypothetical Bayesian participant would have taken more influence from the insusceptible partner. This is actually what participants did when working with a computer partner. However, when working with a

human partner, they followed the opposite strategy and took more influence from the susceptible partner.

Why do people go against an information integration strategy that is more likely to maximize their accuracy? We propose that reciprocity is a pervasive social norm[52], and abiding by norms is sometimes rewarding in itself and could hence become a goal[60,61]. As a consequence, individuals may be ready to pay a cost (in terms of reduced accuracy) to adhere to these norms[62]. This explanation is supported by the finding that participants did not reciprocate with a computer partner as participants did not expect the computer to comply with the reciprocity norm.

In the susceptible condition, participants may reciprocate with their reciprocating partners in order to keep their influence over them and avoid the pain of being ignored[51]. As such, showing reciprocity towards a susceptible partner may be driven by a form of loss aversion. Why would people be aversive to losing influence? Recent works have suggested that influence over others may be inherently valuable both behaviorally and neurobiologically[63]. There is now compelling evidence that others' agreement with our opinion is a strong driver of human brain's reward network[1,64]. In the insusceptible condition, on the other hand, ignoring the insusceptible partner may be motivated by wishing to punish someone who does not comply with the norm of reciprocity.

Following the norm of reciprocity might also improve people's self-efficacy. It is possible that in the insusceptible condition players perceive the experiment as a status competition. In this view, ignoring the insusceptible partners in response to being ignored by them could serve as a signal from the participant that s/he is not willing to accept an inferior position[65,66]. Several studies in behavioral economics (ultimatum game in particular) have shown that one reason why people reject unfair offers is because they want to send the signal that they will not be easily dominated and thereby refuse to accept an inferior social status compared to their peer[67]. Indeed, multiple studies have shown that people do not like to be in an inferior position where their choices are less selected than others' and they use various strategies to compete with their peers in having more influence[63,68,69]. However, people do not engage in a status competition with a computer which is consistent with the difference in reciprocity between human and computer experiments (Figure 3D). In addition to the above, ignoring the non-reciprocating partner may also serve to protect the participant's ''wounded pride''[70,71] and maintain their self-esteem. Our finding that participants rate their own performance higher when playing with the susceptible than with the insusceptible partner (Fig. 3E) is consistent with the hypothesis that having influence over others improves self-efficacy. It should be noted, however, that, there is no one-to-one correspondence between the perception of self-efficacy and reciprocity: while the difference in self-efficacy between the

susceptible and insusceptible conditions is the same for experiment 2A and 2B (see Figure S2 in the supplementary material), the difference in influence is not: in experiment 2A, the influence in the insusceptible condition is less than in the susceptible condition but in experiment 2B, the influence in the susceptible condition is identical to the insusceptible condition (Figure 3C, D). Hence, reciprocity cannot be entirely explained by changes in self-efficacy.

Reciprocating influence is consistent with cognitive balance theory[72] which posits that humans change their preference to be similar to those they like and dissimilar to those they do not. In experiment 1, participants liked the susceptible partner more than the insusceptible one and were more influenced by the partner they liked more. However, in our experiment, the participants were not allowed to change their estimate away from their peers (participants were instructed to make their second choice between their and their partner's initial estimates). This restriction makes it difficult to directly address the relationship between cognitive balance theory and reciprocity observed in the current study.

In the insusceptible condition, participants' perceived performance of themselves dropped significantly (Figure 3E). Previous studies show that humans are good at tracking their accuracy even in the absence of any external feedback[9,73]. It's been argued that people are able to get insight into their accuracy through past experience[74]. However, in social contexts their judgement could be affected by the environment[75,76] depending on whether they compete or cooperate with a peer[77]. Our performance rating results confirm the effects of social context on human performance monitoring. This finding shows that being ignored exerts a devastating impact on self-efficacy. One possibility is that being repeatedly ignored in the insusceptible condition may induce a negative emotional impression on the participant that impairs the participant's self-evaluation. Another possibility is that participants may interpret the partner's revised estimate as the correct position of the target. By definition, the insusceptible partner's revised estimates would fall further from those of the participant. The inevitable conclusion for the ignored participants would be that their opinion must have been less precise in the insusceptible blocks. Future studies could investigate each of these potential explanations.

Previously we showed that when working together in a dyad[38] people tend to operate by an "equality bias" giving equal weight to their own and their partner's decision. Participants fulfilled this goal either by adjusting the weight they assign to each other's opinions[38] or by matching their confidence to the confidence of the other people they worked with[68]. Hence, in both cases, people mutually adapted to each other's behavior when required to make decisions together. Similarly, participants exhibited mutual adaptation of social influence in the present study.

Our results imply that humans do not only consider others' reliability to compute the weight that they assign to others' opinion, but instead they take into account other factors like reciprocity as well. We conclude that reciprocity plays a significant role in human advice taking and social influence which violates the optimal account of human information integration. Reciprocity as a social norm helps people to fulfil objectives of social interaction including maintaining a positive self-image.

Methods:

80 healthy adult participants (39 females, mean age ± std:25 ± 2.9) participated in three experiments after having given written informed consent. Each participant participated in only one of the experiments. Participants were students at the University of Freiburg, Germany. The experimental procedures were approved by the ethics committee of the University of Freiburg. All experiments were performed using Psychophysics Toolbox[78] implemented in MATLAB (Mathworks). The data were analyzed using MATLAB and SPSS.

Experimental task:

Experiment 1: This experiment was designed to investigate whether participants were more influenced by whom they influenced more. Participants first made a perceptual estimate about the location of a target on the screen. Afterwards they were presented with the estimate from their partner regarding the same stimulus. This was followed by making a second choice about the location of the target or observing a second choice of their partner.

In more detail, the experiment went on as follows: participants (N=20, 9 females, mean age ± SD, 25±2.8) were presented with a sequence of 91 visual stimuli consisting of small circular Gaussian blobs (r = 5mm) in rapid serial visual presentation on the screen (resolution = 2560 ×1440 Dell U2713HM 27"). The first item was presented for 30ms and every other stimulus was presented for 15ms each. Participants' task was to identify the location of the first stimulus. Participants were required to wait until the presentation of all stimuli were finished, and then indicate the location of the target stimulus using the computer mouse (Figure 1). The reported location was marked by a yellow dot. After participants reported their initial estimate, they were required to report their confidence about their estimate on a numerical scale from 1 (low confidence) to 6 (high confidence). Afterwards, participants were shown the choice of their partners about the same stimulus (see below, Constructing Partners Section for further details) by a small dot on the screen. Then, either the participant revised her estimate or observed the partner revise theirs. After the second estimate was made, all estimates were presented to the participant for 3 seconds. In this stage,

the first choice was shown by a hexagon to be distinguished from the second choice which was shown by a circle (Figure 1). There was not any time pressure on participants in any stage of the experiment and the experiment did not move to next stage until the participants had registered their responses (Figure 1). Participants were told that their payoff will be calculated based on their first and second estimates. However, everyone was given a fixed amount at the end of the experiment.

During the course of the experiment participants were exposed to three different partners. Partners varied in their susceptibility to the participants' estimates (i.e. the amount of influence the participants' first estimate has on the partner's second estimate). In the baseline blocks, the participant always made the second estimate and the partner never contributed a second estimate. In the susceptible and insusceptible blocks, the partner and the participants made the second estimate in the odd and even trials, respectively. In the susceptible block the partner was influenced strongly by the participant and vice versa in the insusceptible block (see below for details in section 'Constructing Partners'). In each block, the participants worked with only one partner and each block contained 30 trials. Participants worked with each partner for 5 blocks. For example, they worked with baseline partner in block 1, then with the susceptible partner in block 2, then with insusceptible partner in block 3, then again with the baseline partner in block 4 and so on. Participants completed 15 blocks in total. The order of the partners was randomized across participants. The three partners were shown by blue, red, and turquoise markers which were randomly assigned to the different partners at the beginning of each subject's experiment. After finishing the experiment, the participants were required to estimate their own and the three partners' performance on a numerical scale from 1 to 10. They were instructed to only consider the first choice to assess their partners' performance. At the end, we asked them whether they thought they interacted with real people or with a computer algorithm. All participants indicated that they believed they were interacting with real human partners.

Three participants were excluded from the final analyses of this experiment. One participant did not notice she played against three different partners. The other two participants were excluded because they resampled in the second stage, meaning that their second estimates were not between their own and their partner's initial estimates (contrary to the task instruction). In the Supplementary Material we show that our findings remain valid and statistically significant when these three subjects were not excluded from the analysis.

Experiment 2A: This experiment was designed to test whether reciprocity is a dynamic process. The experiment used the same paradigm as experiment 1 but participants (N = 30, 15 females, mean age ±

SD, 25±2.8) were told that they do the task with only one partner which is the same gender as themselves. The partner changed its susceptibility during the course of the experiment, either from susceptible to insusceptible or vice versa. The experiment consisted of 11 blocks in total. Half of the participants were first probed in the susceptible condition which lasted 5 blocks and then with a transition block in between, they switched to 5 insusceptible blocks. The other half completed the opposite order. The average advice/influence that the partner took from our participants is depicted in Figure S1. The transition block was designed in order to avoid a sudden change of the partner's behavior. During the transition block, the partner's advice taking strategy linearly (see below) switched from susceptible to insusceptible or vice versa.

Debriefing: after each session of the experiment, all participants were debriefed to assess to what extent they believed the cover story. We interviewed them with indirect questions about the cover story and all participants stated that they believed they were working with other human participants in neighboring experimental rooms.

Experiment 2B: This experiment differed from experiment2A in one respect: Participants (N = 30, 15 females, mean age ± SD, 24±3.1) were told that their partner in the experiment is a computer. Any other aspects of the experiment were identical to experiment 2A and they received exactly the same task instructions as in experiment 2A except that the human partner was replaced by a computer.

Performance rating: In experiment 1**,** participants rated their own and the three different partners' performance once at the end of the experiment. Note that a different color identified each partner during the experiment. In experiment 2A and 2B, participants rated their own and their partner's performance at the end of each block. This way, we obtained a pair of performance ratings for the self and the susceptible partner and another pair for the self and the insusceptible partner.

Constructing partners: The error distribution of all partners' first choices was modelled from participants' actual estimation errors during a pilot experiment. Ten participants performed an experiment identical to experiment 1 of the current study. We aggregated errors of all participants (N =10) and fitted the concentration parameter kappa of a von Mises distribution centred to the target, yielding the value kappa = 7.4. Then in each trial we drew the first choice of the partner from this distribution. We speculated that participants' assessment of their partners' performance may be strongly influenced by the few trials with high confidence (confidence level of 5 or 6). To avoid this potential problem, the partner's first choice was

not taken from the von Mises distribution in high confidence trials but randomly drawn from a uniform distribution centred on the participants' choice with a width of +/- 20 degrees.

The second choice of the partner was computed differently for susceptible and insusceptible partners. For experiment 1, the influence that the insusceptible partner took from the participants in each trial was chosen with a probability of .65 from a uniform distribution on the interval [0, .2], with a probability of 0.2 randomly from a uniform distribution on the interval [.3, .7], and with a probability of 0.15 randomly from a uniform distribution on the interval [.7, .9]. For experiment 2, The influence that the insusceptible partner took from the participants was chosen randomly from a uniform distribution on the interval [0, .2]. For the susceptible partner, in all experiments, the influence was chosen with a probability of 0.5 randomly from a uniform distribution on the interval [.7, 1], with a probability of 0.2 randomly from a uniform distribution on the interval [.3, .7], and with a probability of 0.3 randomly from a uniform distribution on the interval [0, .3]. In the transition block, the influence of the partner was a linear interpolation between the susceptible and the insusceptible partner:

$$\mathrm{inf} = (1 - \lambda) \times inf_s + \lambda \times inf_{ins}$$

where inf_s and inf_{ins} were the influences of the susceptible and insusceptible partners, respectively (as explained above). λ gradually increased with time from 0 at the beginning to 1 at the end of the transition block for a transition from the susceptible to the insusceptible condition. For the transition from the insusceptible to the susceptible condition, λ decreased gradually from 1 to 0.

In experiment 1, on average the advice that the partners took from the participants was .3 and .55 in the insusceptible and susceptible conditions, respectively. In experiment 2A, on average the advice that the partner took from the participants was .07 and .5 in the insusceptible and susceptible conditions, respectively. The second choice of the partners in experiment 2B was designed exactly the same as in experiment 2A and the average advice that the partner took in the insusceptible and the susceptible condition was identical to experiment 2A.

Data and code availability: The behavioral data that support the findings of this study and the code which was used to generate the findings and to conduct the experiments of this study will be provided to all readers upon request.

Authors Contributions: A.M., B.B., and C.M. designed the experiments. A.M. collected the data. A.M. carried out the data analysis. A.M., B.B., and C.M. interpreted the results and wrote the manuscript.

Acknowledgments:

This work was supported by a PhD scholarship (AM) from the Graduate School Scholarship Program of the German Academic Exchange Service (DAAD), a European Research Council Starting Grant "NeuroCoDec #309865" (BB), the German Research Foundation (DFG, grant no INST 39/1014-1 FUGG) (CM) and the ''Struktur- und Innovationsfonds Baden-Württemberg (SI-BW)'' of the state of Baden-Württemberg (CM). We thank Ulf Toelch for helping to implement the experimental task and Helena Gavrilova, Tobias Pistohl, and Luke Bashford for helping to collect data.

Declaration of conflict of interest: The authors declare no conflict of interest.

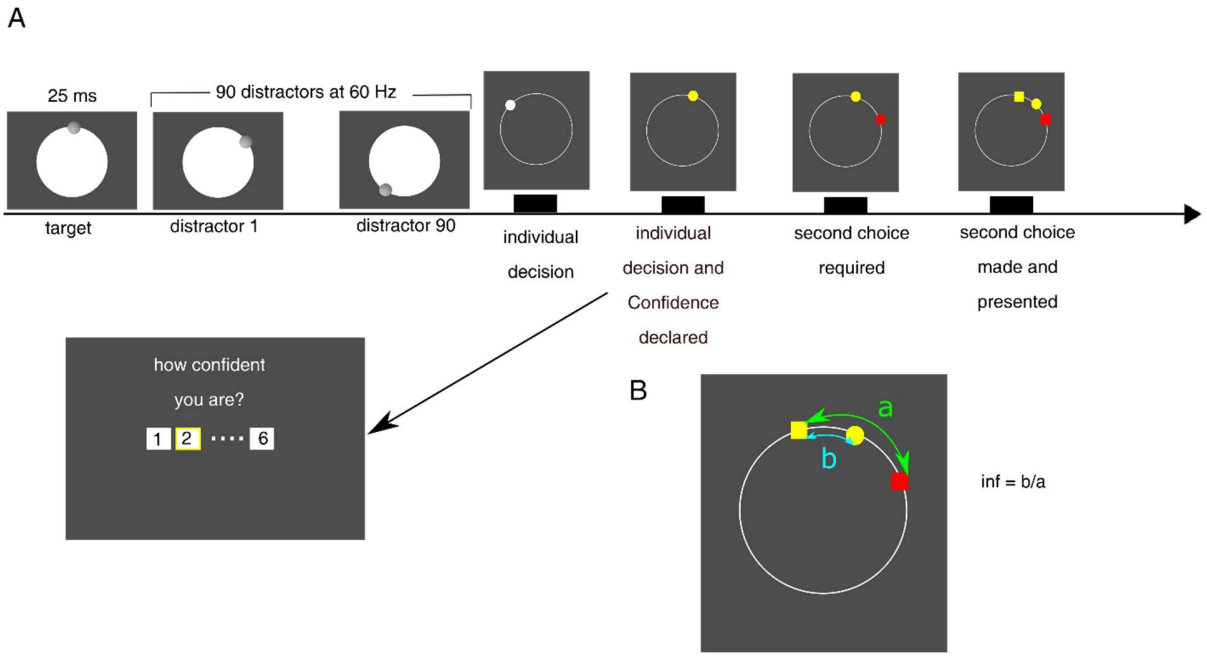

Figure 1: Experimental task: Participants first observed a series of dots on the screen. Participants were required to indicate where they saw the very first dot (yellow dot) and then declare their numerical confidence. After making their individual estimates, they were presented with the estimate of a partner (red dot) concerning the same stimulus. After observing the partner's choice, in some trials the participants and in other trials the partner was given a second chance to revise their initial estimate. Afterwards they were briefly presented with their initial choices and the second choice. In experiment 1 they did the task with three different alleged human partners which only varied in the second choice strategy in different blocks: in the baseline blocks, the participant made all second choices. In the susceptible blocks the partner was very influenced by the participant's first choice, however in the insusceptible blocks, the partner was much less influenced by the participant's first choice compared to susceptible blocks.

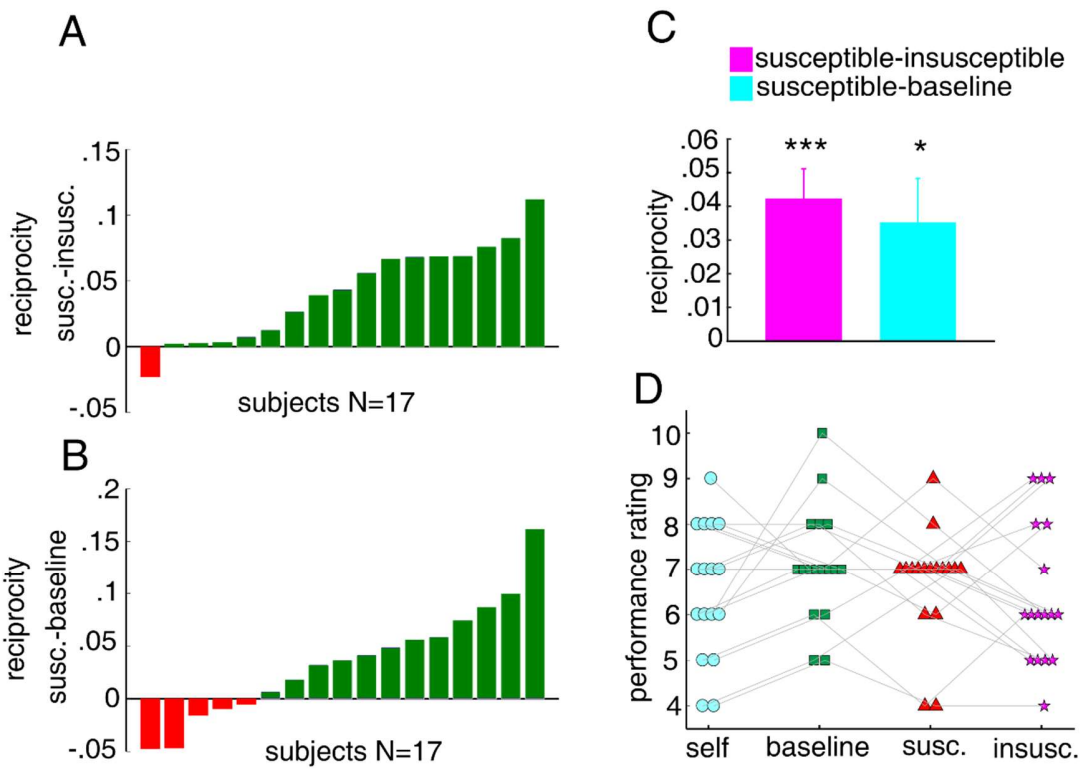

Figure 2: Influence was computed as the angular displacement toward the peer's choice divided by their initial distance from each other's choice. (A) Reciprocity, computed as influence in the susceptible condition minus influence in the insusceptible condition, plotted across participants. (B) Reciprocity, computed as influence in the susceptible condition minus influence in the baseline condition, is plotted across participants. (C) Average reciprocity across participants in insusceptible and baseline conditions compared to the susceptible condition. Standard errors were calculated across subjects. (D) Performance ratings for self and all partners as reported at the end of the experiment.

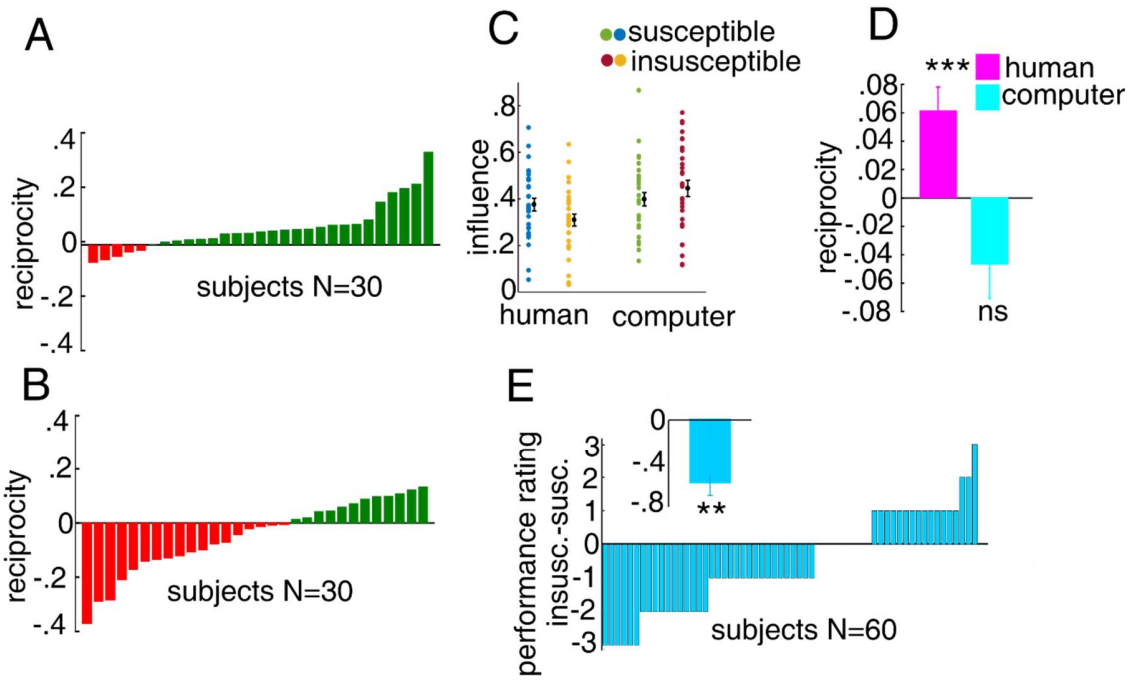

Figure 3: (A) Reciprocity when participants believe they interact with a human partner. (B) Reciprocity when participants believe they interact with a computer partner. (C) Influence for alleged human and computer partners across susceptible and insusceptible conditions. Dots indicate each participant. Error bars show the average and the standard error of the mean. (D) Average reciprocity across participants when participants think they interact with a human or a computer partner. (E) Difference in participants' performance rating for self, insusceptible minus susceptible, plotted for each participant. The inset shows the mean across participants. All standard errors were calculated across subjects.

References:

1. Campbell-Meiklejohn, D. K., Bach, D. R., Roepstorff, A., Dolan, R. J. & Frith, C. D. How the opinion of others affects our valuation of objects. *Curr. Biol.* **20,** 1165–1170 (2010).
2. Berns, G. S., Capra, C. M., Moore, S. & Noussair, C. Neural mechanisms of the influence of popularity on adolescent ratings of music. *Neuroimage* **49,** 2687–2696 (2010).
3. Zaki, J., Schirmer, J. & Mitchell, J. P. Social influence modulates the neural computation of value. *Psychol. Sci.* (2011).
4. Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A. & Fernández, G. Reinforcement learning signal predicts social conformity. *Neuron* **61,** 140–151 (2009).
5. Garvert, M. M., Moutoussis, M., Kurth-Nelson, Z., Behrens, T. E. & Dolan, R. J. Learning-induced plasticity in medial prefrontal cortex predicts preference malleability. *Neuron* **85,** 418–428 (2015).
6. Muchnik, L., Aral, S. & Taylor, S. J. Social influence bias: A randomized experiment. *Science* **341,** 647–651 (2013).
7. Suzuki, S., Jensen, E. L., Bossaerts, P. & O'Doherty, J. P. Behavioral contagion during learning about another agent's risk-preferences acts on the neural representation of decision-risk. *Proc. Natl. Acad. Sci.* **113,** 3755–3760 (2016).
8. Asch, S. E. & Guetzkow, H. Effects of group pressure upon the modification and distortion of judgments. *Groups Leadersh. Men* 222–236 (1951).
9. Bahrami, B. *et al.* Optimally interacting minds. *Science* **329,** 1081–1085 (2010).
10. Toelch, U., Bach, D. R. & Dolan, R. J. The neural underpinnings of an optimal exploitation of social information under uncertainty. *Soc. Cogn. Affect. Neurosci.* **9,** 1746–1753 (2013).
11. Edelson, M., Sharot, T., Dolan, R. J. & Dudai, Y. Following the crowd: brain substrates of long-term memory conformity. *science* **333,** 108–111 (2011).
12. Yaniv, I. Receiving other people's advice: Influence and benefit. *Organ. Behav. Hum. Decis. Process.* **93,** 1–13 (2004).
13. Farrell, S. Social influence benefits the wisdom of individuals in the crowd. *Proc. Natl. Acad. Sci.* **108,** E625–E625 (2011).

14. Harvey, N. & Fischer, I. Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organ. Behav. Hum. Decis. Process.* **70,** 117–133 (1997).
15. Kennedy, J., Kleinmuntz, D. N. & Peecher, M. E. Determinants of the justifiability of performance in ill-structured audit tasks. *J. Account. Res.* **35,** 105–123 (1997).
16. Hartley, P. & Trout, R. Environmental person-organization fit and the importance of promoting organizational policy internally. in *Marketing Dynamism & Sustainability: Things Change, Things Stay the Same…* 71–74 (Springer, 2015).
17. Raafat, R. M., Chater, N. & Frith, C. Herding in humans. *Trends Cogn. Sci.* **13,** 420–428 (2009).
18. Madirolas, G. & de Polavieja, G. G. Improving collective estimations using resistance to social influence. *PLoS Comput. Biol.* **11,** e1004594 (2015).
19. Lorenz, J., Rauhut, H., Schweitzer, F. & Helbing, D. How social influence can undermine the wisdom of crowd effect. *Proc. Natl. Acad. Sci.* **108,** 9020–9025 (2011).
20. Chari, V. V. & Kehoe, P. J. Financial crises as herds: overturning the critiques. *J. Econ. Theory* **119,** 128–150 (2004).
21. Salganik, M. J., Dodds, P. S. & Watts, D. J. Experimental study of inequality and unpredictability in an artificial cultural market. *science* **311,** 854–856 (2006).
22. Festinger, L., Riecken, H. & Schachter, S. *When prophecy fails: A social and psychological study of a modern group that predicted the destruction of the world*. (Lulu Press, Inc, 2017).
23. Fehr, E. & Gächter, S. Fairness and retaliation: The economics of reciprocity. *J. Econ. Perspect.* **14,** 159–181 (2000).
24. Falk, A. & Fischbacher, U. A theory of reciprocity. *Games Econ. Behav.* **54,** 293–315 (2006).
25. Sanfey, A. G. Social decision-making: insights from game theory and neuroscience. *Science* **318,** 598–602 (2007).
26. Tidd, K. L. & Lockard, J. S. Monetary significance of the affiliative smile: A case for reciprocal altruism. *Bull. Psychon. Soc.* **11,** 344–346 (1978).
27. De Quervain, D. J., Fischbacher, U., Treyer, V., Schellhammer, M. & others. The neural basis of altruistic punishment. *Science* **305,** 1254 (2004).
28. King-Casas, B. *et al.* Getting to know you: reputation and trust in a two-person economic exchange. *Science* **308,** 78–83 (2005).

29. Fehr, E. & Gächter, S. Altruistic punishment in humans. *Nature* **415,** 137–140 (2002).
30. Frith, C. D. & Frith, U. Mechanisms of social cognition. *Annu. Rev. Psychol.* **63,** 287–313 (2012).
31. Toelch, U. & Dolan, R. J. Informational and normative influences in conformity from a neurocomputational perspective. *Trends Cogn. Sci.* **19,** 579–589 (2015).
32. Behrens, T. E., Hunt, L. T., Woolrich, M. W. & Rushworth, M. F. Associative learning of social value. *Nature* **456,** 245–249 (2008).
33. Boorman, E. D., O'Doherty, J. P., Adolphs, R. & Rangel, A. The behavioral and neural mechanisms underlying the tracking of expertise. *Neuron* **80,** 1558–1571 (2013).
34. De Martino, B., Bobadilla-Suarez, S., Nouguchi, T., Sharot, T. & Love, B. C. Social Information Is Integrated into Value and Confidence Judgments According to Its Reliability. *J. Neurosci.* **37,** 6066–6074 (2017).
35. Park, S. A., Goïame, S., O'Connor, D. A. & Dreher, J.-C. Integration of individual and social information for decision-making in groups of different sizes. *PLoS Biol.* **15,** e2001958 (2017).
36. Myers, D. Social psychology and the sustainable future. *Soc. Psychol. 11th Ed. McGraw Hill* 586–610 (2013).
37. Soll, J. B. & Larrick, R. P. Strategies for revising judgment: How (and how well) people use others' opinions. *J. Exp. Psychol. Learn. Mem. Cogn.* **35,** 780 (2009).
38. Mahmoodi, A. *et al.* Equality bias impairs collective decision-making across cultures. *Proc. Natl. Acad. Sci.* **112,** 3835–3840 (2015).
39. Hertz, U., Romand-Monnier, M., Kyriakopoulou, K. & Bahrami, B. Social influence protects collective decision making from equality bias. *J. Exp. Psychol. Hum. Percept. Perform.* **42,** 164 (2016).
40. Heyes, C. What's social about social learning? *J. Comp. Psychol.* **126,** 193 (2012).
41. Yaniv, I. & Kleinberger, E. Advice taking in decision making: Egocentric discounting and reputation formation. *Organ. Behav. Hum. Decis. Process.* **83,** 260–281 (2000).
42. Deutsch, M. & Gerard, H. B. A study of normative and informational social influences upon individual judgment. *J. Abnorm. Soc. Psychol.* **51,** 629 (1955).
43. Cialdini, R. B. & Goldstein, N. J. Social influence: Compliance and conformity. *Annu Rev Psychol* **55,** 591–621 (2004).

44. Mackie, D. M. Systematic and nonsystematic processing of majority and minority persuasive communications. *J. Pers. Soc. Psychol.* **53,** 41 (1987).
45. Cialdini, R. B. Science and practice. (2001).
46. Arndt, J., Schimel, J., Greenberg, J. & Pyszczynski, T. The intrinsic self and defensiveness: Evidence that activating the intrinsic self reduces self-handicapping and conformity. *Pers. Soc. Psychol. Bull.* **28,** 671–683 (2002).
47. Tafarodi, R. W., Kang, S.-J. & Milne, A. B. When different becomes similar: Compensatory conformity in bicultural visible minorities. *Pers. Soc. Psychol. Bull.* **28,** 1131–1142 (2002).
48. Williams, K. D., Cheung, C. K. & Choi, W. Cyberostracism: effects of being ignored over the Internet. *J. Pers. Soc. Psychol.* **79,** 748 (2000).
49. Buckley, K. E., Winkel, R. E. & Leary, M. R. Reactions to acceptance and rejection: Effects of level and sequence of relational evaluation. *J. Exp. Soc. Psychol.* **40,** 14–28 (2004).
50. Williams, K. D. Ostracism. *Annu. Rev. Psychol.* **58,** (2007).
51. Eisenberger, N. I., Lieberman, M. D. & Williams, K. D. Does rejection hurt? An fMRI study of social exclusion. *Science* **302,** 290–292 (2003).
52. Gouldner, A. W. The norm of reciprocity: A preliminary statement. *Am. Sociol. Rev.* 161–178 (1960).
53. Cooter, R. Do Good Laws Make Good Citizens? An Economic Analysis of Internalizing Legal Values. (2000).
54. Fehr, E. & Fischbacher, U. Third-party punishment and social norms. *Evol. Hum. Behav.* **25,** 63–87 (2004).
55. Van den Berg, R. *et al.* A common mechanism underlies changes of mind about decisions and confidence. *Elife* **5,** (2016).
56. Bates, J. M. & Granger, C. W. The combination of forecasts. *Or* 451–468 (1969).
57. Nitzan, S. & Paroush, J. Optimal decision rules in uncertain dichotomous choice situations. *Int. Econ. Rev.* 289–297 (1982).
58. Bovens, L. & Hartmann, S. *Bayesian epistemology*. (Oxford University Press on Demand, 2003).
59. Clemen, R. T. & Winkler, R. L. Combining probability distributions from experts in risk analysis. *Risk Anal.* **19,** 187–203 (1999).
60. Axelrod, R. An evolutionary approach to norms. *Am. Polit. Sci. Rev.* **80,** 1095–1111 (1986).

61. Gintis, H. The hitchhiker's guide to altruism: Gene-culture coevolution, and the internalization of norms. *J. Theor. Biol.* **220,** 407–418 (2003).
62. Atran, S. & Ginges, J. Religious and sacred imperatives in human conflict. *Science* **336,** 855–857 (2012).
63. Hertz, U. *et al.* Neural computations underpinning the strategic management of influence in advice giving. *Nat. Commun.* **8,** 2191 (2017).
64. Izuma, K. & Adolphs, R. Social manipulation of preference in the human brain. *Neuron* **78,** 563–573 (2013).
65. Yamagishi, T. *et al.* Rejection of unfair offers in the ultimatum game is no evidence of strong reciprocity. *Proc. Natl. Acad. Sci.* **109,** 20364–20368 (2012).
66. Xiao, E. & Houser, D. Emotion expression in human punishment behavior. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 7398–7401 (2005).
67. Yamagishi, T. *et al.* The private rejection of unfair offers and emotional commitment. *Proc. Natl. Acad. Sci.* **106,** 11520–11523 (2009).
68. Bang, D. *et al.* Confidence matching in group decision-making. *Nat. Hum. Behav.* **1,** 0117 (2017).
69. Mahmoodi, A., Bang, D., Ahmadabadi, M. N. & Bahrami, B. Learning to make collective decisions: the impact of confidence escalation. *PloS One* **8,** e81195 (2013).
70. Burnham, T. C. High-testosterone men reject low ultimatum game offers. *Proc. R. Soc. Lond. B Biol. Sci.* **274,** 2327–2330 (2007).
71. Straub, P. G. & Murnighan, J. K. An experimental investigation of ultimatum games: Information, fairness, expectations, and lowest acceptable offers. *J. Econ. Behav. Organ.* **27,** 345–364 (1995).
72. Heider, F. *The psychology of interpersonal relations*. (Psychology Press, 2013).
73. Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J. & Rees, G. Relating introspective accuracy to individual differences in brain structure. *Science* **329,** 1541–1543 (2010).
74. Bandura, A. Self-efficacy: toward a unifying theory of behavioral change. *Psychol. Rev.* **84,** 191 (1977).
75. Allport, F. H. The group fallacy in relation to social science. *Am. J. Sociol.* **29,** 688–706 (1924).
76. Festinger, L. A theory of social comparison processes. *Hum. Relat.* **7,** 117–140 (1954).

77. Wittmann, M. K. *et al.* Self-other mergence in the frontal cortex during cooperation and competition. *Neuron* **91,** 482–493 (2016).
78. Brainard, D. H. The psychophysics toolbox. *Spat. Vis.* **10,** 433–436 (1997).

Supplementary Material:

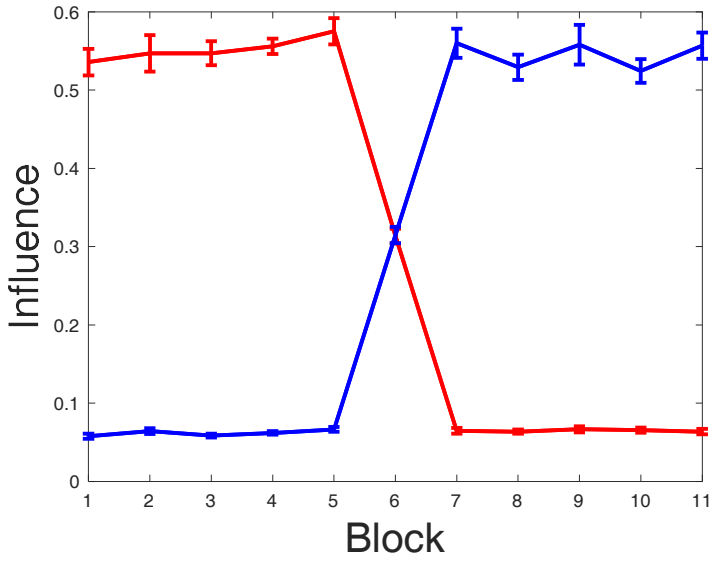

Figure S1: Amount of influence that the virtual partner took from subjects in experiment 2, averaged across all participants. Error bars depict the standard error of the mean across participants.

Effect of actual and perceived performance and liking of the partner on influence in experiment 1: To rule out various potential confounds (perceived and actual precision of partners across conditions, participants' precision and liking of the partner across conditions) we applied the following linear mixed model:

$$I = \beta_{1s} + \beta_2 \times Cond + \beta_3 \times P_{self} + \beta_4 \times P_{partner} + \beta_5 \times PerceivedP_{partner} + \beta_6 \times like_{partner} \quad (1)$$

I is the average influence of the partner on the participant in a specific condition, $Cond$ is the experimental condition (baseline, susceptible, insusceptible), P_{self} is the actual precision of the participant, $P_{partner}$ is the actual precision of the partner, $PerceivedP_{partner}$ is the performance rating, i.e. the perceived performance of the partner, and $like_{partner}$ is how much the participant liked the partner. To account for participant specific baseline influence we let the intercept coefficient vary across participants by including random effects of the form $\beta_{1s} = \beta_{10} + b_{1s}$ where $b_{1s} \sim N(0, \sigma^2)$ and s is the participant id. All other factors were fixed effects.

We found a significant effect of condition (t(45) = 3.59, p = .0007) on influence while the actual precision of the participants (t(45) = -.38 p = .7), the actual partners' precisions (t(45) = -1.89 p = .06), the performance rating for the partners (t(45) = 1.51 p = .13), and the like rating for partners (t(45) = .43 p = .66) had no significant impact on influence.

We also did not find any correlation between the difference in perceived performance and the difference in influence across participants (Pearson correlation coefficient, between susceptible and baseline, r = .19 p = .48, between susceptible and insusceptible r = .2 p = .45).

Moreover, using a repeated measure ANOVA we found that there wasn't any difference neither in the precision of the partners (F(1.7,27) = 1.06, p = .35) nor in the precision of the participants (F(1.77,28.34) = .31, p = .7) across different conditions. Participants' precision after the second estimate didn't vary across conditions either (repeated measures ANOVA, F(1.97,31) = .39, p =.67).

Effect of confidence on influence in experiment 1:

To scrutinize the interaction between confidence and condition on influence, we used a linear mixed model with the influence of the partner on the participant as the dependent variable and the condition (baseline, susceptible or insusceptible) and participant's confidence as independent variables:

$$I = \beta_{1s} + \beta_{2s} \times Cond + \beta_{3s} \times C + \beta_{4s} \times C * Cond \quad (2)$$

I is the influence of the partner on the participant, $Cond$ is the condition, and C is participant's confidence. We categorised our confidences into three levels: low (confidence 1 and 2), intermediate (confidences 3 and 4), and high (confidences 5 and 6). Then for each category we computed the average influence for that category.

The intercept (β_{1s}) and all slopes ($\beta_{2s}, \beta_{3s}, \beta_{4s}$) were allowed to vary across participants by including random effects of the form $\beta_{ks} = \beta_{k0} + b_{ks}$ where $b_{ks} \sim N(0, \sigma^2)$. We first compared the susceptible to the baseline condition. While the effect of confidence and intercept were significant (β_{10} t(964) = 10.22 p = 0, β_{30} t(96) = -5 p = 2×10^{-6}) there was no effect of condition (β_{20} t(96) = 1.36 p = . 17) nor of the interaction between condition and confidence (β_{40} t(96) = -.59 p =.55). Since the effect of the interaction term on the influence was the weakest factor, we removed it from the model to reduce the number of model parameters and therefore increase the statistical power of the remaining model. As a consequence

we obtained a significant effect of condition for the model without interaction term (β_{20} t(97) = 2.43 p = $.01$).

We then compared the susceptible condition to the insusceptible condition. All four factors of the complete model (eq. 2) were significant (β_{10} t(96) = 9.45 p = 2×10^{-15}, β_{20} t(96) = -5.75 p = 1×10^{-7}, β_{30} t(96) = -5.78 9×10^{-8}, β_{40} t(96) = 5. 15 p = 1×10^{-6}).

Effect of confidence on influence in experiment 2A:

We applied the same procedure as above to rule out the effect of confidence on influence in experiment 2A:

$$I = \beta_{1s} + \beta_{2s} \times Cond + \beta_{3s} \times C + \beta_{4s} \times C * Cond \quad (3)$$

where I is the influence of the partner on the participant for each confidence category, $Cond$ is the condition of the experiment (susceptible, insusceptible) and C the confidence category. The intercept (β_{1s}) and all slopes ($\beta_{2s}, \beta_{3s}, \beta_{4s}$) were allowed to vary across participants by including random effects of the form $\beta_{ks} = \beta_{k0} + b_{ks}$ where $b_{ks} \sim N(0, \sigma^2)$. While the effect of intercept and confidence were significant (β_{10} t(167) = 6.75 p = 2×10^{-10}, β_{30} t(167) = -3.15 p = $.001$) there was no effect of condition (β_{20} t(167) = 1.4 p = $.16$) nor of the interaction between condition and confidence (β_{40} t(167) = -.15 p =.87). Since the effect of the interaction term on the influence was the weakest factor, we removed it from the model to reduce the number of model parameters and therefore increase the statistical power of the remaining model. As a consequence we obtained a significant effect of condition for the model without interaction term (β_{20} t(168) = 335 p = $.0009$).

Effect of perceived performance on influence in experiment 2A:

We constructed a linear mixed model

$$I = \beta_{1s} + \beta_2 \times Cond + \beta_3 \times P_{self} + \beta_4 \times P_{partner} \quad (4)$$

where *I* is the average influence of the partner on the participant in a specific condition, $Cond$ is the experimental condition (susceptible, insusceptible), P_{self} is the performance rating for self and $P_{partner}$ is the performance rating for the partner. We included random effects to allow for participant specific intercepts ($\beta_{1s} = \beta_{10} + b_{1s}$ where $b_{1s} \sim N(0, \sigma^2)$ while all other factors were fixed effects. While there was a significant effect of condition (t(56) = -4.1 p = .0001) , neither the performance rating for self (t(56) = -.04 p = .3) nor for the partner (t(56) = .16 p = .87) had a statistically significant impact on influence.

Choosing random or fixed effects:

In all the above models, we included random effects in the intercept to allow for a participant specific baseline influence. In models (2) and (4) we additionally included random effects in the other coefficients to account for repeated measurements from individual participants.

Potential confound of the gender of the experimenter in experiment 2A and 2B:

One could argue that the difference in reciprocity between playing with an alleged human or computer partner could be a confound of the gender of the experimenter. In our experiments, we randomly assigned male and female experimenter to participants. To rule out an effect of the gender of the experimenter, we compared the reciprocity between the participants which were assigned male and female observers in experiments 2A and 2B. Neither for alleged human nor for computer partners a significant effect of the experimenter's gender was found (Wilcoxon rank sum test, experiment 2A p = .62, Z = .48, experiment 2B p = .93, Z= -.07).

Effect of distance of the initial estimates on influence and reciprocity:

To answer this question, we first excluded all trials during which the participant reported a high confidence (confidence levels 5 and 6) as in these trials the estimate of the partner was always around our participants' estimates by construction (see Methods). We then divided our trials into 4 categories depending on the initial distance between the participants' and their partner's estimates. These 4 categories were defined by the four intervals [0°;45°), [45°;90°), [90°;135°) and [135°;180°) which

together spanned the whole range of possible distances. To investigate the effect of distance on influence we then conducted a mixed ANOVA with four distance categories and two conditions as within subjects factors and experiment (2A or 2B) as between subjects factor. We did not find an effect of distance ($F(2.6,151) = 1.17$ $p = .31$) nor condition (susceptible or insusceptible) ($F(1,58) = 2.76$ $p = .1$). We also did not find a significant interaction between distance and condition ($F(2.6,151) = 1.47$ $p = .22$), nor between distance, condition and experiment ($F(2.6,151) = 1$ $p = .38$). The interaction between distance and experiment was around the significance threshold ($F(2.6,151) = 2.7$ $p = .05$). And finally, in line with our previous finding, there was a significant interaction of experiment and condition ($F(1,58) = 4.36$ $p = .04$).

Next we investigated the effect of distance on reciprocity in experiment 2A. We calculated the reciprocity for each distance category and each participant and a repeated measure ANOVA revealed no effect of distance on reciprocity ($F(2.61,75) = 1.63$ $p = .16$). This analysis was only carried out for experiment 2A and not 2B as there was no reciprocity observed in experiment 2B.

Effect of partner's susceptibility on performance rating separately for experiment 2A and 2B:

We computed the difference between the participants' ratings of their own performance in the susceptible and the insusceptible condition (Figure S2). There was no difference between experiments 2A and 2B (Wilcoxon rank sum test $Z = .52$ $P = .6$). We therefore aggregated the data from both experiments and found a significant difference between the performance rating in the susceptible and insusceptible conditions (Figure 3E, Wilcoxon sign rank test, $Z = 3.04$ $p = .002$). For experiment 2A alone the difference between the conditions was just above significance threshold (Figure S2, panel A Wilcoxon sign rank test $Z=1.86$, $p=0.06$), while for experiment 2B alone the effect was significant (Figure S2, panel B, Wilcoxon sign rank test $Z = 2.46$ $P = .01$).

We repeated the above procedure for the participants' ratings of their partners' performance. There was no difference between experiments 2A and 2B (Wilcoxon rank sum test, $Z = -.47$ $P = .63$). We therefore aggregated the data from both experiments and found no significant difference between the performance rating in the susceptible and insusceptible conditions (Wilcoxon sign rank test, $Z = -.9$ $p = .36$).

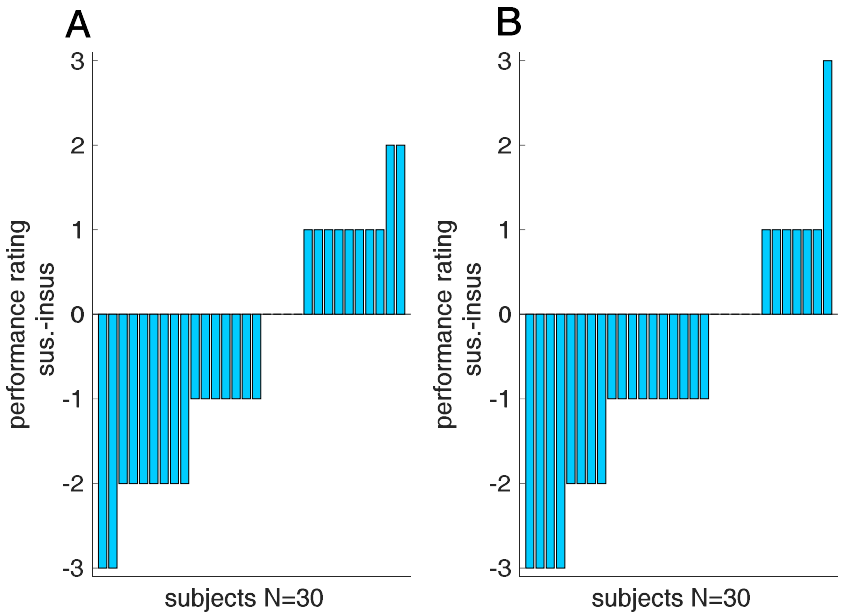

Figure S2*:* Participants' rating of their own p*erformance in* the *insusceptible minus* the *susceptible condition separately for experiment 2A* (A) *and 2B* (B).

Experiment 1 without excluding any participant:

For the results presented in the main part three participants were excluded from the analysis of experiment 1. Two of them got away from their own and their partners' initial estimates in many trials and therefore violated the task instructions. The third one did not pay attention during the experiment and during the debriefing we realised that the participant did not understand the differences between the partners. Here, we reanalysed experiment 1 using the data from all participants including these three participants that were originally excluded. Our results show that reciprocity was still significantly above zero in both cases (Figure S3; Wilcoxon sign rank test, baseline vs susceptible $z = -2.5$ $p = .01$, and insusceptible vs susceptible $z = -3.47$ $p = .0005$).

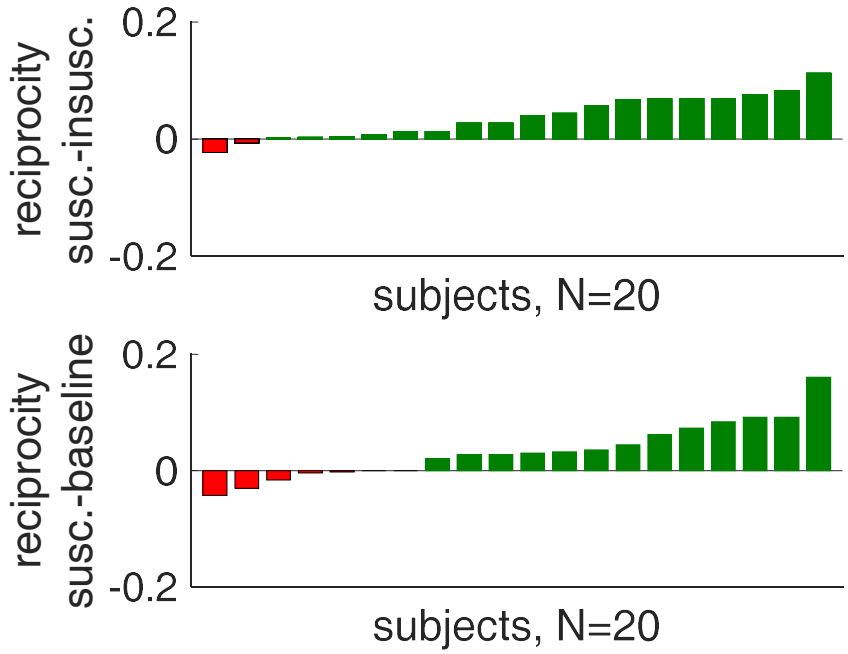

Figure S3: Results of experiment 1 including the three participants which were excluded in the analysis presented in the main text.