Overestimate yourself or underestimate others? Two sources of bias in bargaining with joint production

Quentin Cavalan, Vincent de Gardelle, Jean-Christophe Vergnaud

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Documents de Travail du
Centre d'Economie de la Sorbonne

CES Working Papers

CES
Centre d'Économie de la Sorbonne
UMR 8174

Overestimate yourself or underestimate others?
Two sources of bias in bargaining with joint production

Quentin CAVALAN, Vincent de GARDELLE
Jean-Christophe VERGNAUD

2020.03

Maison des Sciences Économiques, 106-112 boulevard de L'Hôpital, 75647 Paris Cedex 13
https://centredeconomiesorbonne.univ-paris1.fr/
ISSN : 1955-611X

Overestimate yourself or underestimate others? Two sources of bias in bargaining with joint production

Quentin Cavalan[1], Vincent de Gardelle[2], and Jean-Christophe Vergnaud[3]

[1]CES, Université Paris 1, PSE
[2]CES, CNRS, PSE
[3]CES, CNRS

Abstract

Although conflicts in bargaining have attracted a lot of attention in the literature, situations in which bargainers have to share the product of their performance have been rarely investigated theoretically and empirically. Here, by decomposing the well-known overplacement effect, we show that two types of biases can lead to conflict in these situations: players might be overconfident in their own production (overconfidence bias) and/or underestimate the production of others (other-underestimation bias). To quantify these biases, we develop a novel experimental setting using a psychophysically controlled production task within a bargaining game. In comparison to Bayesian agents, participants tend to disagree too often, partly because they exhibit both cognitive biases. We test interventions to mitigate these biases, and are able to increase settlements mainly by reducing the other-underestimation bias. Our approach illustrates how combining psychophysical methods and economic analyses could prove helpful to identify the impact of cognitive biases on individuals' behavior.

Keywords— overconfidence, bargaining, joint production, belief updating

JEL— C91, D03, D74, D81

1 Introduction

On one side, the Winklevoss brothers believed that their original ideas about an online social network were essential. On the other side, Mark Zuckerberg believed that these ideas were of little value without the technical solutions he provided. At stake, a company now estimated at several hundred billions of dollars... When the time comes to split a jointly produced wealth, such as the ownership of Facebook shares, the wealth of a household or the custody of a child, parties who were formerly working together can disagree on their respective contributions. A costly conflict can arise between the former allies, as both claim more than what the other is willing to give.

Many studies in behavioral economics and psychology have established that agents often overestimate their skills relative to others, which can lead to economic failures. Entrepreneurs over-estimate their chances of success when entering a market, leading to excess entry (Camerer and Lovallo (1999)). Individuals who believe to be better than average underweight the advice they get (Gino and Moore (2007)) and exhibit more aggressive behaviors in experimental wargames (Johnson et al. (2006)). However, it is not always clear whether such failures are due to overconfidence in one's own skills (overconfidence bias), the underestimation of others (other-underestimation bias), or both. Distinguishing between these two biases could help us understand and potentially reduce the economic inefficiencies they generate.

In this paper, we focus in particular on situations of bargaining over a joint production where the contributions of both parties are uncertain. Although they are not very well studied in the literature, these situations are in fact ubiquitous in day-to-day life. For instance, in a household, the contributions of each member is uncertain, since they include not only wages but also time, sacrificed career opportunities, participation to chores, etc. As such, in case of a divorce both members may over-estimate their own contributions to the household's wealth and/or they might fail to fully acknowledge the contributions of the other member. Similarly, in collective bargaining between labor unions and management, parties may have incompatible beliefs about their respective merits and about the distribution of wealth between employees and share-holders. Critically, in these situations, uncertainty leaves room for each party to overestimate its own contribution and underestimating the other's.

To the best of our knowledge, only one experimental study has investigated bargaining over a joint production made under uncertainty. Karagözoğlu and Riedl (2015) used an unstructured bargaining procedure, and showed that providing participants with information about their individual performances lead them to increase their entitlements, which impacted the first proposals, the bargaining duration and the agreements reached. This study however did not document precisely participants' overconfidence in their contribution or their underestimation of the other. Thus, the impact of these cognitive biases remains unknown. Our study will aim at addressing this issue.

We note that other biases have been documented as well, in studies that have focused on bargaining over given goods (rather than produced goods). For instance, individuals tend to be too optimistic about their negotiation skills (Neale and Bazerman (1985)). Bargainers have biased beliefs about the value of their outside options if the negotiation fails (Bazerman and Moore (2012), Dickinson (2009)) or about the precision of their judgment in diplomatic negotiations (Bazerman and Sondak (1988)). When they have to predict the outcome of real legal cases, participants judge the case more favorable to the plaintiff when they are assigned the role of the plaintiff, and to the defendant when assigned that role (Babcock and Loewenstein (1997)). Overall, these biases will lead to more negotiation failures as well, but they are not about how individuals estimate their contribution when uncertain.

In the present study, our hypothesis is that individuals believe they have produced more than others in

joint production tasks, because of an overconfidence bias and an other-underestimation bias, both of which should generate bargaining disagreements. If this hypothesis is correct, then treating those biases should increase bargaining efficiency by diminishing disagreements. To test this hypothesis, we design an experiment where participants produce some joint wealth and share it through a particular version of a Nash demand game where payoffs in the case of a disagreement depend on the true individual performance, which is not known with certainty by participants. By eliciting participants' subjective estimation of their performance (i.e. their confidence), both before and after they receive information about the joint production, we are able to evaluate separately how they overestimate themselves and underestimate others. We assess the effect of these two cognitive biases on participants' bargaining behavior and on outcomes of interest.

The primary outcome of the negotiation is whether an agreement was reached or not, and we define the *disagreement index* as the probability that the negotiation fails. Besides, one may also want to check how individuals' claims correspond to their actual contributions. We will thus also consider the *mismatch index* which we define as the distance between claims and contributions. Quite obviously, social planners might want to limit disagreements because they are costly: when the two parties have to resort to court, dead-weight losses are incurred by both parties and by society. However, low levels of disagreements may not be desirable if settlements do not reflect the respective contributions of individuals. Thus, the social planner might want individuals to claim what they have produced in order for them to get what they deserve in case of an agreement.

Note that studying cognitive biases on a real production task introduces a number of methodological challenges. An important contribution of this paper is to propose a way to address them. First, the literature on overconfidence shows that they heavily depend on the difficulty of the task: individuals tend to underestimate their own performance on easy tasks while overestimating it on hard tasks (this is known as the hard-easy effect). Similarly, they tend also to overestimate others' performance on hard tasks and underestimate it on easy tasks (Moore and Healy (2008)). Here, by using a perceptual task and the associated methods of psychophysics, we control the difficulty of each participant's task to obtain comparable conditions of performance for all participants, such that heterogeneity in performance is not a confounding factor when assessing cognitive biases in our setup. Second, because of the uncertainty surrounding the task, we cannot consider that individuals perfectly know what they produce, but based on Signal Detection Theory (Green and Swets (1966)) we can however model what they can infer about their own performance. By doing so, we were able to predict the distribution of confidence and strategic behavior for rational agents which can then be compared to our data. This revealed that even rational and unbiased agents (i.e. Bayesian ideal observers) making optimal claims in the bargaining game would still disagree nearly 25% of the time because of the uncertainty about their production, providing lower bounds to bargaining efficiency level.

The rest of the paper is organized as follows. Section 2 presents the design, methods and measures of the experiment. Section 3 describes our experimental results. First, we show that both an overconfidence and an other-underestimation bias are present in our data, and that they have a detrimental impact on the disagreement index and the mismatch index. Second, by setting up several interventions, we show that through one of them, we significantly increase settlements in our bargaining game and improve the match between participants' claims and their true contribution. Furthermore, we show that the intervention's impact goes through a decrease in the other-underestimation bias. Third, we suggest that these effects might depend on gender, as they appear to be more pronounced for women in our sample. Finally, in section 4 we discuss the methodological advantages of our approach as well as avenues for future research.

2 Design, predictions and measures

We conducted two experiments for the present study: a pilot experiment, and the main experiment. In the pilot experiment, 64 participants played against each other, in 4 different sessions. The data from this experiment was used to program the computer against which participants played in the main experiment. The main experiment involved 234 participants in 15 experimental sessions of 1h30 each. The number of men and women in each session was balanced. All sessions took place at the Paris Experimental Economics Laboratory (LEEP) at the University of Paris 1. Before making compensated choices, participants received a training for all the parts of the experiment. At the end of the experiment, some rounds were randomly drawn to determine participants' earnings.

In this section, we first describe the design of our experiment, before introducing our measures of over-confidence and other-underestimation bias and our predictions based on a SDT model of confidence.

2.1 The bargaining with joint production game

The core of the experiment consists of a bargaining game over a joint production game, summarized in Figure 1. In each round of the game, participants are paired with a computer mimicking the behavior of participants in the pilot study (described in section 2.2). Each round starts with the production task followed by a confidence judgment. Then, the subject learns the exact joint production and is asked to revise his confidence judgment and to make a claim. The round ends with some feedback about the outcome of the game. Participants go through 2×26 rounds of this game with an intervention in between (described in section 2.3). We detail below the steps of the game. Experimental instructions as well as screen-shots of the experiment are presented in Appendix A and B.

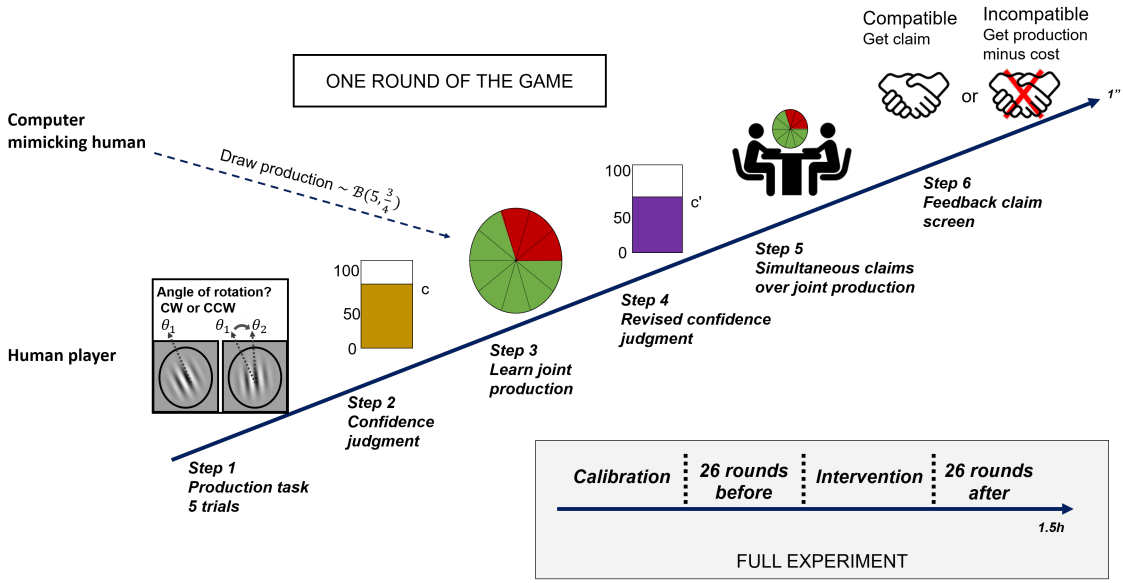

Figure 1: Schematic representation of the experimental procedure for one round of the game

The real-effort production task

The production task consists in 5 trials of a simple visual task. Two gabor patches with different orientations (θ_1 and θ_2) appear sequentially and for a brief period of time on the screen. The task is to judge whether

the second patch is oriented clockwise ($\theta_2 - \theta_1 = \Delta\theta > 0$) or counterclockwise ($\Delta\theta < 0$) with respect to the first patch. Participants respond using the computer's keyboard. After 5 such trials, the number of correct answers denoted X_1 corresponds to the participant's production in this round. Importantly, participants receive no feedback on X_1 at this stage.

The difficulty of the task is determined by $|\Delta\theta|$: the higher, the easier to determine the direction of rotation. Thus, we use $|\Delta\theta|$ to individually calibrate the task's difficulty such that participants are on average 75% correct.[1] Participants were not made aware that difficulty was calibrated.

Confidences elicitation

After the production phase, participants are asked to give their confidence on their answers (denoted c in the following). Confidence is expressed on a continuous scale from 0% to 100% where 100% means that the participant is sure to be correct and 50% means complete uncertainty (i.e. responding at chance). This confidence judgment is incentivized by applying the canonical BDM mechanism (Becker and DeGroot (1974)) to one randomly selected trial out of the 5 perceptual trials. Under expected utility, this mechanism incentivizes participants to give their average confidence across the 5 trials. Note also that this mechanism incentivizes participants to perform the task as well as possible.[2]

Then, participants learn the joint production of the dyad X which is the sum of participant's number of correct answers, X_1 and computer's number of correct answers X_2 that is: $X = X_1 + X_2$. Following this information, participants are asked again to give their confidence on their performance, now knowing X. We call this second judgment, the "revised confidence" and denote it $c|X$ in what follows. This revised confidence is incentivized in the same way as the initial confidence judgment.

The bargaining game

Finally, participants bargain with the computer over the joint production. The bargaining game is a slightly modified version of a one-shot Nash Demand Game where players get what they individually produced if claims cannot be satisfied. More precisely, the participant and the computer make simultaneous claims (Claim_1 and Claim_2) which represent the percentage share of the joint production they want for themselves.

- If claims are compatible i.e. $\text{Claim}_1 + \text{Claim}_2 \leq 100$, the participants have reached an agreement and the joint production is shared in proportion of the two claims. The participant's payoff is thus:

$$V = 100\text{X} \times \frac{\text{Claim}_1}{\text{Claim}_1 + \text{Claim}_2}$$

- Otherwise, if $\text{Claim}_1 + \text{Claim}_2 > 100$, they disagree and have to settle in court where they get their own contribution minus a cost of 50. Their payoff will be:

$$V = 100X_1 - 50$$

In all cases, once the claims are made, participants are informed about the computers' claim and whether an agreement was reached. In case of an agreement, participants also learn the number of points they get. In case of a disagreement, they are simply reminded about the rule that converts their (unknown) performance into points. Overall, participants receive the same amount of information regarding their individual

[1] See Appendix C for the details of the difficulty calibration procedure.

[2] The proof is in Appendix D

performance, irrespective of whether they agree or disagree. Then, once the bargaining game is over, the next round starts.

2.2 Computer's behavior

Participants in the experiment are not matched with each other but with a computer whose behavior mimics that of participants in the pilot experiment, where they were matched with each other. We used this specific protocol to have a controlled environment close to a natural one and still get independent observations for the analysis.

More precisely, at each new round, we draw the computer's number of correct answers X_2 from a binomial distribution $\mathscr{B}(5, 0.75)$. Thus, the computer's performance is at the same level as participants'. For the bargaining phase, the computer's claim Claim_2 is drawn from participants' claims in the pilot experiment in similar conditions, that is, from claims observed in rounds where X_2 and X had the same values as in the current round.[3]

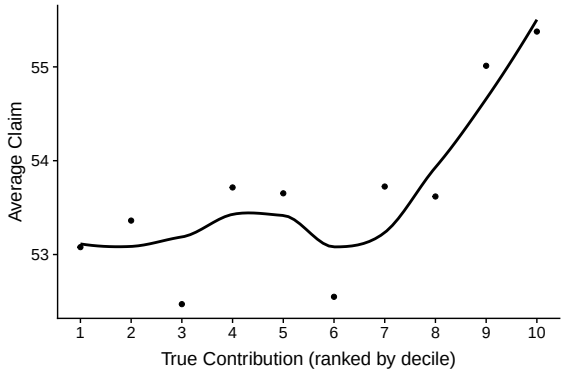

Figure 2: Relationship between participants' claims and true contributions in the pilot study, used to determine computer's behavior in the main study. The true contribution was split in deciles to improve visualization. Dots represent average data accross participants for each decile. The line represents local polynomial regression.

To get some insights about the computer's behavior, in Figure 2 we plot participants' claims and participants' true contribution, in the pilot study. Although computer's claims do not depend solely on the computer's contribution but on both X_2 and X, this representation shows that there is a positive relationship between claims and contributions past a certain level of contribution, which means that computer will tend to make higher claims when its contribution is higher.

At the beginning of the experiment, participants are told that they are playing against machines that play and perform like humans, in the sense that they have been programmed using real participants playing the same game against each other. We present this simpler story to participants to make them understand that there is an underlying real behavior which drives computer's claims without focusing on the complex algorithmic of the computer's behavior.

[3]In Appendix E, we present the mean and standard deviation of computer's claim depending on (X_1, X_2).

2.3 The intervention

In all experimental sessions, after 26 rounds, participants are asked questions and received feedback about one or several of the variables of the game. There are 5 different interventions (4 treatments and a baseline condition) which varied the nature and the number of questions asked.

- In the baseline condition, T_0, we ask Q_0: "In your opinion, what was your average claim in the previous rounds of bargaining?"
- In treatment T_1, we ask Q_1: "In your opinion, what was your average contribution to the joint production in the previous rounds of bargaining?"
- In treatment T_2, we ask Q_2: "In your opinion, how many times was your claim higher than your contribution in the previous rounds of bargaining?"
- In treatment T_3, we ask Q_3: "In your opinion, how many times was your initial confidence higher than your performance in the previous rounds of bargaining?"
- In treatment T_4, we ask Q_1, Q_2 and Q_3 altogether.

The goal of these interventions is to impact participants' beliefs in order to affect their bargaining behavior. Q_1 is targeting participants' tendency to overestimate their contribution to the joint output. It is designed to make participants aware that, on average, they may be contributing less than what they think. Q_2 also targets participants' estimation of their contribution but makes an explicit link with their claims in order to make participants realize that they might be making unreasonable claims with respect to their contribution. Finally, Q_3 is designed to assess more directly how participants' knowledge of their individual performance, independently of the other player, impact their bargaining behavior.

Finally, our baseline condition involves a question about past claims, to control for potential effects of time or increased motivation related to the intervention as we do not expect participants to make systematic mistakes on this question or to change their behavior based on this intervention.[4]

Interventions take place after 26 rounds of bargaining. The experimenter explains in detail the question(s) asked to participants and the feedback they will obtain. Participants then receive and answer the question(s) on the computer screen, and obtain two different types of feedback. First, they see the correct answer along with their actual answer. Second, they receive a visual animation showing their entire distribution of the variable at stake. For instance, in T_1, participants first see their average contribution and can compare it to their answer and then they see their entire distribution of contributions in past rounds. The duration of all interventions was the same (approximately 10 minutes).[5]

2.4 Measures

We used several measures to characterize participants' beliefs about their performance, namely the over-placement bias which is divided in two sub-components that are an overconfidence bias and an other-underestimation bias. In addition, we define two measures regarding the outcomes of the bargaining game:

[4]This can be checked on the data of the pilot experiment where similarly to this baseline condition, we ask this question at the middle of the experiment. We find no systematic mistake: participants slightly underestimate it but this is not significant ($\overline{\text{Error}_{Q_0}} = -2.18\%,\ t(32) = -1.50,\ p = 0.14$). Moreover, we do not find any significant correlation between the mistake made at this question and participant's change in claims after the intervention ($\text{Cor} = 0.09,\ t(31) = 0.48,\ p = 0.63$)

[5]We made sure that the intervention in T_4 was not significantly longer than the other interventions. To do so, less time was spent explaining the questions to participants in T_4 compared to T_0, T_1, T_2 and T_3.

a disagreement index and a mismatch index. Finally, we also define a measure of how claims are sensitive to entitlements.

Subjective entitlements

Given participants' revised confidence, we can retrieve the share of the joint production that participants think they have produced. We call this quantity subjective entitlement.[6] For a given round, the subjective entitlement of participant is defined as follows:

$$\text{Subjective entitlement} = \frac{5 \times \text{revised confidence}}{\text{joint production}}$$

Overplacement

Then, we define a measure of overplacement over a set of rounds as the average difference between subjective entitlements and true contributions. A positive value of overplacement corresponds to overestimating one's own contribution relative to the other. We then decompose this overplacement into overconfidence in one's own performance (overconfidence bias) and underestimation of the performance of the other (other-underestimation bias) as defined bellow.

Overconfidence bias

We compute the overconfidence bias as the average difference between participants' confidence and their performance, as measured in terms of the share of correct answers. A positive overconfidence bias corresponds to an overestimation of one's own performance.

$$\text{Overconfidence bias} = \overline{c - \text{performance}}$$

Other-underestimation bias

We measure the other-underestimation bias indirectly from the way participants revise their confidence after learning the joint production X.[7] In a nutshell, we use participants' revised confidence $c|X$ and compare it to the revised confidence that a Bayesian agent with the same confidence c and a belief about the performance of the other c_{other}, would have. To do so, we hypothesized that participants consider that their own and the other's number of correct answers follow the binomial distributions $X_1 \sim \mathscr{B}(5, c)$ and $X_2 \sim \mathscr{B}(5, c_{\text{other}})$ respectively.

For instance, suppose that a participant who has a confidence of $c = 75\%$, thinks that the other performs at $c_{\text{other}} = 75\%$, and learns that $X = 7$. Then, the revised confidence for this participant should optimally be:

$$c|X = \frac{1}{5}\sum_{k=1}^{5} k \times \mathbb{P}(X_1 = k | c = 75\%, c_{\text{other}} = 75\%, X = 700) = 70\%$$

Considering this Bayesian model of revised confidence, we are able to recover c_{other} from c, X and $c|X$.[8] Then, the other-underestimation bias is simply the difference between the true performance of the computer

[6] Even though participants are not directly asked to state what they think they should get at the bargaining table when they state their revised confidence, we use this term because this is a good proxy of entitlements given that in case of a disagreement the court will share the joint production according to individual productions.

[7] We did not ask directly participants their belief about the probability of success of the other as we thought asking such question round after round would not make sense for them.

[8] In Appendix F, we provide some evidence that participants' confidence revision is consistent with such model and that it outperforms alternative models of confidence revision.

and c_{other}, averaged across rounds. A participant underestimates the performance of the other when this bias is above 0.

$$\text{Other-underestimation bias} = \overline{\text{performance}_{\text{other}} - c_{\text{other}}}$$

Table 1 shows how c_{other} varies depending on c, X and $c|X$ in some specific examples. Mathematically, c_{other} increases with c (compare column 1 vs column 2) and with X (compare column 1 vs column 3) but decreases with $c|X$ (compare column 1 vs column 4).

c	75	60	75	75
X	7	7	5	7
$c\|X$	70	70	70	73
c_{other}	75	60	40	70

Table 1: Calculation of c_{other} (the belief of the performance of the other) given specific values of c, X and $c|X$

Disagreement index

The main outcome of bargaining is whether players reach an agreement or not. Here, rather than using the actual disagreement in each round, that depends on the participant's claim C_1 but also on the random draw of the computer's claim C_2, we consider the expected disagreement corresponding to participant's claim, which we evaluate over the possible draws of the computer's claim, knowing X and X_2.[9] The disagreement index of a set of rounds is the average of the expected disagreement. It lies between 0% and 100%. From a social planner perspective, the closer to 0 the better.

$$\text{Disagreement index} = \overline{\mathbb{E}_{C_2}\Big(\mathbf{1}_{\{C_1+C_2>100\}}|C_1, X_2, X\Big)}$$

Mismatch index

The goal of our interventions is to decrease the disagreement index. However, we control for the fact that we do not achieve this decrease at the expense of another feature of interest: the mismatch index which is defined as the distance between a participant's claims and actual contributions to the joint production. We measure it as the average squared difference between participant's claims and actual contributions, divided by 25 such that it lies between 0 and 100.[10] Here again, from a social planner perspective, the closer to 0, the better.

$$\text{Mismatch index} = \frac{1}{25}\overline{\Big(C_1 - \frac{X_1}{X}\Big)^2}$$

Sensitivity to entitlements

Finally, we measure how participants' beliefs relate to their claims in the bargaining game. Specifically, we define a participant's sensitivity to entitlements as the effect of subjective entitlements on claims, in a linear regression (least-square). The higher this measure, the more sensitive participants are to their entitlements when making claims in the bargaining game.

[9] We still check that our results are similar using the average number of disagreements by participant.

[10] Indeed, since participant are calibrated, their actual contribution to the joint output is equal to 50% on average. If the participant always claimed 100% (or 0%) of the joint production, the squared distance between the claim and the actual contribution would be 2500.

$$\text{Claim sensitivity to entitlements} = \frac{Cov(\text{Claims}, \text{Subjective entitlements})}{Var(\text{Subjective entitlements})}$$

2.5 Model predictions

In this section, we model the link between cognitive biases and bargaining outcomes. Our goal is to make qualitative predictions about how biases impact bargaining outcomes and to provide expected levels of disagreement and mismatch index.[11]

First, we simulate confidence data for a Bayesian observer with a given performance level and a given level of overconfidence bias using Signal Detection Theory. Then, using this confidence as a prior and the information on joint production, we derive the revised confidence of the observer under Bayesian updating and a given level of other-underestimation bias. Finally, we derive the disagreement index and the mismatch index by having these simulated players play the bargaining game against our computer, while considering three potential strategy for producing claims on the basis of revised confidence: an egalitarian strategy (E), a libertarian strategy (L) and a best response strategy (BR). E are equal to 50% of the joint production, regardless confidence and revised confidence levels. Thus this model of claim predicts no relationship between cognitive biases and bargaining outcomes. Claims under L are exactly equal to subjective entitlements: individuals claim what they believe they have produced. Finally, BR claims are those maximizing individual's expected gain, knowing the underlying claim distribution of the computer.

Panel A of Figure 3 shows that, unsurprisingly, subjective entitlements of unbiased Bayesian observers are centered on 50% of the joint production while the ones of biased Bayesian observer are higher. In panel B, we show the relationship between subjective entitlements and claims: claim sensitivity to entitlements is equal to 1 under L, to 0 under E while BR claims are in between, with claims stable at 50% when entitlements are below 52% and increasing with entitlements above this value.

Panel C gives theoretical grounds to our hypothesis that biases should increase disagreements. Indeed, under both BR and L strategies, the disagreement index increases when the Bayesian observer becomes more overconfident or when it becomes more affected by the other-underestimation bias. Notice that even if the Bayesian observer is unbiased (center of the graph), there is an inevitable amount of disagreements: 31% under BR and 59% under L. In the end, except under E, the model predicts that more biased individuals will disagree more.

On panel D, we show how biases affect the mismatch index. Even for an unbiased Bayesian observer, there is an inevitable amount of mismatch: equal to 4 for BR and L claims. The mismatch is also higher when both biases are extremely positive (upper right corner within each plot). Under L, mismatch also increases when both biases are extremely negative (lower left corner). This is less pronounced under BR as BR claims remain quite constant around 50% even when entitlements are low. In the end, except under E, the model predicts that individuals who are more biased (either positively or negatively) will make claims that are further away from their true contributions.

2.6 Summary of hypotheses

To summarize, in this paper we test the following hypotheses:

- H_0: Individuals reach higher levels of disagreement and mismatch indices than BR Bayesian observers.

[11] See Appendix G for the details of the model generating confidence and revised confidence data.

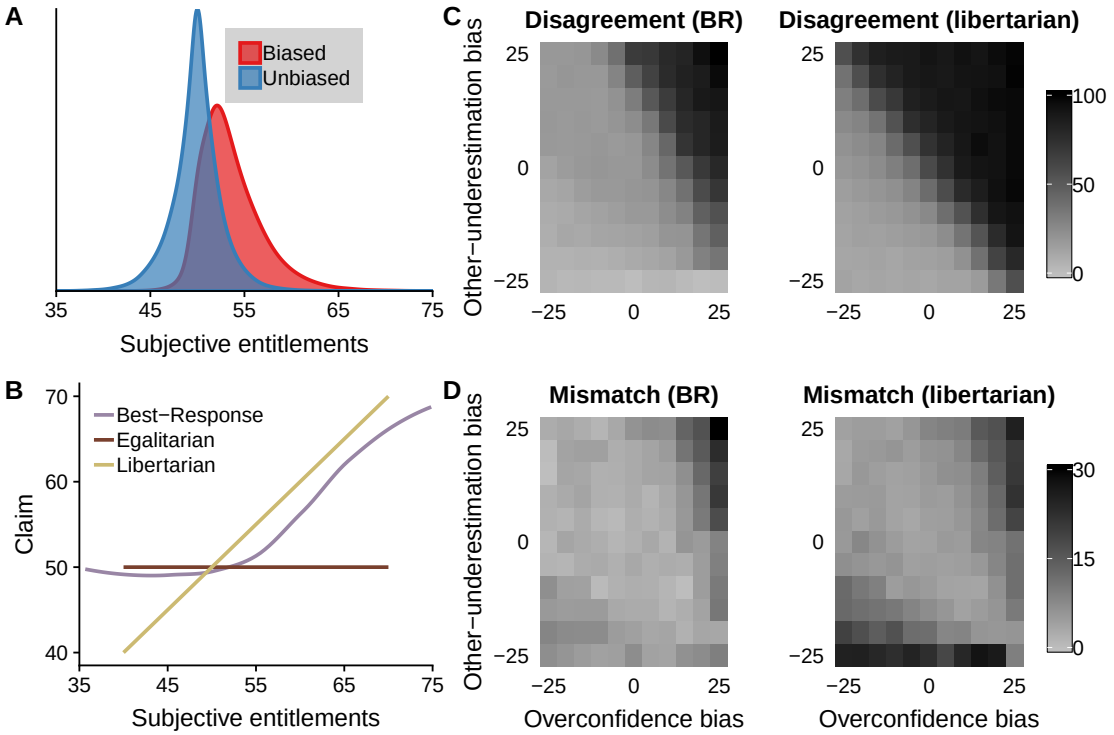

Figure 3: Theoretical impact of biases on mismatch and disagreement index depending on strategy. (A) Simulated distribution of subjective entitlements ($n = 2 \times 10^6$) of an unbiased agent and a biased agent (overconfidence bias = 5%, other-underestimation bias = 5%). (B) Claim of agents following BR, E and L as a function of their subjective entitlements. Claims for BR agents were based on simulations ($n = 2 \times 10^6$) and represented using a local polynomial regression. (C) Disagreement index for BR and E agents as a function of the overconfidence bias (x axis) and the other-underestimation bias (y axis). Each cell is an average over $n = 2000$ simulations. (D) Idem for the mismatch index.

- H_1: Individuals overestimate their contribution when they bargain over a joint production. This comes from the fact that they are overconfident in their abilities but also that they underestimate the ability of the other with whom they bargain.
- H_2: Individuals that have a higher confidence and/or other-underestimation bias will tend to disagree more often.
- H_3: Individuals that have a higher confidence and/or other-underestimation bias in absolute terms will make claims that are further away from their true contribution.

3 Results

In this section, we describe the results of the experiment. First, we test the hypotheses made in section 2.6. Since all treatments are comparable before intervention, we pool all the pre-intervention data and put aside the post-intervention data in the analysis.[12] Second, we study the effect of our interventions. Finally, we

[12]There are indeed no treatment differences before interventions in terms of biases and bargaining behavior.

provide some striking elements about gender differences in our sample.[13]

3.1 Hypothesis testing

Participants disagree more and make claims further away from their contributions compared to ideals agents

Figure 4 provides evidence that the disagreement and mismatch indices are higher for participants than BR unbiased Bayesian observers (hereafter optimal agents). In terms of disagreement, both distributions are particularly dissimilar. Participant's distribution is more spread out compared to optimal agents. Moreover, it is biased towards a high average level of disagreement: participants disagree 51% of the time while optimal agents disagree only 31% of the time. The distribution of mismatch index are closer, even though the dispersion and the average are again higher for participants: their average mismatch index is equal to 10 while the one of optimal agent is equal to 4. Together, those resusts support hypothesis H_0.

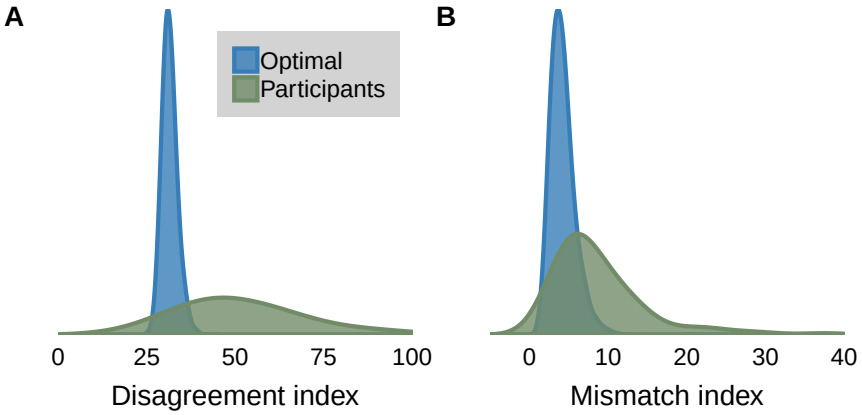

Figure 4: (A) Distribution of disagreement index for participants ($n = 218$) and for optimal agents following BR ($n = 2 \times 10^4$). (B) Idem for the mismatch index.

Without any doubt, participants do not behave optimally whether strategically or cognitively, which increases the level of disagreement and move claims further away from true contributions. Thus, there is room for interventions in order to restore the disagreement rate to a more efficient level.

Participants tend to overestimate their contribution, to be overconfident in their ability and underestimate the others' ability

Figure 5 confirms that participants are indeed subject to cognitive biases (H_1). Participants tend to exhibit an overplacement bias ($\overline{\text{OP bias}} = 4.74$, $t(207) = 12.09$, $p < 0.001$) observed for 81% of them. Indeed, whereas their true contributions are around 50% (which is expected due to difficulty calibration), participants' subjective entitlements are on average equal to 54.46. Then, we decompose this overplacement bias in an overconfidence and an other-underestimation bias (see section 2.4). First, we find that participants exhibit significant overconfidence ($\overline{\text{OC bias}} = 5.54$, $t(207) = 9.39$, $p < 0.001$) observed in 73% of participants.

[13]Out of the 234 participants, we removed 16 of them from the analysis due to difficulty calibration procedure's failure: their average performance was more than two standard deviations away from the average performance in our sample equal to 74%.

Second, we find a significant other-underestimation bias ($\overline{\text{OU bias}} = 2.85$, $t(207) = 5.36$, $p < 0.001$) observed in 63% of participants. Thus, participants tend to (wrongly) think they are better than the other, both because they think they perform particularly well and because they think the other performs particularly badly.

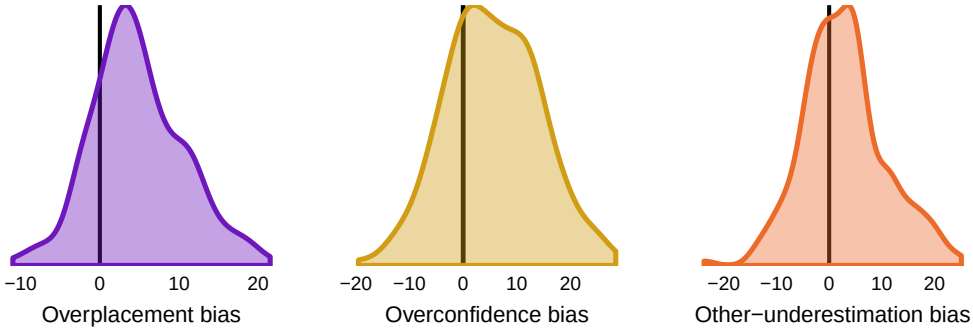

Figure 5: Distribution of overplacement, overconfidence and other-underestimation biases across participants ($n = 218$), measured in the first part of the experiment (before the intervention).

Cognitive biases prevent settlement

We then test the hypothesis that these cognitive biases lead to a higher disagreement index (H_2). This is indeed the case: participants who exhibit a higher overplacement bias also tend to disagree more (Figure 6). Moreover, we find that the two sub-components of overplacement contribute to increase the disagreement index (Table 2). Specifically, an increase in overconfidence by 1 unit leads to an increase in the probability to disagree by 1%, and an increase of other-underestimation by 1 unit increases this probability by 0.6%. Thus, if we debias participants with our interventions, the disagreement index should improve.

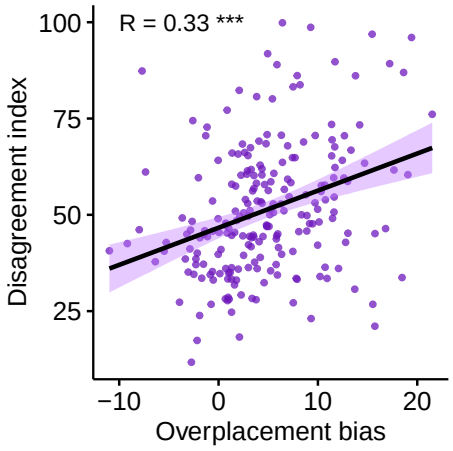

Figure 6: Correlation between overplacement bias and disagreement index. Each dot represents an individual participant. The line represents the least square regression with confidence interval.

	log(Disagreement index)	
	Estimate	Std
Overconfidence bias	0.0099***	0.003
Other-underestimation bias	0.0063*	0.003
Constant	3.807***	0.026
Observations	218	
F Statistic	13.73***	
Adj. R^2	0.11	

*p<0.05; **p<0.01; ***p<0.001

Table 2: OLS regression across participants of disagreement index against overconfidence and other-underestimation biases

Cognitive biases in absolute terms deteriorate the mismatch index

Finally, we test the hypothesis that cognitive biases in absolute value deteriorate the mismatch index (H_3). We indeed find that participants who exhibit a higher overplacement bias also tend to make claims that are further away from their true contributions (Figure 7). Then, we find that the overconfidence and the other-underestimation bias in absolute terms increase the mismatch index (Table 3). More specifically, an increase of overconfidence bias in absolute terms by 1 unit leads to an increase in the mismatch index by 5.3% while an increase of other-underestimation in absolute terms by 1 unit increases the index by 2.4%. Thus, if we want our interventions to not deteriorate the mismatch index, the biases must get closer to zero.

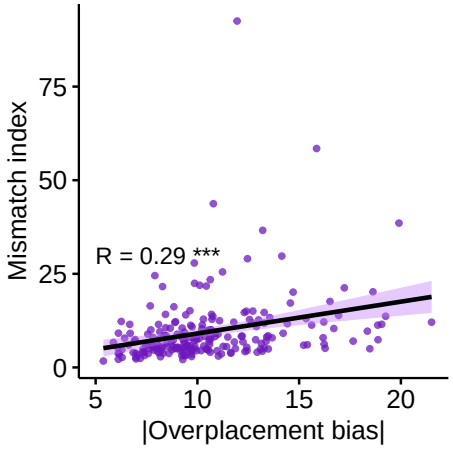

Figure 7: Correlation between overplacement bias and mismatch index. Each dot represents an individual participant. The line represents the least square regression with confidence interval.

	log(Mismatch index)	
	Estimate	Std
|Overconfidence bias|	0.053***	0.013
|Other-underestimation bias|	0.024**	0.009
Constant	0.677**	0.217
Observations	218	
F Statistic	20.73***	
Adj. R^2	0.15	

*p<0.05; **p<0.01; ***p<0.001

Table 3: OLS regression across participants of mismatch index against overconfidence and other-underestimation biases

3.2 The interventions

We now turn to the analysis of the interventions conducted at the middle of the experiment. Recall that there are 5 experimental conditions: a baseline in which we ask participants about their average claim (T_0) and 4 treatments in which we focus on their average contributions (T_1), on how their claims may have exceeded their contributions (T_2), on how their confidence may have exceeded performance (T_3), or on these last 3 variables at the same time (T_4).

Mistakes in the question asked during intervention

First, we present descriptive statistics about participants' answers during the interventions. In the baseline condition, we were not expecting any systematic error but participants marginally overestimated their claims as they reported asking for 54.75% of the joint production while they were only asking for 53.03% on average ($\overline{Error}_{Q_0} = 1.71,\ t(47) = 1.84,\ p = 0.07$). This slight misconception is not a worry though, as it should not impact behavior. By contrast, when asked about their average contribution to the joint production, participants strongly overestimated declaring that they had contributed 60.68% of the output while their true contribution was 49.68% on average. ($\overline{Error}_{Q_1} = 11.06,\ t(82) = 9.80, p < 0.001$). Confirming the robustness of participants' answers, we find that the mistake made on this question correlates across individuals with our measure of overplacement bias ($Cor = 0.51, p < 0.001$). When asked to estimate the number of times their claim was above their contribution, participants answered that it happened 7.84 times out of 26 largely underestimating the true value which was 13.55 times out of 26 ($\overline{Error}_{Q_2} = 5.71,\ t(74) = 8.80, p < 0.001$).

When asked the number of times their confidence was above their individual performance individuals also consistently underestimated this number with an estimated 10.08 times out of 26 instead of 13.18 times out of 26 ($\overline{Error}_{Q_3} = 3.10,\ t(74) = 4.31,\ p < 0.001$). This last observation is consistent with the fact that participants are mainly overconfident and naive with respect to this bias. This is confirmed by the correlation between the mistakes made on this question and participant's overconfidence measured before the intervention ($Cor = -0.50, p < 0.001$). Finally, examining answers to all questions for participants in T_4, we find that all estimation errors significantly correlate in the expected direction.

Intervention effect on disagreement index

We now show how the 4 interventions affected the disagreement index, by regressing the change in the disagreement index (before vs. after the intervention) on an intercept and on treatment effects. We run a similar model for the mismatch index. The results are presented in Figure 8 and in Table 6 and 7, Appendix H.

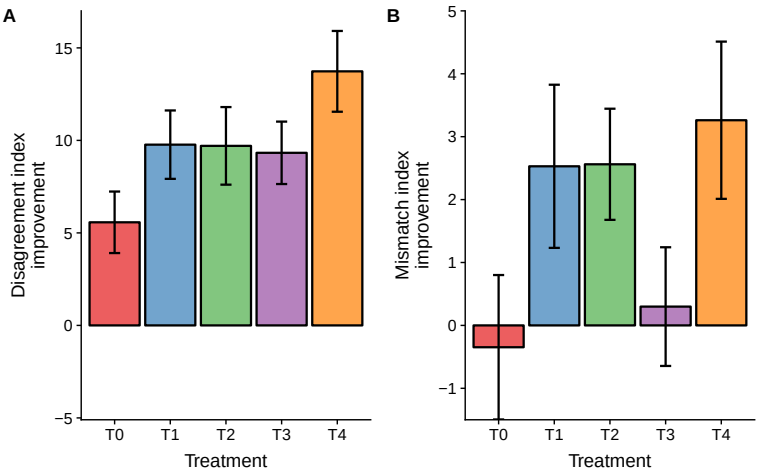

Figure 8: Improvement (i.e. reduction) of disagreement index (panel A) and mismatch index (panel B), after the intervention, for each treatment. Error bars represent mean and s.e.m. across participants.

Regarding the disagreement index, we notice that it significantly falls by 5.57 points after the intervention even in the baseline and that it falls even more in the four treatments. However, the differential effect of T_1, T_2, T_3 are not significant and only T_4 is significantly different from baseline ($p < 0.01$) with a supplementary decrease of 8.16 points. Thus, the only intervention that significantly decreases disagreement is the one where participants have to answer and receive feedback about Q_1, Q_2 and Q_3.

Regarding the mismatch index, the results are somewhat similar. The correspondence between claims and contribution improves after the intervention in all treatments compared to baseline but this is significant only for T_4 where the intervention reduced the distance between claims and actual contributions by 3.61 more points compared to baseline ($p < 0.05$). Thus, with intervention T_4 we achieve our goal of improving the disagreement index while improving the mismatch index.

For completeness, we evaluated pairwise comparisons between all pairs of treatment, and found only one marginally significant difference between T_3 and T_4 in terms of mismatch index ($p < 0.1$). We further check that our results are not driven by pre-intervention differences between treatments, and find no pre-intervention differences between baseline and T_4 in beliefs (overplacement, overconfidence and other-underestimation bias) and in bargaining behavior (claims, sensitivity to entitlements, disagreement and mismatch index). We also

check that this effect is not a measurement artifact and that the intervention led to a decrease in the raw number of disagreements ($p < 0.05$) as well as in claims ($p < 0.01$).

Intervention effect on cognitive biases

To get some insights about the channel through which the intervention impacted the disagreement index, we test whether it impacted participants' overplacement bias and, more specifically, their overconfidence and other-underestimation biases. The results are presented in Figure 9 and Table 8 and 9, Appendix H.

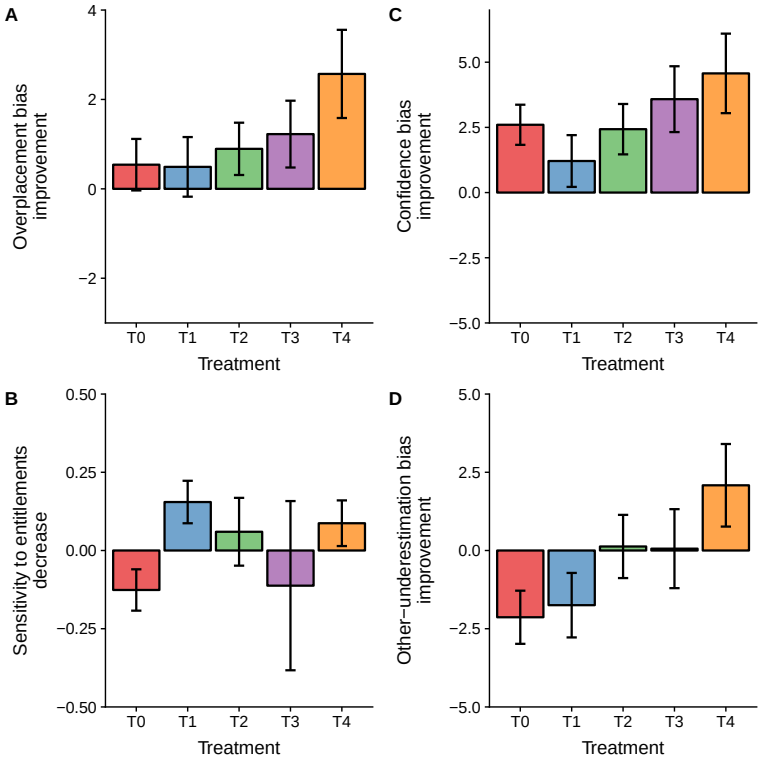

Figure 9: Improvement (i.e. reduction) in overplacement bias (panel A), overconfidence bias (panel C) and other-underestimation bias (panel D), after the intervention, for each treatment. In panel B, similarly, we show the decrease in sensitivity to entitlements (see section 2.4 for details of this measure). Error bars represent mean and s.e.m. across participants.

We find that overplacement decreases significantly more in T_4 than in baseline, by 2.03 points ($p < 0.05$).

When we decompose this effect, we do not find any intervention effect in T_4 on overconfidence bias compared to baseline: the overconfidence bias decreases in all experimental conditions. We find however, an effect on other-underestimation bias. After the intervention in baseline, the other-underestimation bias significantly increases by 2.09 points while it decreases by 2.13 points in T_4. Compared to baseline, we estimate a significant 4.22 supplementary decrease in other-underestimation bias after the intervention in T_4 ($p < 0.01$). Finally, we estimate an 0.21 supplementary decrease in claim sensitivity to entitlements in T_4 compared to baseline although it is not significant. While this decrease in claim sensitivity might have contributed to the decrease in disagreement, this null result suggests that our successful intervention did not work solely because individuals had changed their claim strategy with respect to their entitlements.

Is the effect of T_4 mediated by other-underestimation bias?

Has our debiasing intervention in T_4 induced participants into settling more often? To answer this question, we run two mediation analyses in which we mediate the intervention effect of T_4 (against baseline) on the disagreement index by the decrease in overplacement bias and other-underestimation bias. The results are presented in Figure 10.

We find that the overplacement bias can account for 24.51% of the treatment effect, a mediation that is only marginally significant ($p < 0.1$). When we focus only on the other-underestimation bias, we find a larger and more significant mediation. We find that 30.02% of the direct effect of the intervention on the disagreement index is actually mediated through other-underestimation bias ($p < 0.01$).

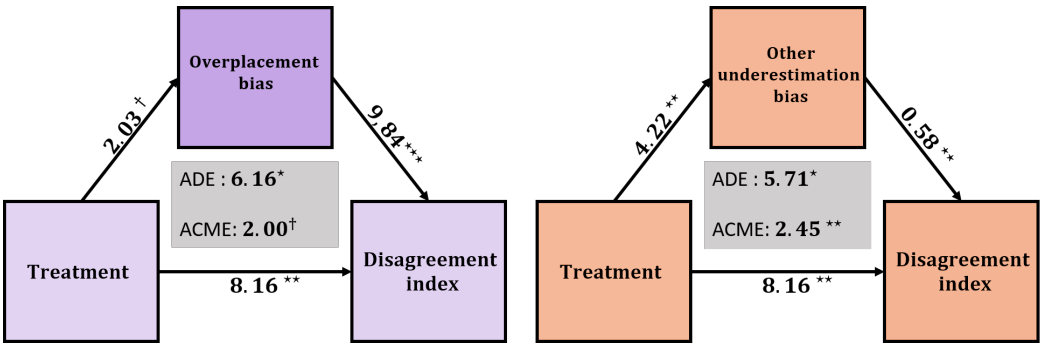

Figure 10: Representation of the mediation analysis, where the disagreement index is affected by treatment (T_4 vs T_0) via overplacement bias (left panel) or via other-underestimation bias (right panel). †p<0.1; *p<0.05; **p<0.01;***p<0.001

Overall, this analysis suggests that our intervention is able to improve the disagreement index, not because we made participants realize that they are not so good, but because we made them realize that the other player may be better than they think.

3.3 Gender differences

We conclude our results by presenting some elements about gender differences.

Gender differences in behavior and beliefs before interventions

Before intervention, we find no difference between men and women regarding the outcomes of the bargaining game (in terms of disagreement index ($F(216) = 0.97, p = 0.33$) or mismatch index ($F(216) = 0.36, p = 0.55$)) or the cognitive biases under study (in terms of overplacement bias ($F(216) = 0.20, p = 0.66$), overconfidence bias ($F(216) = 0.11, p = 0.74$) or other-underestimation bias ($F(216) = 0.52, p = 0.47$)). This similarity in cognitive evaluations is consistent with a recent study (Bordalo et al. (2019)) showing that on a quiz where men and women are expected to perform similarly, they were also similar in how they evaluated themselves or others. This also means that men and women react similarly to the information about joint performance, which contrasts with previous findings that women tend to revise more pessimistically than men when receiving feedback about relative performance (Berlin and Dargnies (2016)). We find one difference between men and women regarding the sensitivity to entitlements: men are more sensitive to their entitlements than women (($F(216) = 6.35, p = 0.01$)). This result is somewhat consistent with D'Exelle et al. (2017) who show that in a Nash demand game the influence of beliefs about what the other will claim is stronger for men than women.

Gender differences in intervention effect

We also consider gender differences with respect to interventions' effect. To do so, we run the same regression than before while introducing a main effect of gender and interaction terms between gender and treatment effects (the regression tables are in Table 10, Appendix H). Interestingly, it appears that the effect goes through women, as we do not find any treatment effect on men on disagreement or mismatch indices. This pattern suggests that women reacted to the interventions while men did not. When we focus on the comparison between baseline and T_4, we find that in T_4 the disagreement index of women improved significantly more than the one of men ($p < 0.01$), with a large difference of 16.38 points between genders. The mismatch index also improved more in women than in men in T_4 compared to baseline but the difference between genders is not significant in this case. The difference between men and women is also present in cognitive biases: when comparing T_4 to baseline, overplacement decreased after the intervention but this decrease was significantly more prononced in women than in men (a difference of 3.96 points, $p < 0.05$). When separating overplacement into its sub-components, we find no significant difference in the decrease of the overconfidence bias between men and women, but a marginal difference in the decrease of the other-underestimation bias, which was more pronounced for women than for men (a difference of 5.52 points, $p < 0.1$).

4 Discussion

4.1 Summary of the results

In this paper, we study how individuals bargain over a joint production. We disentangle two different biases from individuals' confidence: how they overestimate their own performance (i.e. overconfidence) and how they underestimate the performance of others. We show that individuals exhibit both biases and that both biases prevent settlement. However, we find that disagreements are decreased significantly after an intervention in which participants receive information about their actual contribution to the joint production, about whether they claim more than they contributed and about their overconfidence. Furthermore, we disentangle the channel by which this intervention operates, and show that the decrease in disagreements is mediated by the decrease in the other-underestimation bias. Finally, we find evidence that this effect of our intervention is mostly due to women, in that women's disagreement index and cognitive biases decreased following the intervention while men's did not.

4.2 Overconfidence and other-underestimation biases

When they evaluate their contribution to a joint production, individuals should take into account their beliefs about themselves and about others. One methodological advance of the present study is to offer a method to measure these beliefs separately. This contrasts with past studies which have focused on overplacement, that is the product of these beliefs. Before discussing these two biases in more details, we note that their measures are reliable. Indeed, in the baseline condition, where we expect them to be stable over time, we find a strong correlation across participants between the first and second half of the experiment. This is true both for the overconfidence bias (Cor $= 0.79,\ p < 0.001$) and the other-underestimation bias (Cor $= 0.71,\ p < 0.001$). These correlations confirm that both measures are reliable in time, and that our methodological approach is relevant.

Overconfidence has been an important topic of research in both economics and psychology, and numerous studies have shown that it can dramatically affect interactions between agents (Bang et al. (2017), Camerer and Lovallo (1999)). Our measure of overconfidence is relatively straightforward, in that we compare individuals' beliefs about their own performance with their actual performance, as has been done in many previous studies (Massoni et al. (2014), Grossman and Owens (2012), Clark and Friesen (2009)). However,

past studies of bargaining have used a different definition of overconfidence, focusing on agents' overoptimistic beliefs about their success in their interaction with others (Babcock and Loewenstein (1997), Neale and Bazerman (1985))). This contrasts with our definition, which is based on biased beliefs about performance in the production phase, independently of any interaction. Our study therefore addresses a gap in the literature, and shows both theoretically and empirically that this cognitive bias impacts claims and settlements in a bargaining game.

We also measure how an agent may underestimate the contribution of the other player, which also affect individuals' claims and outcomes of the game. To the best of our knowledge, our proposed measure has never been used in the past, so one could argue it deserves further scrutiny. Recall that this measure is based on the assumption that when individuals receive information about the joint production, they revise their estimation of their own performance in a Bayesian manner, using their previous estimate, the joint production, and the estimated performance of the other. Thus, one may question the relevance of this Bayesian model. We address this issue in Appendix F. First, we show that individuals' revised confidence does take into account the joint production and the initial confidence. We also show that this revision is not only driven by extreme cases in which an individual learns that the joint production is incompatible with the initial confidence estimate (e.g. by learning that the total production is 3 out of 10 after an initial estimate of 4 out of 5). Second, we test two alternative models for how individuals may revise their confidence, and we show that our Bayesian model outperforms these models. These elements provide support to the present approach, but we acknowledge that other alternatives may be considered in future investigations.

One other potential concern regarding our measures is that one could argue that participants' estimate of the other's performance could be not static as assumed here, but dynamically adjusted throughout the experiment. Indeed, within each round participants may revise their estimated performance of the other given the information they receive about the joint production, and they might use this revised estimate as a prior for the next round. If so, an overconfident participant may have a correct prior about the other, but will end up with an incorrect (underestimated) posterior about the other's performance. In other words, overconfidence could actually induce other-underestimation. This raises two potential issues. First, we might have overstated participants' level of other-underestimation bias. Second and more importantly, it might be the case that the effect of our intervention on other-underestimation bias actually comes from a small initial decrease in overconfidence bias which in turns reverberates on the other-underestimation bias through this learning process. In Appendix I, we estimate a learning model that controls for this process and show that our results still hold. Indeed, we find that even when controlling for this learning process, participants exhibit a significant other-underestimation bias, and that this bias is significantly reduced with our intervention in T_4.

4.3 The effect of a cognitive intervention on bargaining

The other main aspect of our study is the test of interventions aimed at decreasing cognitive biases in participants' evaluations and improving the outcomes of the bargaining game. Compared to baseline, all interventions decrease overconfidence and other-underestimation biases, but this effect is strongest in our most complete intervention, in which participants receive information about their true contribution as well as information about their overconfidence. Note that overconfidence decreases also in the baseline condition, which might be due to the accumulated information about disagreements throughout the experiment, or to the partial feedback that participants receive through the information about joint production. Indeed, past studies have found that getting feedback on performance reduces overconfidence (Lichtenstein and Fischhoff (1980), Arkes et al. (1987)). The decrease in cognitive biases coincides with a decrease in disagreement rates, so our interventions seem to produce the benefits that were expected. In addition, these benefits are not obtained at the cost of a greater mismatch between claims and contributions. Our work thus points at

a possible way to improve bargaining between individuals from a social planner perspective. Importantly, it remains to be seen whether such interventions could be incorporated to real-life situations and whether they could produce real societal benefits. We are of course aware that there are gaps between the present experimental procedure and what can be done outside the laboratory. For instance, our participants receive feedback about their true production, which is not always feasible in real bargaining situations. A more realistic intervention could be to provide to the two parties information about their confidence biases in tasks for which their true performance is measurable. Assuming that overconfidence is domain-general (West and Stanovich (1997), Ais et al. (2015), Kelemen et al. (2000)), debiasing them for such tasks may correct their bias in the bargaining situation.

Our analyses also pinpoint the mechanism by which our intervention produces its effect. Specifically, we find that 30% of the decrease in disagreement rates is mediated by the change in the other-underestimation bias. Interestingly, we note that our interventions do not seem to affect participants' sensitivity to their entitlements. In other words, individuals' beliefs are affected, but the way individuals use these beliefs to set their claims remains the same. This confirms that participants change their claims primarily because of debiasing at the cognitive level of evaluating the contribution (as opposed to the strategic level). Such results could help refine interventions in future studies.

One important but unexpected aspect of our results is the difference between women and men in our study. Specifically, we find that the intervention had a stronger impact on women than men. Although this effect is clear in our sample, we are aware that it would need to be confirmed in future studies with a larger sample size, as in the present study each treatment only involves 20 men and 20 women. Nonetheless, the fact that men and women may react differently to debiasing procedures would imply that gender-specific procedures have to be implemented in order to be successful. We can only speculate about the possible reasons underlying this gender effect. One possibility is that the competitive aspect of the game played a role, as suggested in other studies (Niederle and Vesterlund (2007)). It is also possible that women reacted more to the intervention because they cared more about the social benefits of reducing conflicts. Indeed, it has been argued that men and women are socialized differently in that boys are taught not to care too much about other people, while girls are encouraged to do so (Gilligan and Snider (2018)). Understanding this gender difference, and designing interventions that are effective for both men and women, constitute an interesting challenge for future research.

4.4 Concluding remarks: uncertainty in economic interactions

When engaged in a production task, economic agents do not always know perfectly what they produce. In other words, when making decisions, participants might be uncertain about the optimal action (Enke and Graeber (2019)). Karagözoğlu and Riedl (2015) have recently shown that when uncertainty about production is reduced, this increases bargaining duration and conflicts. Specifically, in their study, informing individuals about who was the best and who was the worst performer in the production task leads to claims that deviated from the equal-split and that were less compatible. Thus, paradoxically, providing more information to participants created more conflits in this study. One interesting avenue for further research would be to understand the mechanism underlying this effect. For instance, after learning that they are the best, individuals' estimation of their own performance (i.e. their confidence) should increase, but the precision of their estimated performance should also increase, and these two effects remain to be disentangled.

In our study, uncertainty comes from the limited ability of participants to estimate their own performance in a perceptual decision task. We believe that the present study illustrates how psychophysics can be fruitfully used in situations where uncertainty about production is paramount. Indeed, by controlling the difficulty of

the task, we ensure that our confidence measures are not confounded by the hard-easy effect (Moore and Healy (2008)). Moreover, by modeling the link between performance and confidence using Signal Detection Theory (Green and Swets (1966)), we can provide benchmarks for agents' beliefs and behavior. In particular we find empirically that our participants disagree 51% of the time before the interventions, while our theoretical analysis shows that optimal agents would disagree 31% of the time. When trying to reduce this gap, we obtained a 8% point benefit relative to baseline in our most effective intervention. Thus, in perspective, one could argue that we have already solved 40% of the problem. Fortunately, there is still room for more work on this issue.

A Experimental Instructions

You will participate in an experiment lasting about 2 hours. During the experiment, you will earn points that will be converted into euros. You will receive between 0 and 48 euros depending on the way you play. Once the experiment is over, go to the experimenter's office to receive your earnings.

The experiment operates in rounds: it includes a total of 52 rounds. In order to be easily identified, each round is distinguished from the others by a coloured frame. At the beginning of each round you will be randomly matched with a computer that plays "like a human" in the sense that its behavior replicates that of a player who has performed the same experience as you did in a previous session. In each round, you will have several actions to perform:

- Perform 5 trials of a perceptual task.
- Give your confidence in your answers.
- Give your revised confidence in your answers knowing the total number of correct answers of the dyad.
- Negotiate with the computer to allocate the points you earn thanks to the total number of correct answers.

Before you actually play for money, you will do a number of training sessions in which you can familiarize yourself with the experiment.

Perceptual task

Description of the stimulus

Inside the circle located in the center of the screen two stimuli will be displayed successively and for a brief moment. The purpose of the exercise is to give the direction of rotation of the stimuli: if it is clockwise, from left to right, press right using the arrows on the keyboard, if the direction is counter-clockwise, from right to left, press left.

In each round, you will perform 5 trials of this perceptual task. You will not be informed if your answers are right or wrong.

Giving your confidence

After 5 decisions on visual stimuli, the next step is to give your overall confidence in these decisions (how good or bad do you think your answers are?). You will report this confidence on a scale from 0% ("I'm sure all my answers are wrong") to 100% ("I'm sure all my answers are right") with 50% being equivalent to "I answered at random". Note that indicating a confidence level below 50% means that you think your answers are more likely to be incorrect than correct.

Incentive mechanism

The number of points you earn on each set of answers is determined by the drawing of one of the answers in that set. In order to encourage you to reveal your true level of confidence, you are given the opportunity to exchange this answer for a lottery that is necessarily more likely to be winning than your answer. In concrete terms, after having filled in your confidence, the computer randomly draws one of the 5 answers in the set and then draws a number between 0 and 100, which corresponds to the lottery in question.

- If the lottery is less good than your overall confidence then you are rewarded according to your answer: you win 200 points for a correct answer and 0 otherwise.
- If the lottery is better than your overall confidence then you are rewarded according to this lottery: you win 200 points if the lottery is winning and 0 otherwise.

Let's take a few examples to better understand:

- If you are sure you have always been wrong, then you indicate 0%. Thus, whatever the trial that is drawn, the lottery selected by the computer will be better than your confidence, it will determine your reward and you will have at least 1 chance to win
- If you are sure you always have the right answer, then you indicate 100%. Thus, whatever the trial drawn, the lottery selected is less good than your confidence and you avoid lotteries that could make you lose: your earnings will only depend on your answer.
- If you are sure at 60% of your answers, indicate 60% then whatever the random trial drawn, you will only exchange your answer for lotteries better than 60% chance of winning.

Thus, the more confident you are, the more you should indicate a high level of confidence since you will then have less chance of getting a lottery with low chances of winning. On the other hand, the less sure you are of yourself, the more you have an interest in indicating a low confidence: you will then have a better chance of exchanging your answer for a better lottery. Finally, with this mechanism, you maximize your points if you report the true probability that the answer drawn at random in the set is correct.

Giving your revised confidence

Once you have given your confidence, your number of correct answers and the number of correct answers of the computer with which you are paired are summed and revealed to you. Each set contains 5 stimuli, so the total number of correct answers in the pair is between 0 and 10. This number is indicated by vertical bars on the left and right of the screen, each green rectangle corresponding to a correct answer and each red rectangle to a wrong answer. Please note that only the total number of correct answers is revealed to you: we do not give you any information on the number of individual correct answers and therefore on the contribution of each member of the dyad to the total.

You are then asked to give your revised confidence knowing this information. Knowing the number of correct total answers can lead you to change your overall confidence in your answers. If, for example, your overall confidence was 90% but you learn that all the answers in the pair are correct (10/10), it is natural to change your confidence to 100%. If you do not wish to revise your confidence, simply indicate the same confidence as before. Note that as with the first confidence, the lottery exchange incentive mechanism presented above is applied to the revised confidence to determine an additional number of confidence points for each round.

Negociating

The total number of correct answers also determines a number of points Points_{total} that you and the computer you were paired with will then have to share.

$$\text{Points}_{total} = 100 \times \text{Total correct answers}$$

So, if 5 answers out of 10 are correct, the dyad has 500 points to share. To share this total, you are asked to make a claim. To make this claim in practice, use the mouse to move the vertical cursor and click to validate the proposed sharing: in blue, on the left, you see what you want to keep for yourself and in purple,

on the right, what you want to leave to the computer with which you are paired.

Claims are made simultaneously: neither you nor the computer is aware of the other's claim before making its own. Once the two claims are made, they are compared and declared either compatible or incompatible.

- Compatible if the sum of your claim Claim_{P_1} and the computer's claim Claim_{P_2} is smaller or equal than the number of points to share : $\text{Claim}_{P_1} + \text{Claim}_{P_2} \leq \text{Points}_{total}$. The points are then distributed according to the claims and what nobody claimed ($\text{Points}_{remain} = \text{Points}_{totaux}\text{Claim}_{P_1} - \text{Claim}_{P_2}$) is distributed in proportion to the claims you both made. More precisely, the number of points earned in this case is:
$$\text{Claim}_{P_1} + \frac{\text{Claim}_{P_1}}{\text{Claim}_{P_1} + \text{Claim}_{P_2}} \times \text{Points}_{remain}$$
For example, if out of 800 points, I claimed 400 and the other player claimed 200, 200 points were not claimed by anyone. I get my claim back as well as $\frac{400}{400+200} = 75\%$ of the 200 unclaimed points meaning $400 + 150 = 550$ points and the computer gets 250 points.
- Incompatible if the sum of your claim and the computer's claim is bigger than the number of points to share: $\text{Claim}_{P_1} + \text{Claim}_{P_2} > \text{Points}_{total}$. In this case, the sharing is decided by a court. The court knows the number of correct answers made by each player and distributes the points accordingly (each player receives 100 points for each of the correct answers assigned to him) but imposes a fixed cost of 50 points on each player as a court fee.

Let's take a few examples to understand this. Let's imagine a situation where you and the computer made 6 right answers: so you have 600 points to share. Of these 6 correct answers, let's assume that 4 comes from you and 2 from the computer (this information is not directly revealed to you).

- Situation 1: I propose to keep 300 for myself and the computer proposes to keep 150 for himself. The claims are therefore compatible because $300 + 150 = 450 < 600$. I get back my claim, 300, as well as $\frac{2}{3}$ of the 150 points that no one has claimed for a total of 400 points and the computer gets 200 points. Thus, when players do not go to court, the actual contributions of players do not have any impact on the number of points they earn individually, only the claims matter.
- Situation 2 : Each player wishes to keep 400 for himself. The claims are therefore incompatible because $400 + 400 = 800 > 600$. The court knows the individual contributions (4 right answers come from you and 2 from the computer). It therefore allocates to you 400 of the 600 total points and 200 to the computer. But the court is expensive and you each lose 50 as court' fee: so you win $400 - 50 = 350$ points and the computer wins $200 - 50 = 150$ points. Thus, when players go to court, their actual contributions have an impact on their earnings.

Your earnings in euros

At each round, you therefore win a number of points on confidence, revised confidence and negotiation. Your earnings in euros in this experiment are then determined by a draw of two rounds (out of the 52 that are made) and their associated points. Then, these points are converted into euros at the rate of 100 points = 2 euros by rounding up. Thus, for example, if out of the two series drawn you have gained $300 + 530 = 830$ negotiation points, 100 confidence points and 200 revised confidence points, your earnings are $17 + 2 + 4 = 23$ euros.

B Screenshots of the experiment

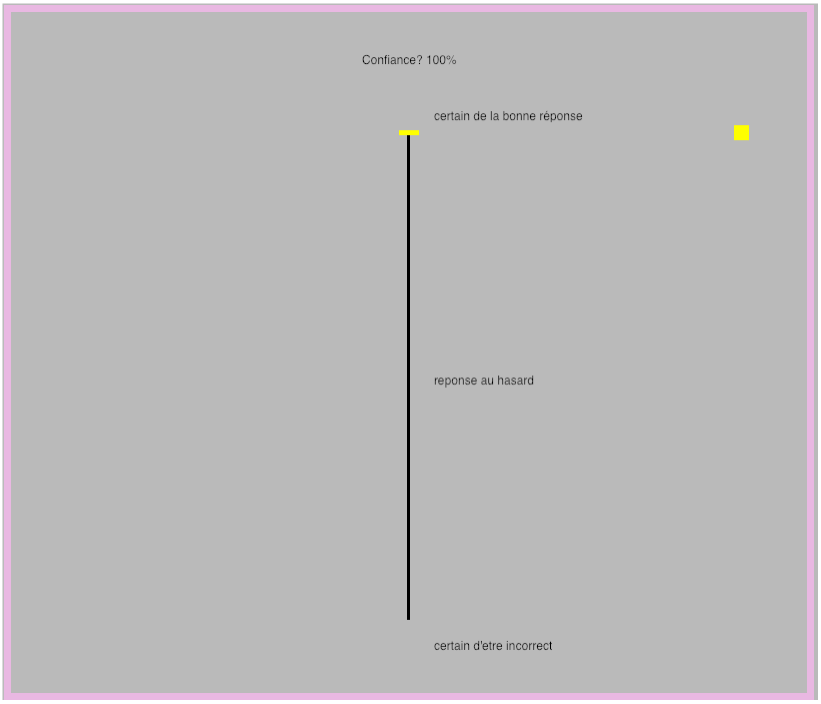

Figure 11: Confidence screen

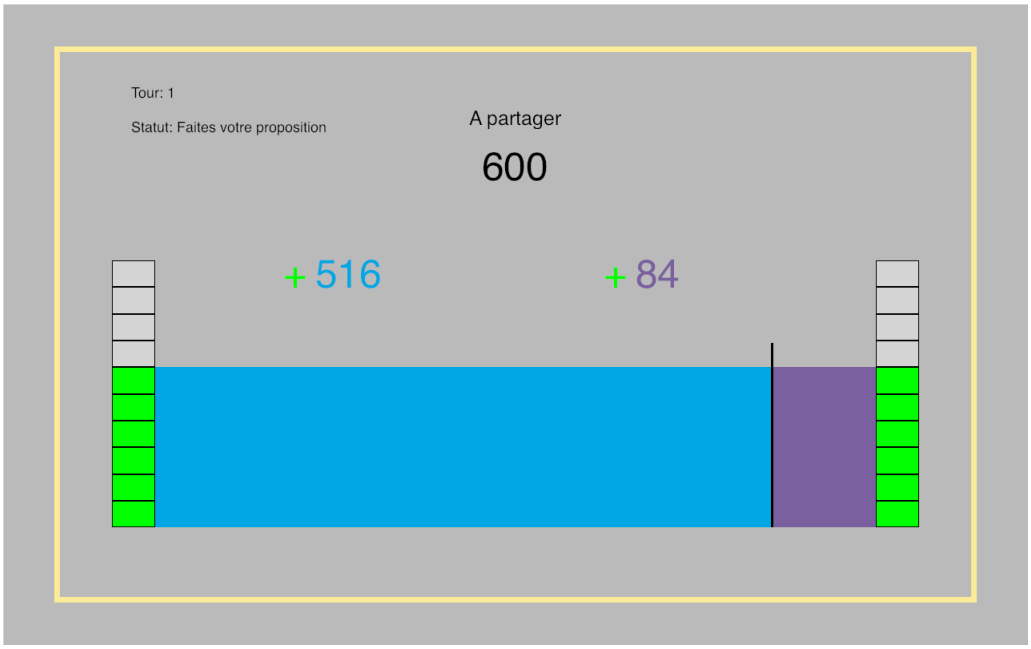

Figure 12: Claim screen

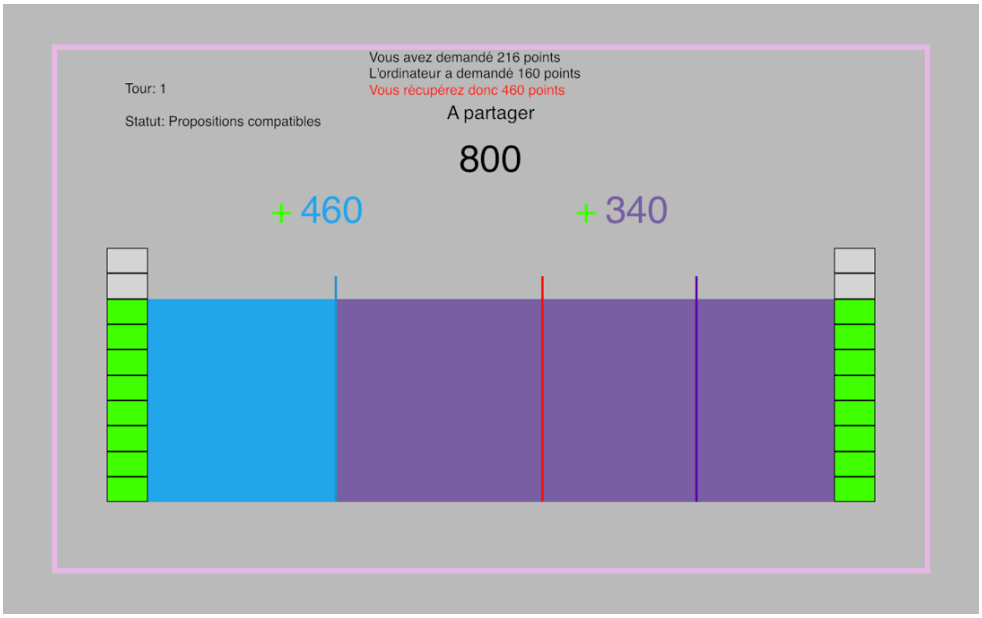

Figure 13: Negotiation success screen

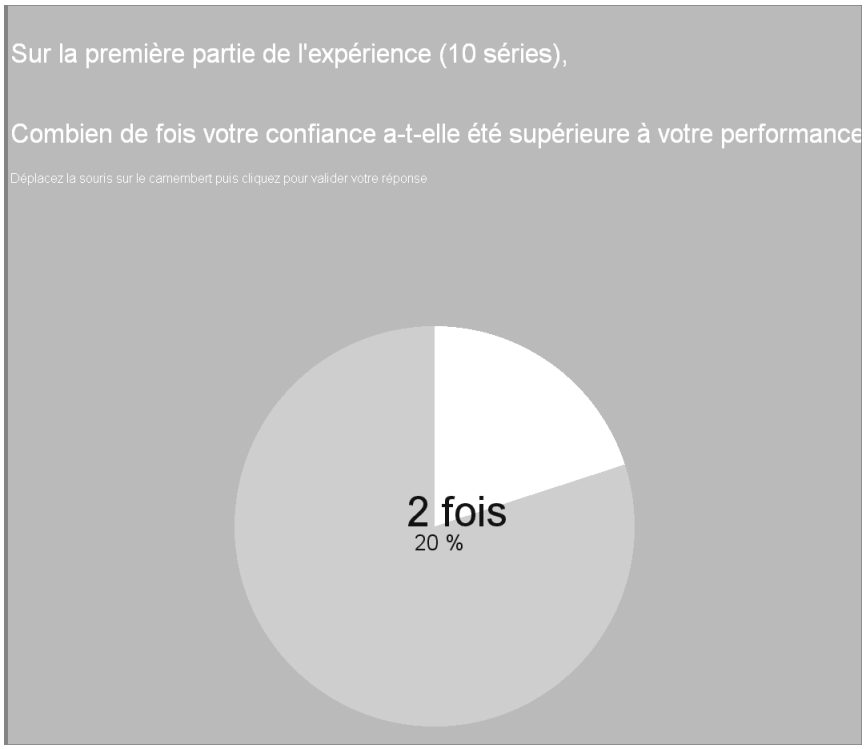

Figure 14: Screen 1 of intervention in T_2

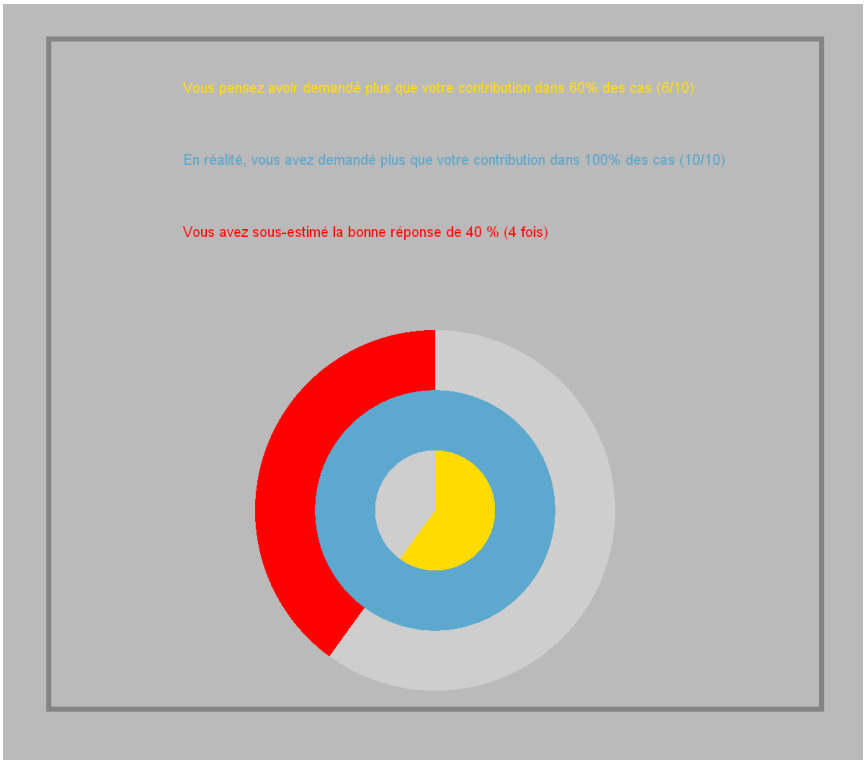

Figure 15: Screen 2 of intervention in T_2

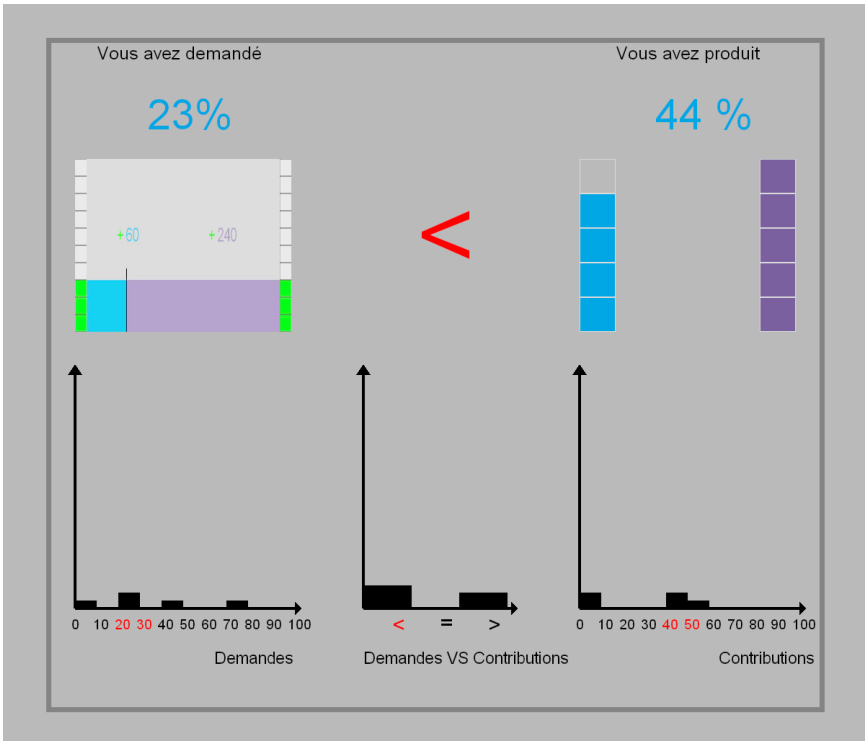

Figure 16: Screen 3 of intervention in T_2

C Difficulty calibration procedure

Recall that we denoted $\Delta\theta$ the rotation angle of the second stimulus with respect to the first and that consequently, $|\Delta\theta|$ corresponds to the difficulty of the task: the higher, the easier for participants to determine whether the rotation is clockwise or counter-clockwise. In our experiment, difficulty calibration is done in two steps. First, participants perform the perceptual task 100 times in a row in what is presented as a training. During this training, difficulty is calibrated using a standard staircase algorithm on $|\Delta\theta|$ (ASA staircase, Kesten (1958)), converging toward a level of difficulty such that the subject would have 75% of correct answers. This first step is a good way to converge towards a performance level close to the targeted one. However, it is highly sensitive to change in participants performance during the whole experiment. In our case, because the experiment is quite long, we had reasons to suspect subject's performance to be unstable during the whole experiment. Thus, in order to keep subjects at 75% correct answers during the whole experiment we used a continuously updated estimate of the function linking the difficulty of the stimulus presented and subject's performance (this function is called the psychometric function). This estimation is straightforward using the answers of the subject to the task ("$\Delta\theta > 0$" or "$\Delta\theta < 0$") as dependent variable and the associated rotation angle ($\Delta\theta$) as independent variable in the following probit model:

$$\mathbb{P}(\text{“}\Delta\theta > 0\text{”}|\Delta\theta) = \phi_{\mathcal{N}(0,1)}(\alpha_0 + \alpha_1\Delta\theta) = \phi_{\mathcal{N}(-\frac{\alpha_0}{\alpha_1}, \frac{1}{\alpha_1})}(\Delta\theta) = \phi_{\mathcal{N}(\widehat{\mu}, \widehat{\sigma})}(\Delta\theta) \tag{1}$$

Every 40 trials (or, simiarly, every 8 rounds) the psychometric function is estimated on the past 120 trials of the subject. Then, using the online estimates, $\widehat{\sigma}$ and $\widehat{\mu}$, for each series of 5 trials of the perceptual task, the subject is presented stimuli $\Delta\theta_1, ..., \Delta\theta_5$ such that

$$\frac{1}{5} \times \sum_{k=1}^{5} \left(\phi_{\mathcal{N}(-\widehat{b}, \widehat{\sigma})}(\Delta\theta_k) \times \mathbb{1}_{\{\Delta\theta_k > 0\}} + \left(1 - \phi_{\mathcal{N}(-\widehat{b}, \widehat{\sigma})}(\Delta\theta_k)\right) \times \mathbb{1}_{\{\Delta\theta_k < 0\}} \right) = 0.75 \tag{2}$$

Note that to avoid heterogeneity between participants in terms of the variability of the difficulty within each set, the series of stimuli are drawn such that the variability of stimuli difficulty is inside some given boundaries. In practice in the experiment, the average standard deviation of the difficulty of stimuli inside a given round in terms of probability to be correct was between 9.42 and 9.87 with a mean at 9.65.

D Confidence elicitation mechanism

The mechanism to incentivize confidence and revised confidence goes as follow. Once the subject has completed 5 trials of the perceptual task, he gives his confidence c between 0 and 100. Firstly, one out of the 5 trials in the set is randomly drawn. Secondly, a lottery l_1, between 0 and 100 is randomly drawn. If $l_1 < c$, meaning the lottery is worse than subject's confidence, it will be subject's answer at the randomly drawn trial that will determine if he wins or looses. If $l_1 > c$, meaning the lottery is better than subject's confidence, it is the lottery that determines if he wins or looses: subject wins with probability l_1 and looses with probability $1 - l_1$. Whatever the way subject wins or looses (through the lottery or through his own answer), he gets w points from winning and l points from loosing.

Firstly, under expected utility, this mechanism incentivizes subjects to give their average confidence. Let's denote p_k the probability that subject's answer k is correct. Given that trials k is randomly drawn, the expected utility of the player is $p_k u(w) + (1 - p_k)u(l)$ if the lottery l_1 is worse than his confidence and $\mathbb{P}(l_2 > l_1 | l_1 > c) \times u(l) + \mathbb{P}(l_2 < l_1 | l_1 > c) \times u(w)$ if the lottery is better. Then, the expected utility of the player given that trial k is randomly drawn is:

$$cp_k u(w) + c(1 - p_k)u(l) + (1 - c) \times \big(\frac{1 - c}{2}u(l) + \frac{1 + c}{2}u(w)\big) \tag{3}$$

Because each trial as the same probability $\frac{1}{5}$ to be drawn, before the draw, the expected utility of the player is:

$$E\big(U(c)\big) = \frac{1}{5}\sum_{k=1}^{5}\Big(cp_k u(w) + c(1 - p_k)u(l) + (1 - c) \times \big(\frac{1 - c}{2}u(l) + \frac{1 + c}{2}u(w)\big)\Big) \tag{4}$$

Then, the first order condition $\frac{\partial E\left(U(c)\right)}{\partial c} = 0$ gives:

$$c \times \big(u(l) - u(w)\big) + \frac{1}{5}\sum_{k=1}^{5} p_k \times \big(u(w) - u(l)\big) = 0 \tag{5}$$

$$\iff c^* = \frac{1}{5}\sum_{k=1}^{5} p_k \tag{6}$$

Secondly, this mechanism incentivizes subjects to be good at the task. It is indeed straightforward to see in equation (4) that the expected utility positively depends on the p_k's, the probability to be correct (because $u(w) > u(l)$)

E Computer's behavior

Participant X_1 / Computer X_2	0	1	2	3	4	5
0	50[a]	30	54.5	45.42	45.69	45.75
	(28.87)	(0)	(27.24)	(8.58)	(13.30)	(7.37)
1	50	50[a]	61.33	58.33	49.88	49.97
	(0)	(28.87)	(14.26)	(14.51)	(16.66)	(8.93)
2	42	52.67	56.65	53.64	53.35	55.58
	(17.89)	(5.96)	(16.25)	(9.08)	(10.74)	(11.37)
3	51.67	53.75	54.79	53.37	52.66	52.44
	(24.58)	(13.70)	(14.11)	(9.76)	(8.05)	(9.23)
4	66.39	59.63	55.99	53.15	53.26	53.79
	(28.21)	(13.76)	(11.24)	(9.12)	(9.86)	(11.33)
5	50	56.92	57.95	54.20	53.27	52.43
	(0)	(7.35)	(13.04)	(10.82)	(10.45)	(10.37)

[a]: *drawn from a uniform distribution due to absence of data*

Table 4: Mean and standard deviation of computer's claim depending on individual productions

F Testing alternative models of confidence revision

To measure the other-underestimation bias, we made the assumption that participants were updating their confidence in a Bayesian way, taking into account the revealed joint production X, their initial confidence c and some prior over the performance of the other c_{other}. An observation that would directly contradict such model would be that participants are actually insensitive to the information given to them on X to revise their confidence. Indeed, if participants don't use the link that exists between an information on the global performance of the dyad and their own performance, their revision should not depend on X. Here, we show evidence for the contrary: participants are indeed sensitive to the information about X when revising, in the way predicted by the model. To do so, we run within participant regressions on the effect of confidence and X on revised confidence.

	Revised confidence		
	Model 1	Model 2	Model 3
Confidence	0.59***	0.45***	0.52***
	(0.02)	(0.02)	(0.02)
Joint production X		6.02***	5.24***
		(0.22)	(0.24)
Observations	5408	5408	4774
F Statistic	2187.34***	5473.49***	5034.75***
Adjusted R. squared	0.27	0.67	0.67

*p<0.05; **p<0.01; ***p<0.001
Standard errors are clustered at the participant level

Table 5: Impact of initial confidence and information about X on confidence revision

The results are presented in Table 5. In model 1, we simply regress revised confidence on confidence and see that the latter positively impact the former. When we introduce the joint production X in model 2, we find the expected positive impact of X on revised confidence: when participants learn that X is higher, they increase their revised confidence. Moreover, note that by introducing X we greatly improve the fit of the model which proves the importance of X as a decision variable for participants (the adjusted R squared goes from 0.27 to 0.67). However, one could say that participants will use X only in the extreme cases when participant's confidence level is incompatible with the information about X he receives (i.e. when it is obvious that X delivers some information).[14] To check that the effect of X is not limited to those extreme cases, we run the same analysis than before without those cases in model 3. Although the coefficient in front of X slightly decreases, we find exactly the same pattern which confirms that participants used X when taking their decisions and not only when the information was obviously in contradiction with their initial confidence.

[14]For instance, suppose that participant thinks he performed at $80\% \approx 4$ correct answers. If he learns that actually $X < 4$, it is quite obvious that he should revise his confidence downward. If he learns that $X = 10$, this time it is obvious that he should revise upward. However, maybe his revision will not be sensitive to X if $X \in [\![4, 9]\!]$.

Another possibility is that participants use the feedback X but don't use Bayesian updating to revise their confidence. Here, we test two different heuristics:

- H_1: a participant with confidence c in his ability and c_{other} in the ability of the other thinks that, *a priori*, he contributed to $\frac{c}{c+c_{other}}$ of the joint production. Then he revises his confidence by only applying this *a priori* share of correct answers to the joint production X he learns.

$$\text{Revised confidence} = \frac{X}{5} \times \frac{c}{c + c_{other}}$$

- H_2: a participant with c and c_{other} who learns X might distribute his prediction error $\frac{X}{5} - (c + c_{other})$ equally between him and the other

$$\text{Revised confidence} = c + \Big(\frac{\frac{X}{5} - (c + c_{other})}{2} \Big)$$

For both models, as for our Bayesian updating model, we estimated the value c_{other} for each trial in to each participants. We then looked at the standard deviation of c_{other} across trials, and found that this was significantly higher with both heuristics than with our Bayesian model ($\overline{Std(Bayes)} = 0.181$, $\overline{Std(H_1)} = 0.227$, $\overline{Std(H_2)} = 0.223$, $t_{H_1}(207) = -13.39$, $p_{H_1} < 0.001$, $t_{H_2}(207) = -12.51$, $p_{H_2} < 0.001$). Moreover, this is true for most participants: 84% for the comparison with H_2 and 86% for the comparison with H_1. In other words, the estimated value of c_{other} was more coherent across trials in the case of our Bayesian updating model. We argue that this analysis supports our Bayesian updating model in comparison to the alternative models.

G Theoretical model

G.1 Model of choice in the production task

Individual productions are based on players' performance in the perceptual task (see section 2.1). In this task, players complete 5 different trials where they have to judge which state of the world is realized, out of two possible states (ω_1 = clockwise rotation, ω_2 = anti-clockwise rotation). As shown by Green and Swets (1966), this task can be modeled using the SDT framework. For each decision, players receive a private noisy signal S, which depends on the state of the world: if $\Omega = \omega_1$ then $S \sim \mathbb{N}(-\frac{d'}{2}, 1)$, while if $\Omega = \omega_2$, $S_i \sim \mathbb{N}(\frac{d'}{2}, 1)$. The value d' is the sensitivity of the signal to the underlying state of the world, and quantifies the amount of the information available to players. We assume that players know this generative process, but that they may misestimate the value d'. The two states of the world are *a priori* equiprobable, which is also known.

Let's denote s^k the realisation of S at trial k: the optimal decision rule i.e. the rule that maximizes players' performance is:

$$\begin{cases} \text{respond “}\omega_1\text{”} & \text{if } s^k < 0 \\ \text{respond “}\omega_2\text{”} & \text{if } s^k \geq 0 \end{cases}$$

Using this optimal rule, the expected performance, i.e. the ex-ante expectation over X_1, that is the random variable coding for whether the response was correct for players, is:

$$\mathbb{E}(X_1) = \phi(\frac{d'}{2})$$

with ϕ the cumulative distribution function of a $\mathbb{N}(0, 1)$

G.2 Model of confidence

Recall that at this stage, players do not receive any feedback on their production X_1. However, they can form a posterior belief about $\mathbb{E}(X_1)$, i.e. a sense of confidence in making the correct decision, based on signal s^k.

If players are ideal Bayesian player, this posterior belief is:

$$\mathbb{P}(\text{Decision is correct}|s^k) = \frac{e^{|d' \times s^k|}}{1 + e^{|d' \times s^k|}}$$

Then, over the 5 trials, the belief about expected performance should be:

$$\mathbb{E}(X_1|s^1, ..., s^5) = \sum_{k=1}^{5} \frac{e^{|d' \times s^k|}}{1 + e^{|d' \times s^k|}}$$

However, players may not be ideal, and their expected confidence denoted c, may differ from that of this ideal agent. In particular, they might misestimate the value of d'. If they overestimate it, c will be higher than $\mathbb{E}(X_1|s^1, ..., s^5)$ and if they underestimate it, c will be lower. We then define the overconfidence bias as follows:

$$\text{Overconfidence Bias} = p_i - \mathbb{E}(X_1|s^1, ..., s^5)$$

In our simulations, we vary the value of d' to generate different levels of overconfidence bias.

G.3 Model of revised confidence

Once confidence is elicited, players learn the joint production X that is the sum of the number of correct answers produced by himself X_1 and the computer X_2. Players will bargain over this joint production. The share that players will claim during bargaining might be based on an estimation of their own production. We describe here how the information should be used by players to update the belief about the individual production and revise their confidence judgment. Optimally, the revised belief of players should be:

$$\mathbb{E}(X_1|s^1,...,s^5,X) = \sum_{n=0}^{5} n \times \mathbb{P}(X_i = n|s^1,...,s^5,X)$$

Using Bayes' rule, we know that:

$$\mathbb{P}(X_1 = n|s^1,...,s^5,X) = \frac{\mathbb{P}(X_1 = n|s^1,...,s^5) \times \mathbb{P}(X_2 = X - n)}{\mathbb{P}(X_1 + X_2 = X|s^1,...,s^5)}$$

Thus,

$$\mathbb{E}(X_1|s^1,...,s^5,X) = \sum_{n=0}^{5} n \times \frac{\mathbb{P}(X_1 = n|s^1,...,s^5) \times \mathbb{P}(X_2 = X - n)}{\mathbb{P}(X_1 + X_2 = X|s^1,...,s^5)}$$

Computer's performance X_2 actually follows a $B(5, c_{other} = 0.75)$ distribution in our experiment. However, players may not estimate correctly c_{other}, and their expected revised confidence may differ from that of this ideal agent. In particular, if $\widehat{c_{other}} < 0.75$, their revised confidence will be higher than the optimal one, and conversely if $\widehat{c_{other}} > 0.75$. Therefore, we define the other-underestimation bias, as the difference between the true probability of success of the computer and its estimation by players:

$$\text{Other-underestimation bias} = 0.75 - \widehat{c_{other}}$$

In our simulations, we vary the level of $\widehat{c_{other}}$ to generate different levels of other-underestimation bias.

H Regression tables

Intervention effect on bargaining outcomes

See Table 6 and 7

	Δ Disagreement		Δ Disagreement (raw)		Δ Claim		Δ Mismatch	
	Estimate	Std	Estimate	Std	Estimate	Std	Estimate	Std
Baseline	*Ref*	(.)	*Ref*	(.)	*Ref*	(.)	*Ref*	(.)
T_1	4.20	(2.67)	5.71†	(3.39)	1.69†	(0.96)	2.88†	(1.56)
T_2	4.13	(2.61)	3.40	(3.31)	1.93*	(0.94)	2.91†	(1.53)
T_3	3.76	(2.81)	2.22	(3.56)	1.29	(1.01)	0.65	(1.64)
T_4	8.16**	(2.66)	8.15*	(3.37)	2.78**	(0.96)	3.61*	(1.65)
Constant	5.57**	(1.84)	6.01*	(2.33)	0.26	(0.66)	−0.35	(1.07)
Observations	218		218		218		218	
F Statistic	2.37†		1.71		2.28†		2.01†	

Note: †p<0.1; *p<0.05; **p<0.01;***p<0.001

Table 6: Regression of disagreement index, average disagreement, average claim and mismatch index decrease by treatment (reference=Baseline)

Intervention effect on cognitive biases and sensitivity to entitlements

See Table 8 and 9

Gender differences

See Table 10 for the regression table

	Δ Disagreement		Δ Disagreement (raw)		Δ Claim		Δ Mismatch	
	Estimate	Std	Estimate	Std	Estimate	Std	Estimate	Std
Baseline	-8.16^{**}	(2.66)	-8.15^{*}	(2.37)	-2.78^{**}	(0.96)	-3.61^{*}	(1.55)
T_1	-3.96	(2.73)	-2.44	(3.46)	-1.09	(0.98)	-0.73	(1.59)
T_2	-4.03	(2.67)	-4.75	(3.39)	-0.85	(0.96)	-0.70	(1.56)
T_3	-4.40	(2.86)	-5.93	(3.63)	-1.49	(1.03)	$-2.96^{\dagger}$	(1.67)
T_4	*Ref*	(.)	*Ref*	(.)	*Ref*	(.)	*Ref*	(.)
Constant	13.73^{***}	(1.92)	14.16^{***}	(2.43)	3.04^{***}	(0.69)	3.26^{**}	(1.12)
Observations	218		218		218		218	
F Statistic	$2.37^{\dagger}$		1.71		$2.28^{\dagger}$		$2.01^{\dagger}$	

Note: †p<0.1; *p<0.05; **p<0.01;***p<0.001

Table 7: Regression of disagreement index, average disagreement, average claim and mismatch index decrease by treatment (reference=T_4)

	Δ OP bias		Δ OC bias		Δ OU bias		Δ Sensitivity	
	Estimate	Std	Estimate	Std	Estimate	Std	Estimate	Std
Baseline	*Ref*	(.)	*Ref*	(.)	*Ref*	(.)	*Ref*	(.)
T_1	−0.05	(1.00)	−1.39	(1.55)	0.39	(1.51)	0.28	(0.18)
T_2	0.35	(0.98)	−0.17	(1.51)	2.26	(1.48)	0.19	(0.18)
T_3	0.68	(1.05)	0.98	(1.62)	2.19	(1.59)	0.01	(0.19)
T_4	2.03*	(0.99)	1.97	(1.54)	4.22**	(1.50)	0.21	(0.19)
Constant	0.54	(0.69)	2.60*	(1.06)	−2.13*	(1.04)	−0.13	(0.12)
Observations	218		218		218		218	
F Statistic	1.42		1.28		2.48*		0.92	

Note: †p<0.1; *p<0.05; **p<0.01;***p<0.001

Table 8: Regression of overplacement bias, overconfidence bias, other-underestimation bias and sensitivity to entitlements by treatment (reference=Baseline)

	Δ OP bias		Δ OC bias		Δ OU bias		Δ Sensitivity	
	Estimate	Std	Estimate	Std	Estimate	Std	Estimate	Std
Baseline	-2.03^{*}	(0.99)	-1.97	(1.54)	-4.22^{**}	(1.50)	-0.21	(0.19)
T_1	-2.08^{*}	(1.02)	-3.36^{*}	(1.58)	-3.83^{*}	(1.55)	0.07	(0.19)
T_2	$-1.68^{\dagger}$	(0.99)	-2.14	(1.54)	-1.96	(1.51)	-0.03	(0.19)
T_3	-1.35	(1.07)	-0.99	(1.65)	-2.03	(1.62)	-0.20	(0.20)
T_4	*Ref*	(.)	*Ref*	(.)	*Ref*	(.)	*Ref*	(.)
Constant	2.57^{***}	(0.72)	4.57^{***}	(1.11)	$2.08^{\dagger}$	(1.09)	0.09	(0.15)
Observations	218		218		218		218	
F Statistic	1.42		1.28		2.48^{*}		0.92	

Note: †p<0.1; *p<0.05; **p<0.01;***p<0.001

Table 9: Regression of overplacement bias, overconfidence bias, other-underestimation bias and sensitivity to entitlements by treatment (reference=T_4)

	Δ Disagreement	Δ Mismatch	Δ OP bias	Δ OC bias	Δ OU bias	Δ Sensitivity
T_1	0.83	2.29	0.33	0.21	−1.04	0.14
	(3.91)	(2.31)	(1.48)	(2.30)	(2.24)	(0.26)
T_2	0.67	3.36	−0.32	−0.31	2.10	−0.04
	(3.69)	(2.11)	(1.35)	(2.11)	(2.05)	(0.24)
T_3	−1.70	0.51	0.32	1.21	2.02	0.08
	(3.75)	(2.21)	(1.42)	(2.20)	(2.15)	(0.25)
T_4	0.39	1.54	0.21	0.47	1.68	0.18
	(3.62)	(2.14)	(1.37)	(2.13)	(2.07)	(0.27)
Women	−7.10†	0.85	−0.39	0.71	−0.61	−0.25
	(3.62)	(2.14)	(1.37)	(2.13)	(2.07)	(0.25)
Women × T_1	6.78	0.89	−0.60	−2.85	2.54	0.29
	(5.31)	(3.13)	(2.00)	(3.12)	(3.04)	(0.36)
Women × T_2	6.91	−0.90	1.42	0.36	0.30	0.48
	(5.16)	(3.04)	(1.95)	(3.03)	(2.95)	(0.35)
Women × T_3	11.67*	0.51	0.79	−0.40	0.29	−0.19
	(5.58)	(3.29)	(2.11)	(3.28)	(3.20)	(0.38)
Women × T_4	16.38**	6.65	3.96*	3.37	5.52†	0.08
	(5.25)	(3.10)	(1.98)	(3.08)	(3.01)	(0.38)
Constant	9.12***	−0.77	0.74	2.25	−1.83	0.00
	(2.56)	(1.51)	(0.97)	(1.50)	(1.47)	(2.17)
Observations	218	218	218	218	218	218
F Statistic	2.30*	1.68†	1.45	1.08	1.77†	0.92
Note:		†p<0.1; *p<0.05; **p<0.01;***p<0.001				

Table 10: Regression of disagreement index, mismatch index, overplacement bias, overconfidence bias, other-underestimation bias and sensitivity to entitlements by treatment interacted with gender

I Robustness to temporal dependance between overconfidence and other-underestimation bias

One concern about the estimation of the other-underestimation bias is that the estimated performance of the other player (c_{other}) at a given period might actually depend on the overconfidence bias at the previous period, if participants were adjusting this estimated performance of the other player from one trial to the next. To understand why, take for instance a participant who is overconfident in his performance but evaluates the other correctly. In round 1, due to overconfidence, he will overestimate his contribution ot the joint output. Then, when getting feedback on X, he will underestimate the contribution of the other participant (although *a priori* he was estimating it correctly) and as a result his posterior about the other's performance will be underestimated. In round 2, if this participant updates his prior about the other using this posterior from round 1, this will make him lower it and thus we will measure a positive other-underestimation bias. Round after round, we will measure an increasing other-underestimation bias which will actually be the result of a learning process using feedback X generated with an initial overconfidence bias.

Here, we check that our results are unchanged if we take into account this trial-to-trial adjustment. Specifically, we estimate, for each participant, the following model based on a classic Q-learning rule, which is commonly used in the learning literature (Sutton and Barto (1998)).

$$\mathrm{c}_{\mathrm{other}}^{t+1,prior} = \mathrm{c}_{\mathrm{other}}^{t,prior} + \theta \times (\mathrm{c}_{\mathrm{other}}^{t,post} - \mathrm{c}_{\mathrm{other}}^{t,prior})$$

with:

- $\mathrm{c}_{\mathrm{other}}^{t,prior}$ the prior belief about the computer's production at the beginning of round t
- $\mathrm{c}_{\mathrm{other}}^{t,post} = \frac{1}{5} \times (1 - \mathrm{entitlement}_t) \times X_t$: given X, it is the posterior belief about the computer's production at the end of round t
- θ the learning rate

Across participants, we find no significant differences between the prior computed with our method and the prior at the first period estimated with this model ($t(217) = -1.43, p = 0.15$). This comes from the fact that the average θ parameter we estimate is quite low, equal to $\overline{\theta} = 0.01$.

We also check that our treatment effect on other-underestimation bias is robust to this updating process, by adding to the previous model a treatment effect T at the period 27 just after the intervention.

$$\mathrm{c}_{\mathrm{other}}^{t+1,prior} = \mathrm{c}_{\mathrm{other}}^{t,prior} + \theta \times (\mathrm{c}_{\mathrm{other}}^{t,post} - \mathrm{c}_{\mathrm{other}}^{t,prior}) + T \times \mathbf{1}_{t=27}$$

We replicate our previous results. In particular, we estimate an average shock on the prior about other of 5.59 in T_4 and of 1.17 in the baseline. This difference between T_4 and baseline is significant ($t(90) = 2.02, p = 0.046$). We also replicate our mediation analysis and find that this shock on the prior about the other mediates the intervention effect in T_4 vs baseline ($p = 0.044$) even though the mediated share is lower (17.55% against 30.02% previously).

References

Ais, J., Zylberberg, A., Barttfeld, P., and Sigman, M. (2015). Individual consistency in the accuracy and distribution of confidence judgments. *Cognition*, 146:377–386.

Arkes, H. R., Christensen, C., Lai, C., and Blumer, C. (1987). Two methods of reducing overconfidence. *Organizational Behavior and Human Decision Processes*, 39(1):133 – 144.

Babcock, L. and Loewenstein, G. (1997). Explaining bargaining impasse: The role of self-serving biases. *Journal of Economic Perspectives*.

Bang, D., Aitchison, L., Moran, R., Herce Castanon, S., Rafiee, B., Mahmoodi, A., Lau, J. Y. F., Latham, P. E., Bahrami, B., and Summerfield, C. (2017). Confidence matching in group decision-making. *Nature Human Behaviour*, 1(6):0117.

Bazerman, M. H. and Moore, D. A. (2012). *Judgment in Managerial Decision Making*. John Wiley and Sons.

Bazerman, M. H. and Sondak, H. (1988). Judgmental limitations diplomatic negotiations. *Negotiation Journal*, 4(3):303–317.

Becker, G. M. and DeGroot, M. H. (1974). *Measuring Utility by a Single-Response Sequential Method (1964)*, pages 317–328. Springer Netherlands, Dordrecht.

Berlin, N. and Dargnies, M.-P. (2016). Gender differences in reactions to feedback and willingness to compete. *Journal of Economic Behavior & Organization*, 130:320–336.

Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2019). Beliefs about gender. *American Economic Review*, 109(3):739–73.

Camerer, C. and Lovallo, D. (1999). Overconfidence and excess entry: An experimental approach. *American Economic Review*, 89.

Clark, J. and Friesen, L. (2009). Overconfidence in Forecasts of Own Performance: An Experimental Study*. *The Economic Journal*, 119(534):229–251.

Dickinson, D. L. (2009). The effects of beliefs versus risk attitude on bargaining outcomes. *Theory and Decision*, 66(1):69–101.

D'Exelle, B., Gutekunst, C., and Riedl, A. (2017). Gender and bargaining. *WIDER Working Paper*.

Enke, B. and Graeber, T. (2019). Cognitive Uncertainty. *NBER Working Paper*.

Gilligan, C. and Snider, N. (2018). *Why Does Patriarchy Persist?*

Gino, F. and Moore, D. A. (2007). Effects of task difficulty on use of advice. *Journal of Behavioral Decision Making*, 20(1):21–35.

Green, D. M. and Swets, J. A. (1966). *Signal detection theory and psychophysics*. John Wiley.

Grossman, Z. and Owens, D. (2012). An unlucky feeling: Overconfidence and noisy feedback. *Journal of Economic Behavior & Organization*, 84(2):510 – 524.

Johnson, D. D., McDermott, R., and Barrett, E. S. (2006). Overconfidence in wargames: experimental evidence on expectations, aggression, gender and testosterone. *Proc Biol Sci*.

Karagözoğlu, E. and Riedl, A. (2015). Performance information, production uncertainty, and subjective entitlements in bargaining. *Management Science*.

Kelemen, W., Frost, P., and Weaver, C. (2000). Individual differences in metacognition: Evidence against a general metacognitive ability. *Memory and cognition*, 28:92–107.

Kesten, H. (1958). Accelerated stochastic approximation. *Ann. Math. Statist.*, 29(1):41–59.

Lichtenstein, S. and Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance*, 26(2):149 – 171.

Massoni, S., Gajdos, T., and Vergnaud, J.-C. (2014). Confidence measurement in the light of signal detection theory. *Frontiers in Psychology*.

Moore, D. A. and Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*.

Neale, M. A. and Bazerman, M. H. (1985). The effects of framing and negotiator overconfidence on bargaining behaviors and outcomes. *Academy of Management Journal*.

Niederle, M. and Vesterlund, L. (2007). Do Women Shy Away From Competition? Do Men Compete Too Much?*. *The Quarterly Journal of Economics*, 122(3):1067–1101.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*.

West, R. and Stanovich, K. (1997). The domain specificity and generality of overconfidence: Individual differences in performance estimation bias. *Psychonomic Bulletin and Review*, 4:387–392.