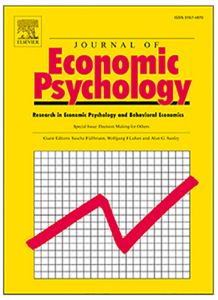
I did most of the work! Three sources of bias in bargaining with joint production

Quentin Cavalan, Vincent de Gardelle, Jean-Christophe Vergnaud

Journal Pre-proof



I did most of the work! Three sources of bias in bargaining with joint production

Quentin Cavalan, Vincent de Gardelle, Jean-Christophe Vergnaud

PII: S0167-4870(22)00077-0
DOI: https://doi.org/10.1016/j.joep.2022.102566
Reference: JOEP 102566

To appear in: *Journal of Economic Psychology*

Received date : 12 May 2021
Revised date : 15 September 2022
Accepted date : 15 September 2022

Please cite this article as: Q. Cavalan, V. de Gardelle and J.-C. Vergnaud, I did most of the work! Three sources of bias in bargaining with joint production. *Journal of Economic Psychology* (2022), doi: https://doi.org/10.1016/j.joep.2022.102566.

I did most of the work!
Three sources of bias in bargaining with joint production*

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

[1]Corresponding author, CES, Université Paris 1, PSE; 57 rue de la Commune de Paris, 93300 Aubervilliers, France; quentin.cavalan@hotmail.fr
[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 less commonly investigated empirically. Here, we show that overplacement leads to conflict in these situations: individuals overestimate their contribution to the joint production and consequently make unreasonable claims. We further decompose overplacement into three types of cognitive biases: overestimation of one's own production (i.e. overconfidence bias), underestimation of others' production (i.e. superiority bias) and biases in information processing. We show that they all contribute to overplacement. To quantify these biases, we develop a novel experimental setting using a psychophysically controlled production task within a bargaining game, where we elicit participants' subjective estimation of their performance, both before and after they receive information about the joint production. In addition, we test several interventions to mitigate these biases, and successfully decrease disagreements and overplacement through one of them. 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

This work was supported by the Agence Nationale de la Recherche [ANR-16-CE28-0002]

*Data are available and can be downloaded at https://osf.io/b8z4m/

1 Introduction

Many studies in behavioral economics and psychology have established that agents often overestimate their skills relative to others, which can lead to economic failures. For example, entrepreneurs overestimate their chances of success when entering a market, leading to excess entry (Camerer and Lovallo, 1999). Individuals who believe themselves 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 where this overplacement bias comes from.

In this paper, we focus in particular on situations of bargaining over a joint production where there is uncertainty surrounding the contributions of both parties. These situations are in fact ubiquitous in day-to-day life. For instance, in a household, the contributions of each member are uncertain, since they include not only wages but also time, sacrificed career opportunities, participation in chores, etc. As such, both members may overestimate their own contribution to the household's wealth and/or they might fail to fully acknowledge the contribution of the other member, leading to disagreements over the merit of each member. Similarly, in collective bargaining between labor unions and management, parties may have incompatible beliefs about their respective merits and consequently about the distribution of wealth between employees and share-holders. Such overestimation on both sides would lead to disagreements and bargaining inefficiencies.

Starting with Gantner et al. (2001) and Cherry et al. (2002), an important literature has recently flourished on bargaining problems over joint production (Cappelen et al., 2007; Konow et al., 2009; Cappelen et al., 2010; Karagözoğlu and Riedl, 2015; Fischbacher et al., 2017; Soldà et al., 2021; Santos Pinto and Colzani, 2021). This literature has mostly focused on fairness issues, using games such as the dictator or ultimatum game, with a very limited strategic component (Cappelen et al., 2010). Moreover, most studies in this literature have used experimental designs where participants have perfect knowledge about their individual production, either because they receive feedback about it (Konow et al., 2009; Fischbacher et al., 2017) or because production is chosen explicitly (Gantner et al., 2001; Cappelen et al., 2007). How individuals bargain when they are uncertain about their own production, by contrast, has been mostly unexplored. Critically though, in these situations uncertainty leaves room for each party to overestimate their relative contribution. Furthermore, the different mechanisms that can generate this overplacement bias have not been clearly identified. Overconfidence in one's own skills might lead individuals to overestimate their contribution but other cognitive biases may also play a role, such as biases in how individuals estimate the production of others, or biases in how they update their beliefs when receiving information about the production. These different biases may all contribute to overplacement and ultimately affect bargaining outcomes.

To the best of our knowledge, three experimental studies have investigated bargaining over a joint production with real effort made under uncertainty. Karagözoğlu and Riedl (2015) have shown that providing participants with information about their relative performance leads them to increase their entitlements, which impact first proposals, bargaining duration and settlements. Soldà et al. (2021) have shown that overplacement generates costly delays and disagreements in bargaining. Similarly, Santos Pinto and Colzani (2021) have found higher rates of bargaining failure when participants exhibit overplacement.

Whereas these previous studies have mainly focused on the impact of overplacement on bargaining outcomes, here we further seek to disentangle the various biases which may lead to overplacement and ultimately bargaining inefficiencies. More precisely, in the present study we shall decompose overplacement in three cognitive biases: 1) an overconfidence bias by which individuals overestimate their production, 2) a superiority bias by which they estimate the production of others as lower than their own, 3) updating biases by which they update their beliefs in a non-Bayesian manner following feedback about the joint production.

In our experiment, participants produced joint wealth by performing a task, and then shared this wealth through a particular version of a Nash demand game. In case of a disagreement, payoffs depended on the true performance of individuals, which was not known with certainty by participants beforehand. If however claims added up to an amount less than the joint production, the remainder was not lost but shared in proportion to the claims of the two players.[1] After the production task (but before making a claim in the bargaining game), participants also provided an initial estimate of their own performance by means of a confidence judgment, then received feedback about the joint production, after which they could revise their confidence. The production and bargaining game were repeated several times, and in the middle of the experiment an intervention was conducted to provide information to participants about their behavior.

Our first goal was to identify different routes to overplacement: overestimation of self, underestimation of others, and belief updating biases. To this end, we built a model in which beliefs about one's own performance are affected by overconfidence, beliefs about the performance of others are affected by a superiority bias and ultimately, updating biases translate those beliefs into a final evaluation of performance. To estimate this model, we directly measured overconfidence by comparing the initial confidence of each participant to their actual performance in the production task. In addition, taking advantage of the multiple rounds in our design, we estimated a superiority bias and updating biases by examining how participants revised their confidence. Our second goal was to evaluate how these biases may change after the intervention, when participants are told about their overall contribution or about their overplacement bias. We evaluated how such interventions affect overplacement, and ultimately bargaining outcomes.

The primary outcome of the negotiation is whether an agreement was reached or not, and we defined the *disagreement index* as the probability that the negotiation would fail. Quite obviously, a social planner might want to limit disagreements because they are costly: when the two parties have to resort to court, dead-weight losses are incurred by society. However, settlements might not always be desirable if they do not reflect the respective contributions of individuals: a social planner might want individuals to claim what they have produced so that they get what they deserve in case of an agreement. Thus, to check how individuals' claims correspond to their true contribution, we also considered the distance between claims and contributions, hereafter labelled *mismatch index*.

With respect to previous studies on bargaining, our paradigm also introduced two original methodological features. Firstly, our production task was based on a perceptual task and we used psychophysical methods to obtain comparable conditions of performance for all participants. Given that overconfidence heavily depends on the difficulty of the task (Kruger and Dunning, 1999; Grieco and Hogarth, 2009; Santos Pinto and Colzani, 2021), this ensured that heterogeneity in performance was not an important confound in our setup when assessing cognitive biases. Secondly, our participants interacted with a computer-simulated human. Indeed, because multiple Nash equilibria often coexist in bargaining games, the process through which an equilibrium is selected might be an important confound of participants' behavior. To mitigate this issue, we first performed a pilot study, in which participants interacted with other humans. Then, we performed the main study in which we used the behavior of participants in the pilot to program the computer's behavior. This allowed us to better isolate the impact of biases on bargaining behavior by reducing strategic uncertainty in the bargaining game.

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 document overplacement and its effect

[1] We introduced this feature to be closer to a real legal proceeding in which what is not claimed is certainly not lost. Moreover, this might reduce the appeal of always opting for fifty-fifty splits which reduce statistical power if they occur too much.

on bargaining outcomes. Second, we decompose overplacement into overconfidence bias, superiority bias and updating biases and find that overplacement is driven by these three components. Third, we evaluate the effect of our interventions on bargaining outcomes and overplacement. Fourth, we highlight gender differences in response to our interventions. Finally, section 4 discusses our results and avenues for future research.

2 Design, predictions and measures

2.1 Participants and computers

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 main experiment involved 234 participants in 15 experimental sessions of 1 hour and 30 minutes each.

In the main experiment, participants played against a computer, which mimicked the behavior of participants in the pilot study. This design helped us obtain more homogenous conditions across participants in our main study, thereby increasing statistical power. Indeed, this design reduces the strategic complexity of the game from the participants' perspective (as the behavior of the other player is known to be stationary), and simplifies the process by which participants may converge (or not) on one of many possible equilibria in the game.[2] This allowed us to focus on the main question that is the role of cognitive biases. Besides, although we are aware that participants may adopt different behavior when interacting with humans vs. machines (Devaine et al., 2014; March, 2021), the data from the pilot experiment (reported in Online Appendix 5) showed similar effects of overplacement on bargaining outcomes as in our main experiment.

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 training for all parts of the experiment. At the end of the experiment, 2 rounds were randomly drawn to determine participants' earnings.

2.2 The bargaining with joint production game

The core of the experiment consists of a bargaining game over a joint production, briefly summarized here and illustrated in Figure 1, with details for each step presented in the following subsections. In short, in each round of the game, participants were paired with a computer mimicking the behavior of participants in the pilot study (computer's behavior is detailed below in section 2.3). Participants were aware that they were playing against a programmed computer. In each round, participants first completed the production task, and then rated their confidence about their own production. Then, participants learnt the exact joint production over the pair, they reported a revised confidence judgment, and they made a claim over the joint production. The round ended with some feedback about the outcome of the game. Participants completed 2 sets of 26 rounds of this game, with an intervention (described in section 2.4) in between the 2 sets.

[2]For instance, always claiming 100% of the joint production, or claiming a fixed high or low amount depending on whether one's own production is estimated to exceed a certain threshold, are already simple Nash equilibria, but there are many more equilibria. When participants play against each other, the selection of one of these equilibria could take a long time and render the link between cognitive biases and behavior virtually unidentifiable. The case is even more problematic when assessing the effect of a cognitive intervention, as any intervention effect could result from a change in how participants expect others to behave following the intervention, rather than an effect on cognitive biases. Although human players' beliefs and expectations can still lead to multiple equilibria in our main experiment, introducing a computer player, whose strategy is known to be constant, helps mitigate those issues.

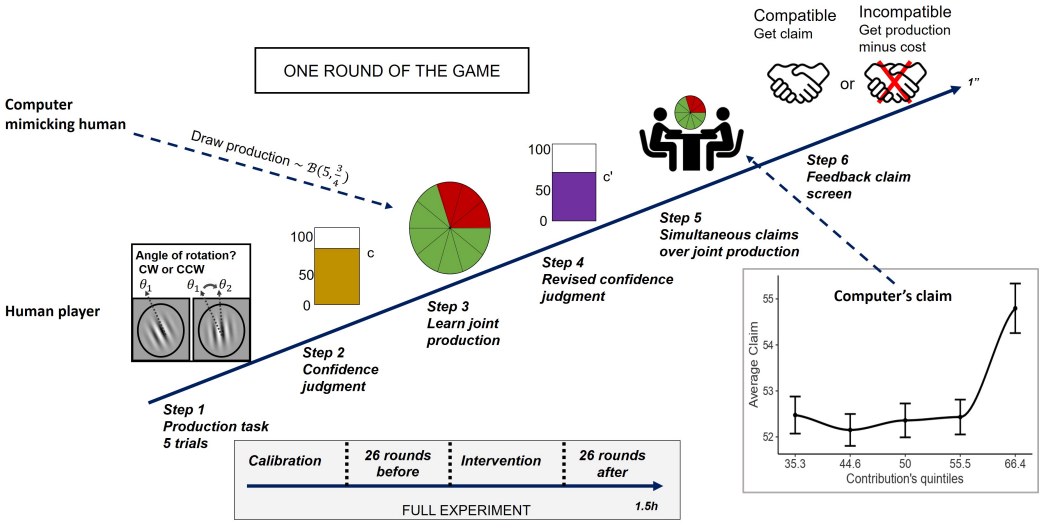


Figure 1: Schematic representation of the experimental procedure for one round of the game and timeline of the experiment. See section 2.3 for details about computer's claim.

The real-effort production task

In each round, the production task consisted of 5 trials of a simple visual task. On each trial, two stimuli (Gabor patches) with different orientations (θ_1 and θ_2) appeared sequentially and for 500 ms on the screen. Participants had to judge whether the second stimulus was oriented clockwise ($\theta_2 - \theta_1 = \Delta\theta > 0$) or counter-clockwise ($\Delta\theta < 0$) with respect to the first stimulus. They responded 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 received no feedback about X_1 at this stage.

The difficulty of the task was determined by $|\Delta\theta|$: the larger the difference between the two angles, the easier. We calibrated this quantity for each participant to aim for 75% of correct responses, based on individual performance in the 100 practice rounds.[3] Participants were not made aware that difficulty was calibrated.

Confidences elicitation

In each round, after the production phase, participants had to give their confidence regarding their answers, on a continuous scale from 0% to 100%. This confidence judgment was incentivized with a canonical BDM mechanism (Becker et al., 1964) applied to one randomly selected perceptual decision in the round. Under expected utility, this mechanism incentivized participants to report the probability that a randomly chosen trial in the round was correct, or equivalently, the average probability of being correct across trials in the round. Additionally, it incentivized them to perform the task as accurately as possible.[4] Participants were trained with this incentive mechanism before starting the main experiment.

Then, participants learnt the joint production of the dyad X for that round, which is the sum of the number of correct answers for the participant, noted X_1, and for the computer , noted X_2, that is: $X = X_1 + X_2$.

[3] See Online Appendix 1 for further details on the difficulty calibration procedure.
[4] See Online Appendix 2 for further details on the mechanism.

Following this information, participants had to give again their confidence in their performance, now knowing X. This "revised confidence", was incentivized in the same way as the initial confidence judgment.[5]

The bargaining game

Finally, participants bargained with the computer over the joint production X. The bargaining game was a slightly modified version of a one-shot Nash Demand Game, where payoffs were proportional to individual production (minus a cost) if claims couldn't be satisfied. More precisely, the participant and the computer made simultaneous claims (Claim_1 and Claim_2) which represent the share of the joint production they wanted for themselves. If claims were compatible i.e. $\text{Claim}_1 + \text{Claim}_2 \leq 100\%$, an agreement was reached and joint production was shared in proportion to the two claims. Otherwise, if $\text{Claim}_1 + \text{Claim}_2 > 100\%$, players settled in court, where they received their own contribution minus a cost of 50.

In both cases, participants were informed about the claim of the computer, and whether an agreement was reached. In case of an agreement, participants also learnt the number of points they received. In case of a disagreement, they were simply reminded about the rule that converted their (unknown) performance into points. Overall, participants received the same amount of information regarding their individual performance, irrespective of whether they agreed or disagreed. Then, once the bargaining game was over, the next round started.

2.3 Computer's behavior

In each round, the number of correct answers for the computer X_2 was drawn from a binomial distribution $\mathscr{B}(5, 0.75)$. Computer's accuracy was thus similar to participants'. For the bargaining phase, the computer's claim Claim_2 was drawn from participants' claims in the pilot experiment in similar conditions, that is, in rounds where X_2 and X_1 had the same values as in the current round. The inset in Figure 1 illustrates claims as a function of true contribution for participants in the pilot experiment (and thus for computers in the main experiment), showing a clear increase in claims when contributions exceed 55% of the total. The mean and standard deviations of the computer's claims conditional on (X_1, X_2) are presented in Online Appendix 4.

2.4 The intervention

In all experimental sessions, after 26 rounds of the bargaining game, participants were asked questions and received feedback about one or several of the variables of the game. There were 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 asked Q_0: "In your opinion, what was your average claim in the previous rounds of bargaining?"
- In treatment T_1, we asked 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 asked Q_2: "In your opinion, how many times was your claim higher than your contribution in the previous rounds of bargaining?"

[5] Because participants' earnings from confidence and revised confidence are drawn from the same round, one might worry that risk averse subjects hedge with their stated belief for one confidence elicitation against the other. However, such hedging is unlikely in our experiment as it would only be expected for extreme levels of risk aversion (see Online Appendix 3 for further details).

- In treatment T_3, we asked 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 asked Q_1, Q_2 and Q_3 altogether.

In each session, the experimenter explained in detail the question(s) asked to participants and the feedback they would obtain. Participants then answered the question(s), and obtained two different types of feedback: one screen showed the correct answer along with their actual answer, and the subsequent screen showed the entire distribution of the variable at stake (e.g. in T_1 the distribution of the participant's past contributions) with a visual animation. The duration of all interventions was the same (approximately 10 minutes).

The goal of these interventions was to impact participants' beliefs in order to affect their bargaining behavior. In T_1, we targeted participants' tendency to overestimate their contribution to the joint output, to make participants aware that, on average, they were contributing less than what they thought. In T_2 we also targeted participants' estimation of their contribution, but now with an explicit link with their claims, to make them realize that they might have made unreasonable claims with respect to their contribution. Finally, T_3 was designed to assess more directly how participants' knowledge of their individual performance, independently of the other player, impacted their bargaining behavior.

Finally, our baseline condition involves a question about past claims, to control for potential effects of time or increased motivation after the intervention, as we do not expect participants to make systematic mistakes on this question or to change their behavior based on this intervention.[6]

2.5 Main measures

We used two measures for beliefs about performance: overconfidence and overplacement. In addition, we defined two measures regarding the outcomes of the bargaining game: a disagreement index and a mismatch index. Finally, we also defined a measure of how claims are sensitive to participants' beliefs about their contribution to the joint production.

Overconfidence and overplacement biases

For each participant, we defined overconfidence bias as the mean difference between the initial judgment of confidence and the actual performance.

$$\text{Overconfidence bias} = \overline{\text{confidence} - \text{actual performance}}$$

Overplacement bias was defined as the mean difference between revised confidence once the joint production is known and actual performance, a positive overplacement bias indicating that participants overestimated the relative share they produced.

$$\text{Overplacement bias} = \overline{\text{revised confidence} - \text{actual performance}}$$

[6] This can be checked on the data of the pilot experiment where similarly to this baseline condition, we ask this question in the middle of the experiment. We find no systematic mistake: participants slightly underestimate it but this is not significant ($\overline{\text{Error}_{Q_0}} = -2.18\%,\ t(32) = -1.50,\ p = 0.143$). Moreover, we do not find any significant correlation between mistakes made at this question and participants' change in claims after the intervention ($\text{Cor} = 0.09,\ t(31) = 0.48,\ p = 0.633$).

Disagreement index

The main outcome of bargaining was whether players reached an agreement or not. Here, to smooth out the effect of the randomly drawn claim C_2 of the computer, we considered the expected disagreement corresponding to the claim of the participant given X_1 and X_2 the production of each player.[7] The disagreement index is the average of the expected disagreement over bargaining rounds. It lies between 0% and 100%. From a social planner perspective, the closer to 0 the better.

$$\text{Disagreement index} = \overline{\mathbb{E}_{C_2}\Big(\mathbf{1}_{\{C_1+C_2>100\}}|C_1, X_1, X_2\Big)}$$

Mismatch index

The goal of our interventions was to decrease the disagreement index. However, another measure of interest was the distance between participants' claims and their actual contributions. We thus defined a mismatch index as the average squared difference between participant's claims and actual contribution, divided by 25, such that on average it lied between 0 and 100.[8] Here again, from a social planner perspective, the closer to 0, the better.

$$\text{Mismatch index} = \frac{1}{25}\overline{\Big(\text{actual contribution} - \text{claim}\Big)^2}$$

Sensitivity to perceived contribution

Finally, we measured how participants' beliefs related to their claims in the bargaining game. First, given participants' revised confidence, the perceived contribution is defined as the estimated production (5 times their revised confidence) divided by the joint production.

Then, participant's sensitivity to their perceived contribution was defined as the effect of these perceived contributions on claims, in a linear regression (least-square). The higher this measure, the more participants were sensitive to their perceived contributions when making claims in the bargaining game. The objective of this measure was to check whether the interventions changed participants' claims, regardless of their beliefs.

$$\text{Claim sensitivity to perceived contribution} = \frac{Cov(\text{claim}, \text{perceived contribution})}{Var(\text{perceived contribution})}$$

[7] Our results were similar when using the average number of disagreements by participants instead.

[8] Indeed, since participants were calibrated, actual contribution was equal to 50% on average. Thus, when claiming 100% (or 0%) of the joint production, the squared distance between claim and actual contribution would be 2500.

3 Results

In this section, we first describe the effect of overplacement bias on disagreement and mismatch indices, using the pre-intervention data.[9] We then introduce a model for revised confidence in which the role of several cognitive biases (overconfidence bias, superiority bias and biases in belief updating) can be quantified. We then study the intervention effect on biases and bargaining outcomes. Finally, we provide some striking elements about gender differences in the reaction to interventions.

3.1 Overconfidence, overplacement and bargaining outcomes (before intervention)

Figure 2 illustrates the distributions of overconfidence, overplacement and bargaining outcomes (disagreement and mismatch indices) across participants, before the intervention. Before receiving feedback about the joint production, participants exhibited an overconfidence bias as their confidence ($M = 79.79, SD = 8.31$) was significantly higher ($t(217) = 9.394$, $p < .001$) than their actual performance ($M = 74.24, SD = 3.11$). After receiving feedback, there was still a clear overplacement bias, as participants' revised confidence ($M = 79.90$, $SD = 7.43$) was significantly higher ($t(217) = 10.342$, $p < .001$) than their actual performance ($M = 74.24$, $SD = 3.11$). In the bargaining game, we found 51.21% of disagreement and a mismatch of 9.54 on average. As a comparison, for virtual agents who would maximize their expected earnings, best respond to the feedback about joint production and perfectly estimate their own production, disagreement and mismatch would still be present (even in the absence of overplacement, due to strategic uncertainty) but they would be significantly lower than in our data (disagreement: $t(217) = 14.685$, $p < .001$; mismatch: $t(217) = 9.764$, $p < .001$).

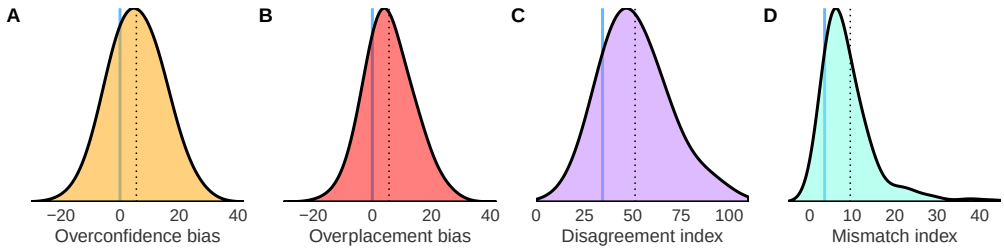


Figure 2: (A) Distribution of overconfidence for participants ($n = 218$). The line represents the absence of bias. (B) Same as A but for overplacement. (C) Distribution of disagreement index. The line represents the benchmark for unbiased agents making best-response claims. (D) Same as C) but for mismatch index. Dotted lines represent averages across participants.

We then investigated the relation between overplacement and bargaining outcomes. Under the assumption that claims should reflect revised confidence, we expected i) that disagreement would increase with overplacement and ii) that mismatch, would be minimal in the absence of overplacement. These two predictions were confirmed in panel regressions (Table 1), both across and within participants. Figure 3 illustrates these relations.

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

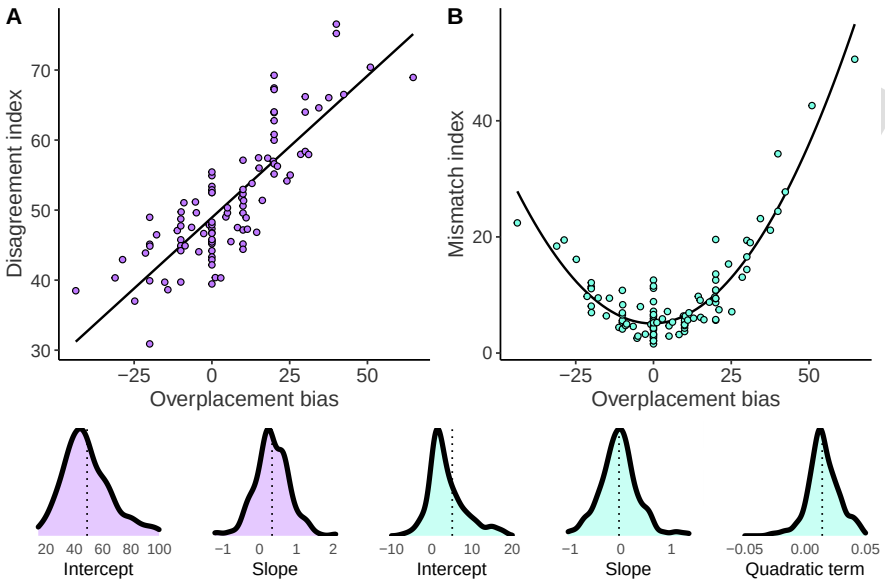


Figure 3: (A) Disagreement index regressed against overplacement bias. Each dot represents the average of observations over a percentile of overplacement. The black line represents a linear regression. Below the main plot are shown the distributions across participants of the regression coefficients (slope and intercept), with dotted lines representing the mean across participants. (B) Same as A but for mismatch index, and with a quadratic regression, with intercept, slope, and coefficient for the quadratic term.

	Disagreement		**Mismatch**	
	BE	FD	BE	FD
Intercept	46.97^{***}	-0.81	5.18^{***}	-0.035
	(1.34)	(0.54)	(1.23)	(0.301)
Overplacement	0.75^{***}	0.36^{***}	-0.058	0.054^{***}
	(0.14)	(0.02)	(0.093)	(0.015)
Overplacement^2	x	x	0.013^{***}	0.010^{***}
	x	x	(0.003)	(0.000)
Observation used	218	5450	218	5450
F Statistic	28.90^{***}	226.28^{***}	8.20^{***}	459.62^{***}
Note:	*p<0.05; **p<0.01;***p<0.001			
	BE = between estimator			
	FD = first difference estimator			

Table 1: Panel data regressions of disagreement index and mismatch index against overplacement bias

3.2 A generative model of revised confidence

To investigate more precisely the determinants of overplacement, we considered a model in which revised confidence is predicted from initial confidence and joint production, while taking into account biases in the estimation of others' performance and biases in belief updating.[10] By doing so, we could decompose overplacement into three specific components: one related to beliefs about own performance (i.e. confidence) which are affected by an overconfidence bias, one related to beliefs about the performance of the other player (thereafter other-confidence) which are affected by a superiority bias and one related to belief updating which is affected by biases in information processing (conservatism and asymmetry). A schematic representation of this model is shown in Figure 4. We present below evidence supporting the existence of the different components of the model before reporting our empirical estimation.

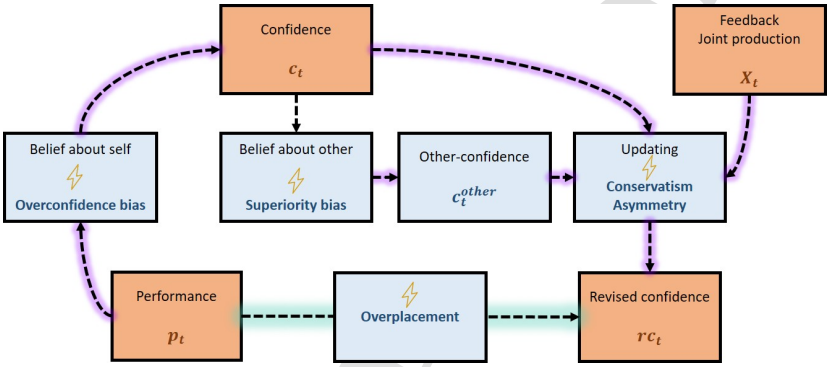


Figure 4: Schematic representation of the generative model of revised confidence. Orange boxes represent the empirically observed variables: performance, confidence, feedback and revised confidence. Blue boxes represent the internal variables of the model, namely the performance of the other player c_t^{other} as estimated by the individual, and the different cognitive biases of the individual. The green arrow illustrates our measure of overplacement. The purple arrows illustrate the decomposition of overplacement into three cognitive biases: overconfidence (overestimation of one's own performance), superiority (estimation of the performance of the other player lower than one's own), and updating biases (conservatism and asymmetry in belief updating after feedback).

Updating

For a Bayesian agent, the revised confidence in a given round t (noted rc_t^*) can be calculated from Bayes' rule once the joint production of the pair X_t is revealed, under the assumption that the number of correct answers (the production) was generated by a binomial distribution with success probability c_t for the participant and c_t^{other} for the other player (equation 1).

$$rc_t^* = \frac{\sum_{k=0}^{5} k \times \mathbb{P}(k|X_t, c_t, c_t^{\text{other}})}{5}$$
$$\text{with } \mathbb{P}(k|X_t, c_t, c_t^{\text{other}}) = \frac{\mathbb{P}(X_t \cap k|c_t, c_t^{\text{other}})}{\mathbb{P}(X_t|c_t, c_t^{\text{other}})} \tag{1}$$

In our model, belief updating could depart from the Bayesian benchmark: updating could be under-

[10] We thank an anonymous reviewer who suggested incorporating non-Bayesian updating into our model. Initially, we had planned to recover beliefs about others by comparing confidence and revised confidence which would have merged biases about others with updating biases.

sensitive or over-sensitive to feedback and this could depend on whether this feedback represented good or bad news compared to the expected joint production. Indeed, an important literature shows that in comparison to Bayesian agents, individuals often exhibit conservatism (Edwards, 1968; Fischhoff and Beyth-Marom, 1983; Huck and Weizsäcker, 2002) and asymmetry (Ertac, 2011; Eil and Rao, 2011; Charness and Dave, 2017; Coutts, 2019; Coutts et al., 2021; Drobner and Goerg, 2021; Möbius et al., 2022) in their updating.[11] As seen in equation 2, we captured these biases with two parameters, β which is the responsiveness to feedback following bad news (when Bayesian revised confidence is below the initial confidence) and γ which is the difference in responsiveness between good and bad news. For the Bayesian agent, β is equal to 1, and γ is equal to 0.

$$\begin{gathered} rc_t - c_t = \beta \times (rc_t^* - c_t) + \gamma \times \text{Good News} \times (rc_t^* - c_t) \\ \text{with } rc_t^* \text{ the ideal Bayesian revision and} \\ \text{Good News} = 1 \text{ when } rc_t^* > c_t \text{ and } 0 \text{ otherwise} \end{gathered} \tag{2}$$

Figure 5A illustrates belief updating biases in our data, under the hypothesis that other-confidence is constant during the experiment and equal to .75 (the actual performance of the other player). The figure presents participants' updating $rc_t - c_t$ against the Bayesian benchmark $rc_t^* - c_t$. Overall, observations form a flatter line than the Bayesian benchmark and particularly so in the case of bad news. In other words, even without estimating our full model (which involves two additional parameters described below, jointly estimated with β and γ) we could obtain evidence for conservative and asymmetric updating in our data.

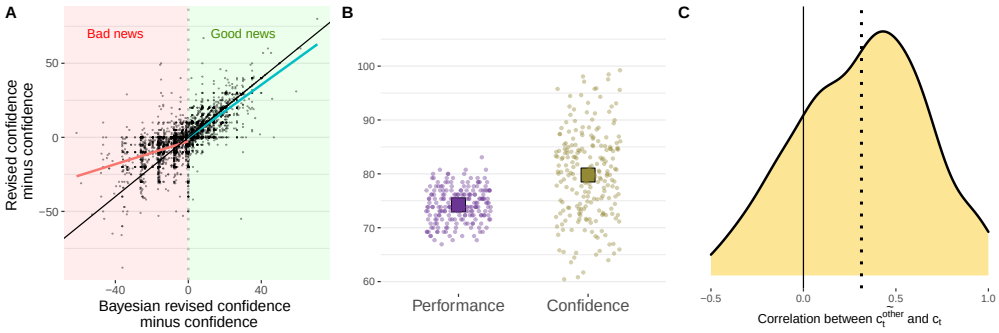


Figure 5: (A) Participants' updating (revised confidence minus confidence) against Bayesian updating (Bayesian revised confidence - confidence) for bad news (left side of the plot, in red) and good news (right side of the plot, in green) before the intervention. Each dot corresponds to a given participant in a given round. For this illustration, Bayesian revised confidence is defined under the assumption that $c_t^{other} = .75$. The red and green lines correspond to separate linear regression for bad and for good news. (B) Performance and confidence before the intervention. Each dot corresponds to a single participant. Squares represent means across participants. (C) Distribution across participants of correlation between $\tilde{c}_t^{other}$ and c_t. For this illustration, $\tilde{c}_t^{other}$ is defined under the assumption that the updating process is Bayesian, that is $\beta = 1$ and $\delta = 0$ (see main text for details). The dotted line represents the average correlation across participants.

[11] We note that the evidence for asymmetric updating is actually mixed in the literature. While most studies find oversensitivity to good news (Eil and Rao, 2011; Charness and Dave, 2017; Coutts et al., 2021; Drobner and Goerg, 2021; Möbius et al., 2022), other studies find the opposite asymmetry (Ertac, 2011; Coutts, 2019), or no asymmetry (Grossman and Owens, 2012; Buser et al., 2018; Barron, 2021). Barron (2021) recently concluded that differences in information structure, priors, domains and stake sizes between the different studies could not account for those mixed results. A detailed investigation of the mechanims by which deviations may (or may not) arise, however, is outside of the scope of the present paper.

Other-confidence

Belief updating requires individuals to entertain a belief c_t^{other} about the performance of the other player. We modelled this other-confidence by means of two parameters. One parameter b captured the superiority bias, which is the extent to which participants estimated that the other player's performance was lower than their own. The other parameter δ captured the degree to which participants relied on their confidence in each round to form their other-confidence in that round. Putting together these two elements, participant's other-confidence in round t could be written as

$$c_t^{\text{other}} = (1-\delta) \times \overline{c} + \delta \times c_t - b \tag{3}$$

with $\delta \in [0,1]$ and $\overline{c}$ is participant's average confidence during the experiment.

Notice that averaging across rounds equation 3, b corresponds to how much participant's average estimation of the performance of the other player ($\overline{c^{\text{other}}}$) is lower than his own ($\overline{c}$). The parameter δ by contrast does not relate to overplacement. We introduced δ to account for the possibility that participants' other-confidence may not have been constant during the experiment and instead depended on their own confidence, as appeared to be the case in our data. For instance, when setting β and γ in equation 1 to fixed values, we could compute the other-confidence $\tilde{c_t^{\text{other}}}$ that would be consistent with the observables X_t, c_t and rc_t, and we found that $\tilde{c_t^{\text{other}}}$ was positively correlated with c_t across trials. This gives some indication that participants use their belief about their own performance to form their estimates about the performance of the other. Figure 5C illustrates this phenomenon for $\beta = 1$ and $\delta = 0$ (i.e. Bayesian benchmark for belief updating). In this case, the correlation between $\tilde{c_t^{\text{other}}}$ and c_t, estimated for each participant, was significantly positive across participants ($\overline{Cor} = .32$, $t(217) = 13.764$, $p < .001$).

Estimation of the full model

In the full model, revised confidence was thus determined for each participant on each round by participant's confidence c_t, the joint production X_t, and 4 free parameters characterizing the formation of beliefs about the performance of the other (b, δ) and the updating of beliefs following feedback (β, γ). Assuming normally distributed noise on the predicted revised confidence, we estimated our model using maximum likelihood, separately for each participant, using the 26 observations available for each participant before the intervention.

In Figure 6, we show the distribution of these 4 parameters across participants. In addition, we plot the average revised confidence predicted by our fitted model against the actual average revised confidence. On average, b was slightly higher than zero. Indeed, participants' estimated other-confidence ($M = 73.77, SD = 17.98$) was on average lower than their confidence ($M = 79.79, SD = 8.31$). Furthermore, the extent to which participants' confidence influenced their other-confidence was highly heterogeneous: δ was equal to 1 for 34% of participants, to 0 for 13% of them and fell between these two values for 53% of participants. Finally, participants processed feedback about the joint production in a biased manner, exhibiting conservatism ($\beta < 1$) and asymmetry ($\gamma > 0$) in their belief updating.

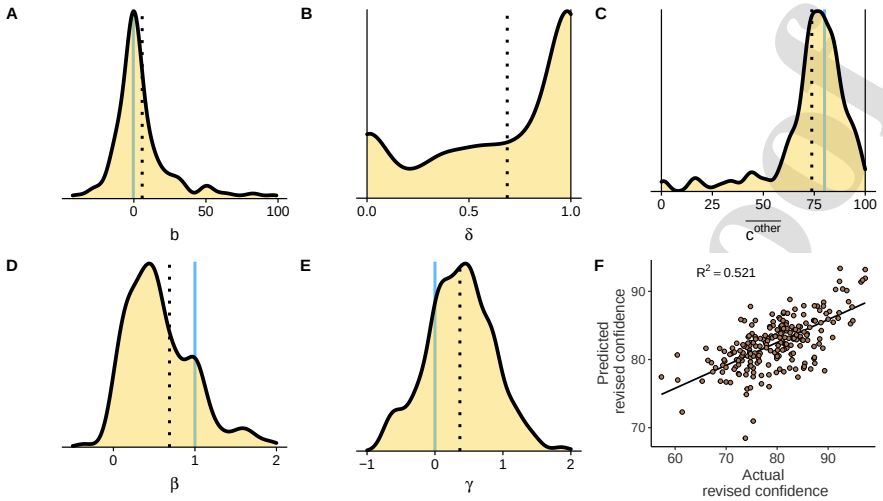


Figure 6: (A-E) Distribution across participants of the estimated parameters: b quantifies the superiority bias, δ the extent to which participants' confidence influences their other-confidence on a round-by-round basis, $\overline{c^{other}}$ is the average other-confidence across rounds as defined by parameters b and δ in equation 3 and β and γ characterize conservatism and asymmetry in belief updating. Dotted lines represent averages across participants. When relevant, blue solid lines represent the absence of bias as defined in the main text. (F) Actual and predicted value of the average revised confidence for each participant in the full model ($N = 218$).

Decomposition of overplacement

To evaluate the contribution of the different biases to overplacement, we examined how the revised confidence predicted by the model would change if a particular bias was to be removed. Eliminating overconfidence bias was done by setting confidence in each round to the actual performance, that is setting for each round t $c_t = p_t$. Eliminating the superiority bias was done by setting b to the true difference between the participant's performance and the performance of the other (such that on average, the difference in beliefs would be equal to the actual difference in performance between the participant and the other player). Eliminating biased belief updating corresponded to setting $\beta = 1$ and $\gamma = 0$. Starting from the full model, we eliminated one bias at a time such that after three steps the model would be fully unbiased. There were 6 possible orders for doing so, as illustrated on Figure 7. When removing a bias, we then evaluated how revised confidence should change in the model. Finally, for each bias, we considered the average effect of removing this bias (across the 6 different paths) to quantify its contribution to overplacement.

For the full model, predicted revised confidence was 79.71% on average (close to the 79.90% in the actual data). For the fully unbiased model, it dropped to 74.88%. On average, eliminating overconfidence in our model reduced predicted revised confidence by 1.40 points, eliminating superiority bias reduced it by 1.68 points, and eliminating biased updating by 1.75 points. In other words, each bias contributed about equally (by 29%, 35% and 36% respectively) to overplacement.

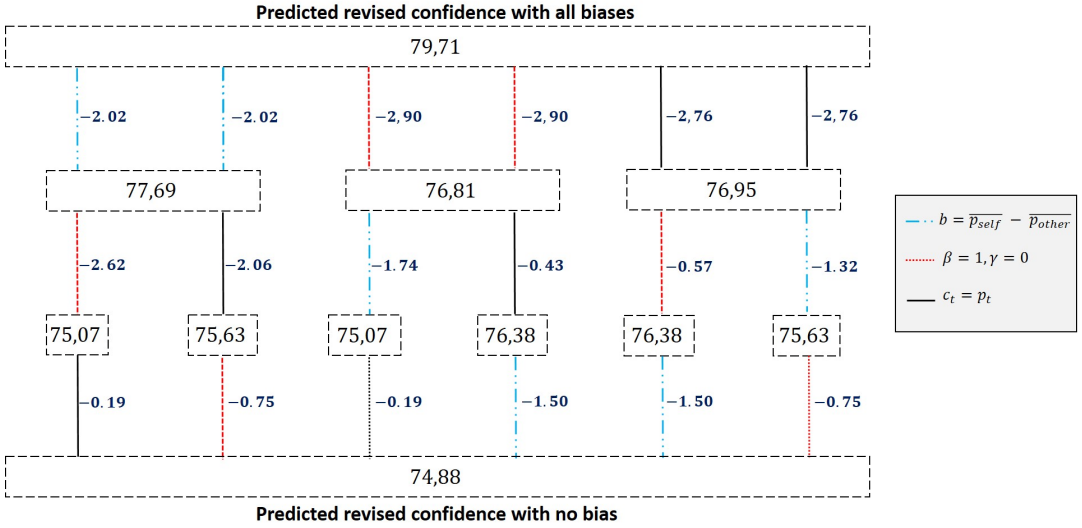


Figure 7: Representation of each bias' contribution to overplacement. At the top is indicated the predicted revised confidence in our full model (with the 3 biases) and at the bottom the predicted revised confidence in the fully unbiased case (when all biases are removed from the model). Each line corresponds to the elimination of one type of bias (blue: superiority bias, red: belief updating bias, black: overconfidence bias), and indicates the change in the predicted revised confidence when doing so. See main text for details.

3.3 The interventions

We now turn to the analysis of the interventions conducted at the middle of the experiment. Recall that there were 5 experimental conditions: a baseline in which we asked participants about their average claim (T_0) and 4 treatments in which we focused 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 slightly overestimated their past claims as they reported asking for 54.75% of the joint production while they were only asking for 53.03% on average ($\overline{Error}_{Q_0} = 1.71,\ t(47) = 1.84,\ p = 0.072$). 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 (in T_1 and T_4), participants strongly overestimated it, declaring that they had produced 60.68% of the joint production while their actual 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, mistakes made on this question were correlated with overplacement bias ($Cor = 0.51, p < 0.001$). In T_2 and T_4, participants largely underestimated the number of times their claim was above their contribution, with an average estimate of 7.84 times, compared to the true value of 13.55 times out of 26 ($\overline{Error}_{Q_2} = 5.71,\ t(74) = 8.80,\ p < 0.001$). In T_3 and T_4, participants also consistently underestimated the number of times their confidence was above their individual performance, with an estimate of 10.08 times 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 participants being not only overconfident but also naive with respect to

this bias. This was confirmed by the correlation between the mistakes made on this question and participant's overconfidence ($Cor = -0.50, p < 0.001$). Finally, examining answers to all questions for participants in T_4, we found that all estimation errors were significantly correlated in the expected direction.

Intervention effect

The effect of the intervention on bargaining outcomes and overplacement bias are presented in Figure 8 and in Table 2 and 3. Disagreement index significantly fell by 5.57 points after the intervention in the baseline condition, and even more in the four treatments. However, only T_4 significantly differed from baseline ($p = .002$) with an additional decrease of 8.16 points. Results were similar for the mismatch index. The distance between claims and contribution was reduced after the intervention in all treatments, but only for T_4 was the intervention effect significant in comparison to baseline ($p = .021$), with a reduction of mismatch by 3.61 points. In sum, intervention T_4 successfully reduced disagreements and brought claims closer to actual contributions, thereby improving both bargaining outcomes. Overplacement bias showed a similar pattern, and was reduced in comparison to baseline only in T_4, with a drop of 2.90 points ($p = .027$). As a robustness check, we confirmed that this intervention effect in T_4 (compared to baseline) was directly observable on the raw number disagreements ($p = .016$) as well as in claims ($p = .004$).

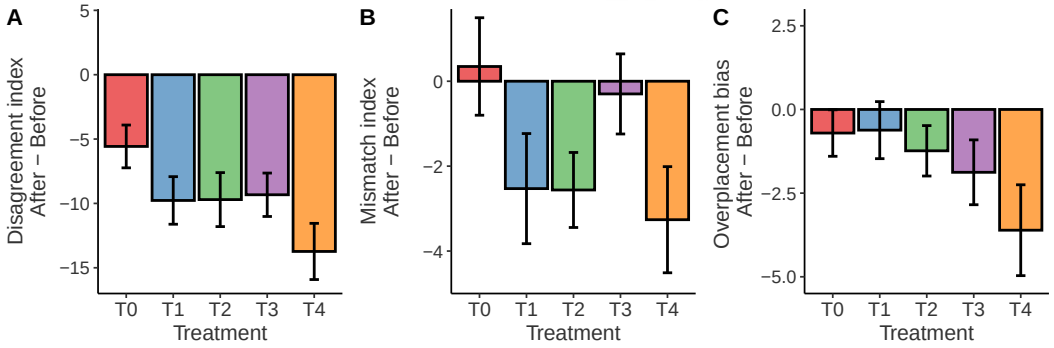


Figure 8: Change in disagreement index (panel A), mismatch index (panel B) and overplacement bias (panel C) after (vs. before) the intervention, for each treatment. Positive values indicate an increase in the variable of interest after the intervention. Error bars represent mean and s.e.m. across participants.

For completeness, we further checked that our results were not driven by pre-intervention differences between treatments: we found no pre-intervention differences between baseline and T_4 in beliefs (overplacement, overconfidence) and in bargaining behavior (claims, claim sensitivity, disagreement and mismatch indices). We also found that the change in participants' claim sensitivity following the intervention was not significantly different in T_4 compared to baseline, suggesting that our intervention did not work solely because individuals changed their claim strategy with respect to their revised confidence.

	Δ Disagreement		Δ Disagreement (raw)		Δ Claim		Δ Mismatch	
	Estimate	Std	Estimate	Std	Estimate	Std	Estimate	Std
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<0.05; **p<0.01;***p<0.001					

Table 2: Regression of disagreement index, average disagreement, average claim and mismatch index's change (After − Before) by treatment (reference=Baseline)

	Δ OP bias		Δ Sensitivity	
	Estimate	Std	Estimate	Std
T_1	0.09	(1.31)	−0.28	(0.18)
T_2	−0.53	(1.28)	−0.19	(0.18)
T_3	−1.17	(1.38)	−0.01	(0.19)
T_4	−2.90*	(1.54)	−0.21	(0.19)
Constant	0.71	(0.90)	0.13	(0.12)
Observations	218		218	
F Statistic	1.71		0.92	
Note:	*p<0.05; **p<0.01;***p<0.001			

Table 3: Regression of overplacement bias and sensitivity to perceived contribution's change (After − Before) by treatment (reference=Baseline)

Decomposition of the intervention effect on overplacement

To evaluate how the change in overconfidence bias, superiority bias and updating biases may have contributed to the intervention effect on overplacement, we extended the methodology used above in our decomposition of overplacement. Specifically, we first estimated our model for revised confidence as before but now on all observations, allowing the parameter b capturing the superiority bias, and the parameters β and γ capturing updating biases to change after the intervention. To evaluate how much each bias contributed to the intervention effect, in our model we then eliminated the intervention effect on each bias, one at a time, as if the bias was unaffected by the intervention. For updating biases, we simply set β and δ after the intervention to their values estimated before the intervention. To eliminate the change in overconfidence bias, we translated the reported confidence after the intervention such that participants would exhibit the same overconfidence

bias as the one measured before the intervention. We proceeded in a similar way to eliminate the change in superiority bias, by setting b after the intervention to its value estimated before the intervention, with a correction for the potential change in the difference in actual performances between the participant and the other player.[12] From this, we could evaluate how the intervention effect on revised confidence would be affected if each bias was kept constant across the experiment and could not participate in this effect. As before, there were 6 possible different paths to consider (i.e. there were 6 possible orders in which biases could be removed) and we calculated the average effect of removing a given bias across these 6 paths. The results of this decomposition for each treatment are summarized in Figure 9A.

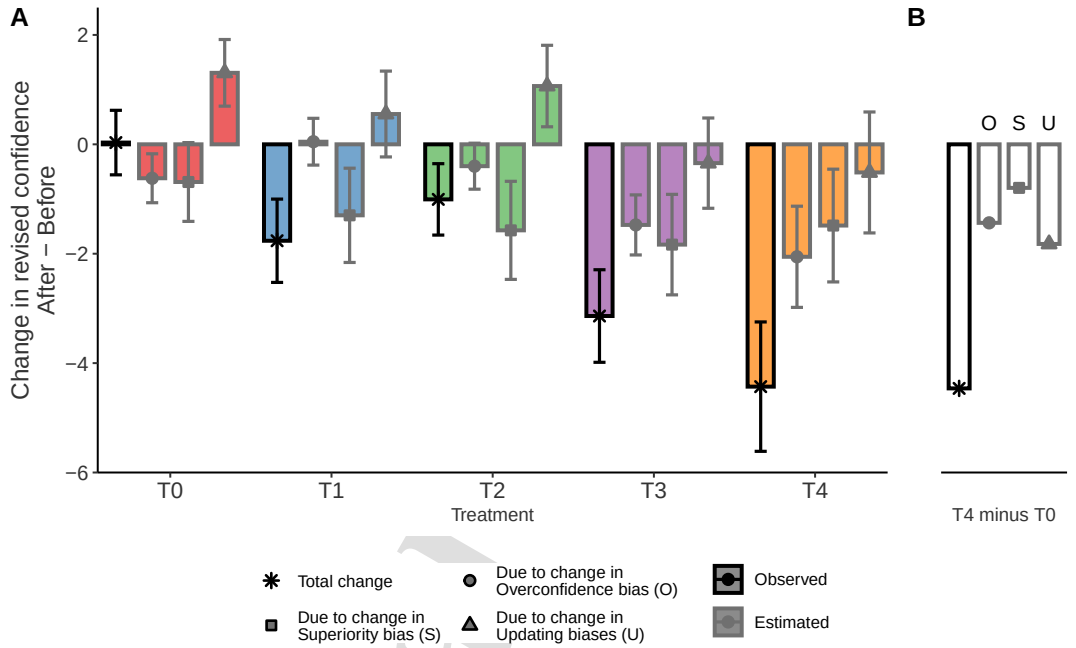


Figure 9: (A) Decomposition of the change in revised confidence after (vs. before) the intervention by each bias, for each treatment. Error bars represent mean and s.e.m. across participants. (B) Decomposition of the treatment effect, that is, the change in T_4 relative to the change in T_0. Asterisks represent the change in revised confidence observed in our data. Circles, squares and triangles represent the estimated change in revised confidence due to a change in overconfidence bias, superiority bias and updating biases, respectively (see main text for details).

In particular, in T_4, where the intervention significantly decreased overplacement and disagreements compared to baseline, predicted revised confidence dropped by 4.22 after the intervention (4.43 in the actual data). We estimated that 51% (2.06 points) of this effect was due to a decrease in overconfidence bias, 37% (1.49 points) to a decrease in superiority bias and 13% (0.52 points) to a decrease in updating biases.

Finally, we also carried out this procedure to the treatment effect, that is, the comparison of the intervention effect between T_4 and baseline treatments. The predicted treatment effect on revised confidence was

[12] That way, if a participant became relatively better (worse) than the other player after the intervention, we accounted for the fact that ideally the superiority bias should have increased (decreased) accordingly.

4.2 points in the full model (4.46 points in the data), and 0.15 points in the model where all changes in biases were eliminated. As can be seen in Figure 9B, eliminating overconfidence reduced the predicted treatment effect by 1.42 points (35%), eliminating superiority bias reduced it by 0.80 points (20%), and eliminating biased updating reduced it by 1.82 points (45%). Notice that the relative contributions of overconfidence and superiority biases were higher when we decomposed the intervention effect in T_4 alone than when we decomposed the treatment effect (T_4 vs. baseline): this was expected since in the baseline overconfidence and superiority biases slightly decrease while updating biases increase (see Figure 9A).

3.4 Gender differences

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

Before the intervention, we found no difference between men and women regarding the outcomes of the bargaining game (gender effect on disagreement index: $F(216) = 0.966, p = 0.327$; mismatch index $F(216) = 0.361, p = 0.549$), regarding the elicited performance beliefs (overplacement bias: $F(216) = 0.197, p = 0.583$; overconfidence bias: $F(216) = 0.113, p = 0.737$) and regarding biases estimated in our model (superiority bias: $F(216) = 0.187, p = 0.665$; conservatism: $F(216) = 2.607, p = 0.108$; asymmetry $F(216) = 1.955, p = 0.163$). This similarity in cognitive evaluation between men and women is not surprising. Indeed, although men are often more overconfident than women (Lundeberg et al., 1994; Hügelschäfer and Achtziger, 2014), gender differences in confidence were found to be highly task dependent (Deaux and Farris, 1977; Lundeberg et al., 1994; Bordalo et al., 2019). For instance, Bordalo et al. (2019) showed that on a quiz where men and women were expected to perform similarly, they were also similar in how they evaluated themselves or others. On the other hand, the fact that men and women reacted similarly to the joint production feedback contrasts with previous findings that women tend to revise more pessimistically than men when receiving feedback about relative performance (Berlin and Dargnies, 2016).[13]

We found one difference between men and women regarding the sensitivity to perceived contribution: men were more sensitive to their perceived contribution than women ($F(216) = 6.35, p = 0.012$). This result is somewhat consistent with D'Exelle et al. (2017) who showed that in a Nash demand game the influence of beliefs about what the other will claim is stronger for men than women and with Durante et al. (2014) who showed that the difference in individuals' demand for redistribution when initial earnings are arbitrary vs earned is much more pronounced for men than women. One other possible interpretation is that our sensitivity measure might have captured risk or fairness principles (risk averse participants' should have made claims that are more disconnected from their confidence and favoured safer claims at 50%) on which men and women may differ (e.g. women being more risk averse than men (Eckel and Grossman, 2008)).

We also considered gender differences with respect to the effect of interventions on bargaining outcomes (see Table 4). Interestingly, it appeared that the effect of the intervention was mostly driven by women, as for men we did not find any treatment effect on disagreement or mismatch indices. This pattern suggests that women reacted to the interventions while men did not. When focusing on the comparison between baseline and T_4, we found that in T_4 the disagreement index improved significantly more for women than for men ($p = .002$), with a large difference of 16.38 points between genders. The mismatch index also improved more for women than for men in T_4 compared to baseline, but the difference between genders was not significant in this case. The difference between men and women was also present in overplacement: when comparing T_4 to

[13] This may be due to the fact that our perceptual task was more gender neutral than the mathematics task in Berlin and Dargnies (2016). However, although we expect the gender gap on prior beliefs to depend on the type of task (Deaux and Farris, 1977; Lundeberg et al., 1994; Bordalo et al., 2019), it is not clear why it would also impact the gender gap in information processing.

baseline, overplacement decreased after the intervention but this decrease was significantly more pronounced for women than men (a difference of 5.50 points, $p = .035$).

	Δ Disagreement	Δ Mismatch	Δ OP bias	Δ Sensitivity
T_1	−0.83	−2.29	−0.07	−0.14
	(3.91)	(2.31)	(1.94)	(0.26)
T_2	−0.67	−3.36	0.19	0.04
	(3.69)	(2.11)	(1.77)	(0.24)
T_3	1.70	−0.51	−1.01	−0.08
	(3.75)	(2.21)	(1.86)	(0.25)
T_4	−0.39	−1.54	−0.37	−0.18
	(3.62)	(2.14)	(1.79)	(0.27)
Women	7.10	−0.85	0.66	0.25
	(3.62)	(2.14)	(1.79)	(0.25)
Women × T_1	−6.78	−0.89	0.18	−0.29
	(5.31)	(3.13)	(2.63)	(0.36)
Women × T_2	−6.91	0.90	−1.50	−0.48
	(5.16)	(3.04)	(2.55	(0.35)
Women × T_3	−11.67*	−0.51	−0.27	0.19
	(5.58)	(3.29)	(2.76)	(0.38)
Women × T_4	−16.38**	−6.65	−5.50*	−0.08
	(5.25)	(3.10)	(2.60)	(0.38)
Constant	−9.12***	0.77	−1.04	−0.00
	(2.56)	(1.51)	(1.27)	(2.17)
Observations	218	218	218	218
F Statistic	2.30*	1.68	1.57	0.92
Note:		*p<0.05; **p<0.01;***p<0.001		

Table 4: Regression of disagreement index, mismatch index, overplacement bias and sensitivity to perceived contribution's change (After − Before) by treatment (reference=baseline) interacted with gender

4 Discussion

In this paper, we show that overplacement bias prevents settlement when individuals bargain over a joint production, and we decompose this overplacement into several cognitive biases. In practice, by eliciting participants confidence in their performance both before and after they receive feedback about the joint production, and repeating the bargaining game a large number of times, we decompose overplacement into three different biases: overestimation of one's own performance (overconfidence bias), estimation of others' performance lower than one's own (superiority bias), and non-Bayesian information processing. We show that these three biases are present in our data, and contribute in roughly the same proportions to the overplacement bias.

Our paper thus contributes to the empirical literature that identifies cognitive biases at the origin of bargaining failures when individuals have to share a jointly produced surplus under uncertainty (Karagözoğlu

and Riedl, 2015; Santos Pinto and Colzani, 2021; Soldà et al., 2021). Among the previously cited studies, only Soldà et al. (2021) measured beliefs twice as we do, before and after participants receive a noisy signal about their relative performance. They found no biases for prior beliefs (consistent with participants knowing that they were paired with equally performing players), but overplacement in posterior beliefs, due to biased updating.[14] Our results on biased updating are in line with these findings, but in our case updating was based on exact information about the joint production, not on noisy feedback regarding relative performance. Furthermore, our study goes beyond the updating issue, by examining the contribution of other biases as well.

To the best of our knowledge, our approach (i.e. evaluating the contribution of different cognitive biases to overplacement) has never been used in the past, and one could argue that our generative model of revised confidence deserves further scrutiny. In this regard, note that our model was parsimonious in terms of hypotheses. We used a classic model of belief updating to capture non-Bayesian information processing and for other-confidence our specification has the main advantage of directly incorporating the superiority bias with a single parameter. We agree that having other-confidence vary across rounds depending on confidence could be criticized as unnecessary a priori, but we provided evidence to support this specification. An interesting avenue for further research would be to measure this belief regarding the performance of the other player directly. This should allow assessing more precisely whether participants incorporate their own confidence in their belief about the other player.

The other novelty of our study was to test interventions aimed at decreasing cognitive biases and improving bargaining outcomes. In practice, we tested 4 interventions and showed that the most comprehensive one successfully reduced disagreements, mismatch and overplacement. In this intervention, participants received information about their true contribution to the joint production, about whether they claimed more than what they contributed and about their overconfidence, and after this intervention overplacement decreased significantly, compared to the baseline treatment. This treatment effect was in fact driven not only by participants in the treatment group whose overconfidence and superiority biases decreased, but also by participants in the baseline group who became increasingly biased in their updating (more conservative and asymmetric), which also raises an interesting question about the evolution of belief updating biases in competitive environments. Importantly, this decrease in overplacement coincided with a decrease in disagreement, so our intervention has produced the benefits that were expected. In addition, these benefits were not obtained at the cost of a greater mismatch between claims and contributions. In this regard, we note that the intervention in which participants only received feedback about their overconfidence (T_3) seems to have reduced cognitive biases almost as much as our most successful intervention (T_4), but with no reduction in disagreements and mismatch indices.

Our work thus points at a possible way to improve bargaining between individuals from a social planner perspective. We are of course aware that there are gaps between the present experimental procedure and what can be done outside the laboratory. For instance, here participants received feedback about their true production, which is not always feasible in real bargaining situations. More realistically, one could provide the two parties information about their overconfidence in tasks for which their true performance is measurable. Assuming that overconfidence is domain-general (West and Stanovich, 1997; Kelemen et al., 2000; Ais et al., 2015), de-biasing them for such tasks may correct their bias also in the bargaining situation.

Finally, one important but unexpected aspect of our results is the difference between women and men in our study. Specifically, we found that the intervention had a stronger impact on women than men. Although this effect was 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.

[14] Soldà et al. (2021) also measured absolute overestimation biases, but they did not analyse them in detail.

Nonetheless, the fact that men and women may react differently to de-biasing 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 strategic aspect of the game played a role, as suggested in other studies (Niederle and Vesterlund, 2007; Dreber and Johannesson, 2008; Malik et al., 2021). 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). In any case, understanding this gender difference, and designing interventions that are effective for both men and women, constitute an interesting challenge for future research.

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.

Barron, K. (2021). Belief updating: does the 'good-news, bad-news' asymmetry extend to purely financial domains? *Experimental Economics*, 24(2):31–58.

Becker, G., DeGroot, M. H., and Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral science*, 9(3):226–232.

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.

Buser, T., Gerhards, L., and van der Weele, J. (2018). Responsiveness to feedback as a personal trait. *Journal of Risk and Uncertainty*, 56(2):165–192.

Camerer, C. and Lovallo, D. (1999). Overconfidence and excess entry: An experimental approach. *American Economic Review*, 89(1):306–318.

Cappelen, A. W., Hole, A. D., Sørensen, E. O., and Tungodden, B. (2007). The pluralism of fairness ideals: An experimental approach. *American Economic Review*, 97(3):818–827.

Cappelen, A. W., Sørensen, E. O., and Tungodden, B. (2010). Responsibility for what? Fairness and individual responsibility. *European Economic Review*, 54(3):429–441.

Charness, G. and Dave, C. (2017). Confirmation bias with motivated beliefs. *Games and Economic Behavior*, 104:1–23.

Cherry, T. L., Frykblom, P., and Shogren, J. F. (2002). Hardnose the dictator. *American Economic Review*, 92(4):1218–1221.

Coutts, A. (2019). Good news and bad news are still news: experimental evidence on belief updating. *Experimental Economics*, 22(2):369–395.

Coutts, A., Gerhards, L., and Zahra, M. (2021). What to blame? Self-serving attribution bias withmulti-dimensional uncertainty. *Working Paper*.

Deaux, K. and Farris, E. (1977). Attributing causes for one's own performance: The effects of sex, norms, and outcome. *Journal of Research in Personality*, 11(1):59–72.

Devaine, M., Hollard, G., and Daunizeau, J. (2014). The social bayesian brain: Does mentalizing make a difference when we learn? *PLOS Computational Biology*, 10(12):1–14.

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

Dreber, A. and Johannesson, M. (2008). Gender differences in deception. *Economics Letters*, 99(1):197–199.

Drobner, C. and Goerg, S. (2021). Motivated belief updating and rationalization of information. *Working Paper, Munich Papers in Political Economy*.

Durante, R., Putterman, L., and van der Weele, J. (2014). Preferences for redistribution and perception of fairness: an experimental study. *Journal of the European Economic Association*, 12(4):1059–1086.

Eckel, C. C. and Grossman, P. J. (2008). Men, women and risk aversion: Experimental evidence. *Handbook of experimental economics results*, 1:1061–1073.

Edwards, W. (1968). Conservatism in human information processing. In Kleinmuntz, B. and Symposium on Cognition. Annual, editors, *Formal representation of human judgment*. Wiley, New York, NY; London.

Eil, D. and Rao, J. M. (2011). The good news-bad news effect: Asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, 3(2):114–38.

Ertac, S. (2011). Does self-relevance affect information processing? Experimental evidence on the response to performance and non-performance feedback. *Journal of Economic Behavior & Organization*, 80(3):532–545.

Fischbacher, U., Kairies-Schwarz, N., and Stefani, U. (2017). Non-additivity and the salience of marginal productivities: Experimental evidence on distributive fairness. *Economica*, 84(336):587–610.

Fischhoff, B. and Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review*, 90(3):239–260.

Gantner, A., Guth, W., and Konigstein, M. (2001). Equitable choices in bargaining games with joint production. *Journal of Economic Behavior & Organization*, 46(2):209–225.

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

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

Grieco, D. and Hogarth, R. M. (2009). Overconfidence in absolute and relative performance: The regression hypothesis and bayesian updating. *Journal of Economic Psychology*, 30(5):756–771.

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

Hügelschäfer, S. and Achtziger, A. (2014). On confident men and rational women: It's all on your mind(set). *Journal of Economic Psychology*, 41:31–44.

Huck, S. and Weizsäcker, G. (2002). Do players correctly estimate what others do? Evidence of conservatism in beliefs. *Journal of Economic Behavior & Organization*, 47(1):71–85.

Johnson, D. D., McDermott, R., and Barrett, E. S. (2006). Overconfidence in wargames: experimental evidence on expectations, aggression, gender and testosterone. *Proceedings. Biological sciences*, 273(1600):2513–2520.

Karagözoğlu, E. and Riedl, A. (2015). Performance information, production uncertainty, and subjective entitlements in bargaining. *Management Science*, 61(11):2611–2626.

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

Konow, J., Saijo, T., Akai, K., et al. (2009). Morals and mores: Experimental evidence on equity and equality. *Levine's Working Paper Archive*.

Kruger, J. and Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6):1121–1134.

Lundeberg, M. A., Fox, P. W., and Punćcohaŕ, J. (1994). Highly confident, but wrong: Gender differences and similarities in confidence judgments. *Journal of Educational Psychology*, 86:114–121.

Malik, S., Mihm, B., Mihm, M., and Timme, F. (2021). Gender differences in bargaining with asymmetric information. *Journal of Economic Psychology*, 86:102415.

March, C. (2021). Strategic interactions between humans and artificial intelligence: Lessons from experiments with computer players. *Journal of Economic Psychology*, 87:102426.

Möbius, M. M., Niederle, M., Niehaus, P., and Rosenblat, T. S. (2022). Managing self-confidence: Theory and experimental evidence. *Management Science*.

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.

Santos Pinto, L. and Colzani, P. (2021). Does overconfidence lead to bargaining failures? *Working Paper, Université de Lausanne, Cahiers de Recherches Economiques du Département d'économie*.

Soldà, A., Ke, C., von Hippel, W., and Page, L. (2021). Absolute versus relative success: Why overconfidence creates an inefficient equilibrium. *Psychological Science*, 32(10):1662–1674.

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.