What to blame? Self-serving attribution bias with multi-dimensional uncertainty*†

Alexander Coutts   Leonie Gerhards   Zahra Murad

October 2020

Abstract

People often receive feedback that depends on factors beyond their ability, yet little is known about how this alters the scope for self-serving biases. In a theory-guided experiment, individuals receive a noisy signal about *their ability*, which comes bundled with another source of uncertainty – a *teammate's ability.* In this environment individuals can attribute the feedback across these two dimensions, updating in a self-serving fashion, leveraging the additional flexibility from multi-dimensional uncertainty. In the experiment, rather than blaming their teammate, they process information about them in a positively biased way. This reduces costs associated with over-attribution towards own performance, but later impedes learning by decreasing willingness to change teammates. These results suggest that individuals distort their perceptions of the environment in order to arrive at self-serving beliefs.

***Coutts**: Nova School of Business and Economics, Universidade NOVA de Lisboa, Campus de Carcavelos, Rua da Holanda 1, 2775-405 Carcavelos, Portugal (email: alexander.coutts@novasbe.pt); **Gerhards**: Department of Economics, Universität Hamburg, Von-Melle-Park 5, 20146 Hamburg, Germany (e-mail: leonie.gerhards@uni-hamburg.de); **Murad**: Economics and Finance, The University of Portsmouth, Portsmouth, PO1 2UP, United Kingdom (email: zahra.murad@port.ac.uk).

†We are very grateful for useful comments from Kai Barron, Thomas Buser, Tingting Ding, Boon Han Koh, Yves Le Yaouanq, Robin Lumsdaine, Cesar Mantilla, Luis Santos Pinto, Giorgia Romagnoli, Adam Sanjurjo, Marcello Sartarelli, Peter Schwardmann, Sebastian Schweighofer-Kodritsch, Séverine Toussaert, Joël van der Weele, and Georg Weizsäcker, as well as helpful comments from seminar and conference participants at University of Alicante, University of Amsterdam, Bayesian Crowd Conference, briq Workshop on Beliefs, CEA Banff, ECBE San Diego, ESA Berlin, HEC Lausanne, IMEBESS Utrecht, Lisbon Game Theory Meetings, M-BEES, NASMES Seattle, NYU CESS, NYU Shanghai, University of Portsmouth, RWTH Aachen, Schulich School of Business, SHUFE, THEEM, TRIBE Copenhagen, and WZB. We gratefully acknowledge financial support from the Hamburgische Wissenschaftliche Stiftung and the University of Hamburg.

1 Introduction

Researchers have amassed a wealth of evidence suggesting that people hold self-serving beliefs, regarding personal traits such as ability, beauty, or health (Benoît et al., 2015; Eil and Rao, 2011; Oster et al., 2013). The motives for holding these overly-rosy beliefs are typically thought to relate to their hedonic, signalling, or motivational value (Bénabou and Tirole, 2002).[1] Yet the production and persistence of such inflated beliefs is not well understood, and is especially puzzling considering that individuals often receive informative feedback about these traits, suggesting some degree of reality denial in processing this information.[2]

Surely there must be consequences or limits to denying incoming information which contradicts desired beliefs.[3] Indeed, the literature has studied two mechanisms thought to be responsible for mediating and limiting the production of excessively biased self-serving beliefs. The first is that poorly calibrated individuals would make materially costly mistakes (Brunnermeier and Parker, 2005), for example overconfident traders or entrepreneurs would lose money (Barber and Odean, 2002; Camerer and Lovallo, 1999). In line with this, Zimmermann (2019) finds that material consequences affect the scope for biased recall of negative feedback about performance on an IQ test. The second is that it can be cognitively costly to internally justify these deceptions, i.e. there are cognitive limits on how much individuals can deceive themselves (Bénabou and Tirole, 2002; Bracha and Brown, 2012). For example, Engelmann et al. (2019) find that self-deception about the likelihood of an impending electric shock decreases when the evidence for this shock is presented in a less ambiguous way.

In this paper we find and discuss a new mechanism. In many facets of life, individuals receive feedback or information that comes bundled with other fundamental (non-transitory) sources of uncertainty, such as specific features of the decision environment or a coworker's ability. We highlight the critical role played by these fundamentals, by analyzing how they affect the scope of the two cost mechanisms, and subsequently the production of self-serving beliefs. We distinguish these multi-dimensional settings from the one-dimensional settings that the empirical economics literature has mainly focused on so far, reviewed by Benjamin (2019). In these typical studies on motivated information processing, subjects receive noisy (transitory) feedback about an ego-relevant trait. This literature has been invaluable in gathering evidence on whether self-serving biases alter how we process information. However we argue that the

[1] Specifically, benefits may arise from: (i) direct utility from holding overconfident beliefs for example arising from self-esteem or ego-protection (Möbius et al., 2014; Brunnermeier and Parker, 2005), (ii) benefits to personal motivation or self-signalling (Bénabou and Tirole, 2002, 2009, 2011), or (iii) strategic signalling motives and persuasion of others (Burks et al., 2013; Schwardmann and van der Weele, 2019). These three explanations have long been a part of the core motivation for attribution theory of social psychology, corresponding to (i) self-enhancement/protection (ii) belief in effective control, and (iii) positive presentation of self to others; see Kelley and Michela (1980) and Tetlock and Levi (1982).

[2] While we focus on biases in information processing, there is evidence for other self-serving strategies such as avoiding negative information (e.g. for health (Oster et al., 2013); see Golman et al. (2017) for a broader review), or biased recall (Zimmermann (2019)).

[3] Bénabou (2015) refers to this as *reality denial*, which is our focus in this paper. Beyond this, he discusses two further supply side categories: *willfull blindness* and *self-signalling*. The former refers to strategic selection or avoidance of information sources, while the latter refers to the possibility of "manufacturing signals" for example through inference about past choices (Prelec and Bodner, 2003; Mijovic-Prelec and Prelec, 2010).

one-dimensional case is likely to limit the scope for self-serving distortions, since it lacks the added flexibility to arrive at desired beliefs that multiple dimensions offer. These constraints on distorting beliefs in the one-dimensional case could in part explain why the literature which studies ego-relevant belief updating has not found robust evidence for self-serving information processing.[4] Specifically, we study how different dimensions of uncertainty interact with the two aforementioned cost mechanisms to offer additional flexibility in enabling self-serving belief distortions.

The following example illustrates our theoretical and experimental setting. An individual is placed in a team of two, and must choose an allocation between the two teammates, where the optimal allocation weight is proportional to their respective abilities. She receives noisy joint feedback about her own and the teammate's ability. Assuming she would intrinsically benefit from holding overconfident beliefs, the material costs of self-serving information processing about her ability are that she will bias the allocation towards her own performance, and subsequently end up with a worse outcome.

In accordance with the two cost mechanisms, a typical approach would be to model the optimal level of biased information processing about own ability as trading off the presumed benefits from overconfidence against the material costs, accounting for any cognitive costs of engaging in distorted updating. However, our main insight is that the individual has an additional tool to arrive at self-serving beliefs, which is to bias how she perceives information about her teammate. As we will show, in our context, and more generally, this bias could go in either direction.

On the one hand, following theories of self-serving attribution bias (Hastorf et al., 1970), one might blame others for poor performance (and/or take credit for good performance).[5] This strategy will increase self-confidence, as it reduces one's own responsibility for failure. On the other hand, by engaging in positively biased information processing about their teammate's ability, individuals will end up with a more balanced weighting allocation, which reduces the material costs of belief distortion, thus providing a means of hedging against overconfidence.

Our primary interest is in fact not to pin down the direction of the bias, but to study the

[4]This empirical literature is typically focused on asymmetry in updating: whether positive signals about ability are over-weighted relative to negative signals. Different authors have found: Positive asymmetry (Eil and Rao, 2011; Möbius et al., 2014), no asymmetry (Grossman and Owens, 2012; Buser et al., 2018), and negative asymmetry (Coutts, 2019a; Ertac, 2011) have all been observed. Buser et al. (2018) do find positive asymmetry in some sub-samples. Reactions to feedback have also been studied in less comparable or non ego-relevant settings, see Barron (2020), Burks et al. (2013), Charness and Dave (2017), Eberlein et al. (2011), Erkal et al. (2019), Gotthard-Real (2017), Pulford and Colman (1997), Ertac and Szentes (2011), and Wozniak et al. (2014).

[5]If it were "easier" to blame others, this could be modeled as a lower cognitive cost of distorting information about others; see our concluding discussion. The study of self-serving attribution biases within psychology has naturally focused on environments with multi-dimensional uncertainty. That people attribute outcomes to more salient sources such as other individuals was noted by Heider (1944, 1958) and later studied by Pryor and Kriss (1977); Lassiter et al. (2002). This type of attribution has clear parallels to availability bias of Tversky and Kahneman (1973). While the overall evidence suggests significant evidence in favor of the existence of self-serving attribution biases (Mezulis et al., 2004), the resulting studies of attribution were focused on general principles rather than tractable models, discussed in Kelley (1973) and Weiner (2010). Moreover, the study of self-serving biases in psychology is generally framed as one of trade-offs for how to manage blame in order to maintain desirable beliefs (Campbell and Sedikides, 1999).

mechanics of this distortion more generally. Regardless of its direction, our theoretical insight is that in environments with multiple dimensions of uncertainty, individuals subconsciously use this tool of distortion to arrive at higher levels of overconfidence than individuals who do not manipulate their beliefs about the external fundamental (teammate's ability). To help develop this intuition, our theoretical contribution is to present a micro-founded quasi-Bayesian model of self-serving information processing with two dimensions of uncertainty. In the model, for a given signal, individuals distort the likelihood of being in a particular state (i.e. scoring in the top half) for either of these two dimensions. In other words, they may mis-perceive how indicative a signal is of their own high performance, and additionally, may mis-perceive the analogous likelihood for a teammate's performance.

In the Main treatment of our experiment, a two-person team's output depends on the ability of both members, measured through an IQ-style test, where one of the team members is the subject themselves. To properly assess whether information processing exhibits self-serving biases, an otherwise identical Control treatment removes ego-relevance, by matching a subject with another two-person team. Individual payoffs depend on the team output as well as on the decision weight that an individual places on own (Control: teammate 1's) ability relative to the other teammate's ability. Individuals receive noisy aggregate feedback, and can attribute the feedback to both their own (Control: teammate 1's) and the other teammate's ability. The updating problem is then one of joint inference; however the feedback from these two sources cannot be disentangled. Feedback is materially relevant, as the weight must be chosen immediately after it is received.

Our first result is that relative to the Control (and relative to a Bayesian benchmark), individuals in our Main treatment engage in significant self-deception in the face of feedback. Using a structural analysis we find a high degree of self-serving bias in updating about own performance: individuals significantly under-weight negative compared to positive feedback when updating beliefs. These effects are confirmed in a non-parametric matching strategy which conditions on initial priors. After receiving four rounds of feedback, individuals in our Main treatment end up 8.5 percentage points more confident about their performance than comparable individuals in Control. By matching on the ratio of positive to negative signals received, the result is strongest for those receiving a greater proportion of negative signals.

Our second result is that information processing is also significantly positively biased about teammate performance in the Main treatment. Using the same strategy which matches on initial prior beliefs, after receiving feedback, individuals end up 5.2 percentage points more optimistic about their teammate's performance in the Main treatment, relative to what individuals in the Control treatment believe about their teammates. Similar to updating about own performance, our structural estimations show these updating patterns are driven by under-weighting of negative feedback. This suggests that in the process of nurturing self-serving beliefs, individuals engage in distorted information processing about other uncertain qualities such as the ability of their teammate.

While this presents, to our knowledge, some of the first empirical evidence in economics on self-serving attribution biases with two-dimensional uncertainty, existing theoretical work has studied similar types of environments.[6] Bénabou and Tirole (2009) study a context where individuals benefiting from anticipatory feelings may engage in biased recall over their contribution to output generated by a two-person partnership, where individuals can perfectly recall their prior actions, but not information about about what type of person they are. Although able to generate self-serving beliefs, the mechanism is different from our setting, which centers on the immediate biased interpretation of feedback.[7] Our focus on biased information processing also distinguishes our model and results from recent work by Heidhues et al. (2018) and Hestermann and Le Yaouanq (2020), who examine the long run consequences of initial biases in confidence in environments with two dimensions of uncertainty with Bayesian information processing.[8]

While the long run is not our focus, crucially our results do point to a broader set of implications. A first order effect is that individuals who end up biased about other states of the world will subsequently make sub-optimal decisions. In a superficial sense, this relates to the finding of Heidhues et al. (2018), who show that overconfident individuals will subsequently make poor decisions, leading to a cycle of self-defeating learning and worse outcomes, which the individuals increasingly attribute to an external fundamental (e.g. a teammate). Yet the dynamics in our setting are very different, as we shut down the link between weighting decisions and feedback, which precludes this type of self-defeating learning. In our experiment, these sub-optimal outcomes can only occur through the channel of biased inference, not through the link between weighting decisions and outcomes. Importantly, we do find evidence that the biased information processing about teammates in our experiment leads to subsequent worse decision making. Specifically, when given a surprise opportunity to change teammates, individuals in our Main treatment are 34% less likely to be willing to pay to change teammates than their Control counterparts, who switch at optimal levels on average.

Beyond this, there are prominent second order effects. Individuals who are less likely to change environments will face fewer opportunities to learn about their true ability. This dampens learning, and as a result can exacerbate overconfidence even further. Importantly this goes contrary to the long run predictions of Bayesian inference as shown by Hestermann and Le Yaouanq (2020), who study the consequences of initial mis-calibration in confidence in a world where individuals are matched with some fundamental but can change their environment, i.e. match with a new fundamental at some cost. They show that only underconfidence will

[6]We are only aware of one other empirical study, the contemporaneous work of Goette and Kozakiewicz (2020), who conduct an experiment to test a model of self-defeating learning (Heidhues et al., 2018). Different from our paper, their main focus is to examine the link between actions and feedback, a channel which we intentionally shutdown. Beyond this, our focus on the mechanics of information processing distinguishes our work from their approach which focuses on the aggregate evolution of beliefs.

[7]While imperfect recall may be realistic in some settings, we focus on the short-term; hence our experiment was constructed to rule out opportunities to bias memory. The extent to which actions are biased to maintain self-serving beliefs is an interesting question (Bénabou and Tirole, 2011), however, we take the stance that absent these longer-run considerations, the direction of influence runs from beliefs to actions.

[8]As their primary focus, Heidhues et al. (2018) focus on an extreme form of overconfidence, where individuals believe with certainty that their ability is higher than it really is, and use Bayes' rule to update their beliefs. They do relax this assumption to show how a particular form of biased updating does not change the core predictions of their theory. In this extended framework individuals receive continuous signals about ability which are biased upwards by a fixed amount. In contrast, our setting allows for more flexible belief distortion.

persist in the long run, as initially overconfident individuals will be unsatisfied with outcomes and subsequently more likely to change environments. Importantly, our results show the exact opposite – that overconfidence could persist for similar reasons. This is consistent with real world evidence, which has found overconfidence to be more prevalent than underconfidence (Dunning, 2005).

The rest of the paper proceeds as follows. In the next section, we outline our experimental context and design. This is followed by our model, which focuses on self-serving attributions with an additional source of uncertainty. Subsequently we describe our predictions, followed by results, and conclude with a discussion.

2 Experimental Design

2.1 Overview

The experiment was conducted at the WiSo experimental laboratory at the University of Hamburg. All decisions were computerized, using z-tree (Fischbacher, 2007). A total of 426 student subjects (52% of them female) participated in 17 sessions, across two waves in the 2017-18 academic year; 192 subjects participated in wave 1, 234 subject in wave 2. Experimental sessions in the first wave lasted approximately 1 hour, in which subjects received an average payment of €14. The second wave was for the most part identical to the first but had a slight difference in the belief elicitation, and comprised an additional experimental part in which individuals could switch teammates. Experimental sessions in wave 2 lasted approximately 1.5 hours in which subjects earned on average €19.[9] Table 1 summarizes the structure of the experiment, full experimental instructions are presented in the Online Appendix Section 9.

We now describe the components of the experiment in the framework of the Main treatment. Afterwards we present the design features in which the Control treatment differs from the Main treatment. At the beginning of the experiment we provided subjects with the instructions for Part 1 and announced that they would receive the instructions for the other parts as the experiment progressed. In Part 1 subjects had 10 minutes to complete a trivia and logic test consisting of 15 questions. A timer in the upper right corner of the screen continuously informed subjects how much time was remaining on the test. The instructions stated: "Questions similar to these are often used to measure a person's general intelligence (IQ). Your task is to answer as many of these questions correctly as possible." Our priority was to emphasize the importance of the test to subjects, so that they would care about their ranking. Our intention was not to actually measure their IQ. In order to examine hard-easy effects in information processing, subjects were assigned to one of two versions of the test, one harder and one easier, randomized at the session level.[10] Subjects were unaware of these differences and were incentivized the same

[9]Earnings included a €5 show-up fee. In one session of wave 2 a fire alarm went off at the end, invalidating only data for Part 3. Due to a small glitch, some subjects inadvertently skipped entering beliefs, which leaves us with 3155 out of 3170 observations.

[10]See Larrick et al. (2007) and Moore and Small (2007) on the hard-easy effect. This effect stipulates that individuals will be more upwardly biased in estimating relative performance on easy rather than hard tasks.

Table 1: Experimental Flow

Part 1	• IQ task (10 minutes) with monetary incentives
Part 2	• Teammate 1 is matched at random to a teammate 2 • Observe # of attempted questions for teammate 2 • Report prior beliefs about teammate 1 and teammate 2 • Submit first weight **Repeated × 4 times:** • Receive feedback • Report posterior beliefs about teammate 1 and teammate 2 • Submit the weight
Part 3: Wave 2 only	• Willingness to pay to switch teammate 2 • BDM style lottery determines whether teammate 2 is switched or not • Observe # of attempted questions for (new) teammate 2 • Report beliefs about teammate 1 and teammate 2 • Submit the weight **Repeated × 4 times:** • Receive feedback • Report posterior beliefs about teammate 1 and teammate 2 • Submit the weight

way in both versions: each correct answer would earn 2.5 points while an incorrect answer would be penalized by 1 point. Unanswered questions did not affect the final score. These incentives ensured that subjects attempted a question only if they were relatively sure that they knew the answer such that the attempted number of questions (which we use in later parts of the experiment) would carry some informational value.[11] Subjects could not score below zero and were paid €0.10 per point earned in Part 1 at the very end of the experiment. At this stage no feedback on performance was given.

At the beginning of Part 2, subjects were paired into teams of two that remained constant throughout this part. Subjects' individual performances on the test from Part 1 jointly defined their "team performance" in Part 2. We neither provided subjects with any information about their teammates' identity nor about their teammates' actual test scores. Subjects only received information on the number of questions that their teammate *attempted* on the test. This figure

[11] If women are more risk averse this could lead to gender differences in the number of attempted questions, see Baldiga (2014). We do not find evidence for this in our experiment.

provided some limited information about the teammate's performance, generating variation in initial prior beliefs.

We designed the team formation protocol such that both teammates' test scores were compared to the same randomly selected group of 19 other test scores from the experimental session. Each subject could either score in the top 10 (top half) or the bottom 10 (bottom half) of this comparison group of 20, with ties broken randomly. Our main measure of interest is the degree to which subjects believe that they and their teammate score in the top half of performances. Subjects neither learned their absolute score nor whether they themselves or their teammate belonged to the top or bottom half until the end of the experiment. Not comparing teammates' scores to each other, but to the same comparison group, ensured that the teammates' individual rankings were independent of the other's score.

It was also critical for us to conduct a fully powered comparison group as a control. To this end, randomized across sessions, we varied whether subjects themselves were members of the team and hence were reporting beliefs about themselves and their teammate or whether they play the role of a third party who must report beliefs for a team composed of two different individuals. That is, in the Main treatment (226 subjects) subjects' beliefs and subsequent earnings depended on subjects' own performance, while in the Control treatment (200 subjects) own test performance was not relevant.

In Control, at the beginning of Part 2 each subject was assigned to a team consisting of two randomly selected other subjects (the teammates) from the same session. Subjects in Control were shown the screenshot of the submitted answers to the IQ quiz of one of the teammates (*teammate 1*) and were provided with information about the number of attempted questions of the other teammate (*teammate 2*). In this way, we ensured that the subjects in the Control treatment had identical information about all decision-relevant variables as the subjects in the Main treatment. As a result, by comparing reported beliefs across the Main and Control treatments, we are able to isolate biases driven by reasons of ego-protection and to abstract from other sources of belief updating biases. In the following we will consistently denote beliefs reported about own performance (in Main) and teammate 1's performance (in Control) as performance beliefs about teammate 1 and similarly, denote beliefs reported about the teammate's performance (Main) and teammate 2's performance (Control) as performance beliefs about teammate 2.

2.2 Weighting Decision and Belief Elicitation

Subjects were informed that their earnings from Part 2 would depend on their team's performance which was determined by the teammates' relative rankings in Part 1 as well as by a weighting decision that they would take during Part 2. We emphasized in the instructions that the weighting decision depended on subjects' reported beliefs and only affected subjects' own earnings. This ensured that social preferences played no role in their decisions.

The weighting decision and its direct relationship with earnings provided subjects with a transparent monetary incentive to truthfully report their beliefs about the probabilities of the two teammates scoring in the top half of performances on the IQ task. Based on subjects' reported beliefs, the computer then calculated the optimal weight and recommended how much to weight one teammate's performance relative to the other teammate's performance, using graphical tools and an explanation of which weight would give them the highest expected payoffs (see Figure 1).

Assuming subjects can form subjective beliefs, as long as they strictly prefer a higher probability of earning €10, it is in their best interest to truthfully report those beliefs. This procedure is thus novel in its indirect implementation, but shares the same incentive compatibility properties of other elicitation procedures such as matching probabilities (Holt and Smith, 2009; Karni, 2009), or the binarized scoring rule (Hossain and Okui, 2013). Like these other methods, our procedure does not require the assumption of risk-neutrality, and only requires minimal assumptions of probabilistic sophistication, see Machina (1982).[12]

Subjects were given complete information about the structure of expected payoffs. If both of the teammates were ranked in the top half of the comparison group (unknown to subjects at this point of the experiment), the subject would earn an amount of €10 for sure. Analogously, if both of the teammates were ranked in the bottom half, the subject would earn an amount of €0 for sure. If, however, one teammate was ranked in the top and the other was ranked in the bottom half, a subject's probability of earning €10 would depend on his or her weighting decision $\omega_t \in [0, 1]$. Specifically, the probability of earning €10 was given by $\sqrt{\omega_t}$ if teammate 1 scored in the top half and teammate 2 in the bottom half and $\sqrt{1-\omega_t}$ if teammate 1 scored in the bottom half and teammate 2 in the top half.

For each elicitation, subjects entered beliefs for the probability that teammate 1 scored in the top half, and the probability that teammate 2 scored in the top half. Calculating the optimal weight requires knowledge of the probabilities of the two payoff relevant states: whether teammate 1 is top and teammate 2 is bottom, and vice-versa, see Section 3.2.

In wave 1 we assumed independence between beliefs about performance of the teammates, in order to calculate the probabilities of these states. In wave 2 beliefs were additionally elicited about the probabilities of all four possible states: both top, both bottom, and teammate 1 top and teammate 2 bottom (and vice-versa). Subjects had full freedom to re-allocate these probabilities to the four relevant states as they saw fit. Screenshots of the procedure can be seen in Figure 1 (and in Online Appendix Section 9 for wave 1). Reassuringly, 90% of the time subjects chose not to alter beliefs in the four states, that is, they followed the independence assumption.[13] Strictly speaking, when faced with the 2×2 set which corresponds to each

[12]If subjects choose to enter different weights from those suggested, we are no longer able to claim incentive compatibility. Reassuringly, only 7% of weights did not correspond to the suggested optimal. Results are not affected excluding these observations. Note that theoretically there are different combinations of beliefs (in particular, sharing the same ratio) that lead to the same optimal weight. It is thus possible that subjects can arrive at the optimal weight, but intentionally report different combinations of beliefs to deceive the experimenter. We do not find this likely.

[13]Independence fails to hold after feedback, which creates dependencies between beliefs about performance of the two teammates. For the 10% that reported beliefs that were inconsistent with the independence assumption, the average difference in the belief reported was less than one percentage point. Results are robust to excluding these observations. Piloting suggested it was not intuitive for subjects to initially think about the probabilities of these four states. For this reason we first asked about the probability of teammate 1 and 2 being in the top

teammate being either in the top or bottom half, our elicitation procedure is only incentive compatible for the two payoff relevant states (in which only one of the two teammates ranked in the top half and the other in the bottom half). However, given that the vast majority of subjects do not alter beliefs in the four states, it suggests that subjects were not strategically mis-representing beliefs in the other two states. Finally, in Online Appendix Section 1 we show beliefs are nearly identical across the two waves, which additionally suggests that subjects did not alter their behavior in response to these theoretical subtleties. This is sensible, as they are hard to perceive, but beyond this, they do not generate any additional strategic motivation to not tell the truth.

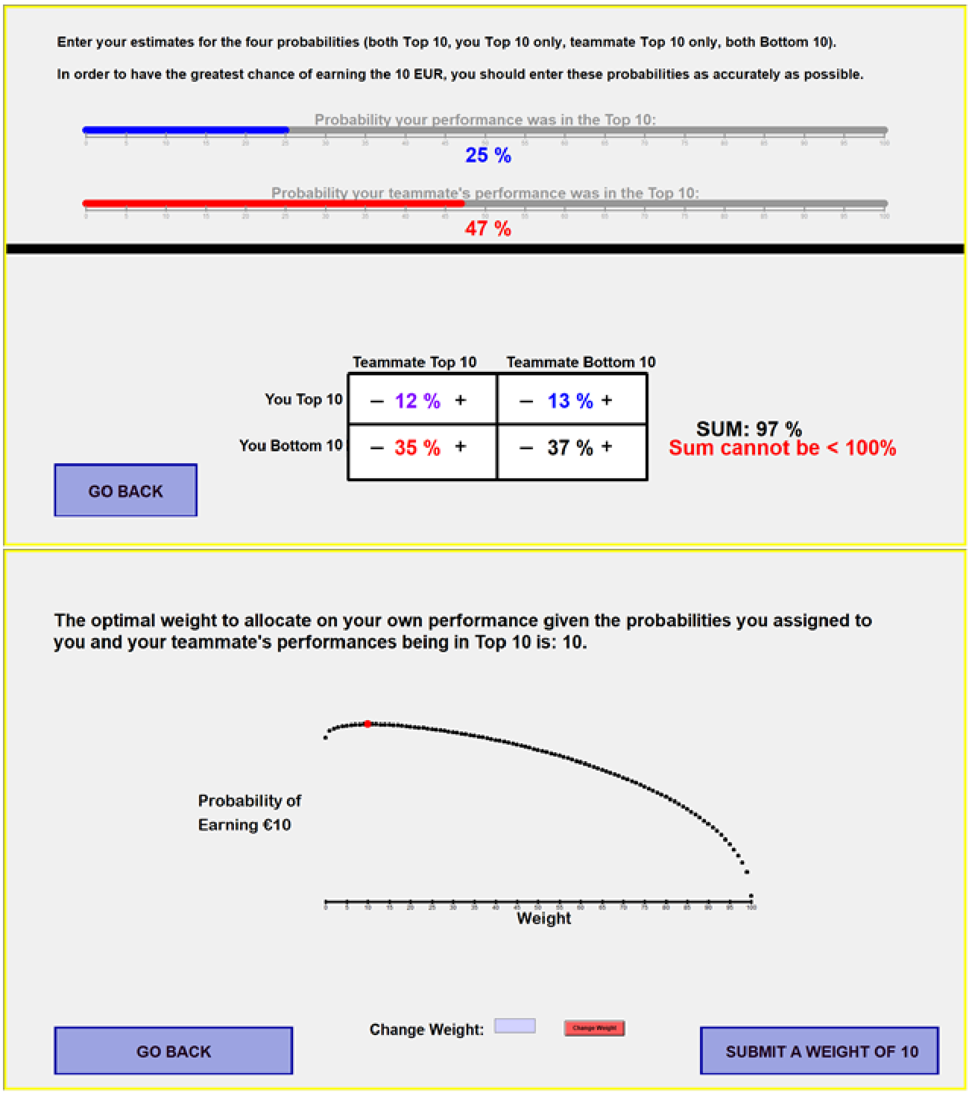


Figure 1: Screenshot of the mapping from chosen weight to probability of winning €10 which was calculated for every subject, conditional on the beliefs they entered.

2.3 Feedback

Once their weight was submitted, subjects received feedback in the form of binary signals from a "Team Evaluator", represented as a cartoon figure. Positive or negative team feedback

half.

corresponded in the experiment to the Team Evaluator giving a "Green Check" or "Red X" respectively. If both teammates scored in the top half, the Team Evaluator gave a Green Check with 90% probability and a Red X with 10% probability. If one teammate scored in the top half and the other scored in the bottom half, then the Team Evaluator gave a Green Check or a Red X with 50% probability. If both teammates scored in the bottom half, then the Team Evaluator would give the Red X with 90% probability and a Green Check with 10% probability.

Note that the feedback received from the Team Evaluator was (i) independent across feedback rounds, (ii) related to the actual performance of the teammates in Part 1 of the experiment, and (iii) depended neither on the beliefs reported by subjects nor on the previous weights submitted. This ensured that subjects did not have incentives to "experiment" with their chosen beliefs and weights to learn more about their rankings.

After receiving the Team Evaluator's feedback, subjects entered the next elicitation stage where they had to again report their beliefs that the teammates scored in the top half. Subsequently, the computer gave them a new weight recommendation which they could review and submit. This process was repeated four times. In total, subjects reported their beliefs about the teammates' performances and submitted a weight five times and received feedback from a Team Evaluator four times.

At the beginning of the Part 2, subjects were told that one of the five weighting decisions they were going to take would be selected at random and the probability of winning the €10 would depend on the selected weighting decision as well as on the teammates' performances as explained above.[14] Before the start of Part 2, subjects had to answer five control questions that were aimed at ensuring their understanding of the payment calculation, the Team Evaluator's feedback, and the weighting function. Subjects were only allowed to start Part 2 of the experiment and enter their first belief when the experimenter had checked that the answers provided were correct.

2.4 Part 3

In wave 2, at the end of Part 2, we presented subjects with a surprise opportunity to switch teammates. Specifically, we asked for their maximum willingness to pay (WTP) to be randomly re-matched with a new teammate 2 for Part 3. Our interest in WTP stems from understanding the consequences of biases in attribution for decisions to change one's environment.

Part 3 otherwise was identical to Part 2. We elicited WTP using the BDM mechanism of Becker et al. (1964). The mechanism asked subjects to enter any amount between €0 and €5 as their maximum willingness to pay to switch their teammate. The lottery would then choose a random price in the [€0, €5] interval and subjects would switch their teammate if their maximum WTP was above the chosen price and keep their teammate if this maximum WTP is below that price. Our focus is on differences in WTP across Main and Control.

[14] For more discussion on incentive compatibility of paying for one randomly selected decision in experiments see Azrieli et al. (2018). Note that in wave 2 there is an additional paid Part 3, however subjects are not aware of its structure until completing Part 2.

3 Model

3.1 Preliminaries

We first setup the theoretical framework which follows from the experimental design. An individual faces an environment with two sources of uncertainty: (i) the ability of teammate 1 (own ability in Main) and (ii) the ability of teammate 2. Following the experiment, our interests are in the discrete 2×2 state space of the ability of both teammates. Teammate 1's unknown ability is given by $A_1 \in \{B, T\}$, corresponding to either low ability (bottom half of the performance distribution) or high ability (top half). The unknown fundamental of interest $A_2 \in \{B, T\}$ is defined analogously. In the experiment this will correspond to whether teammate 2 is in the bottom half or top half of performances respectively. This leads to the four relevant states:

$$A_1A_2 = \begin{cases} TT & \text{if } A_1 = T \text{ and } A_2 = T \\ TB & \text{if } A_1 = T \text{ and } A_2 = B \\ BT & \text{if } A_1 = B \text{ and } A_2 = T \\ BB & \text{if } A_1 = B \text{ and } A_2 = B \end{cases}$$

At time t, the individual holds beliefs about the probability that the ability of teammate 1 and teammate 2 are T, given by b_t^1 and b_t^2 respectively. As in the experiment, at each time period t, individuals take an action, by choosing how much to weight the performance of teammate 1 relative to teammate 2, ω_t. Monetary payoffs at time t, are awarded probabilistically, with the possibility of earning a payment $P > 0$ or nothing. The individual will optimize by considering the payoffs of each period, which are determined according to the lottery $(P, 0; \sqrt{\omega_t})$ that pays P with probability $\sqrt{\omega_t}$ and 0 otherwise.

$$\Pi^t(\omega_t, A_1, A_2) = \begin{cases} P & \text{if } TT \\ (P, 0; \sqrt{\omega_t}) & \text{if } TB \\ (P, 0; \sqrt{1-\omega_t}) & \text{if } BT \\ 0 & \text{if } BB \end{cases} \tag{1}$$

3.2 Optimal weight

We assume that individuals are subjective expected utility maximizers, with strictly increasing utility function $u(\cdot)$. Individuals form subjective beliefs about the probabilities that teammate 1 and 2 are in the top half. Section 3.4 will describe the subconscious process underlying the formation of beliefs, however for now we take them as given. Denote beliefs about the four states at time t by $b_t^{A_1A_2}$. Thus, individuals have beliefs $b_t^1 = b_t^{TT} + b_t^{TB}$ and $b_t^2 = b_t^{TT} + b_t^{BT}$, respectively about the probability that $A_1 = T$ and $A_2 = T$ at time t.

The optimization problem of individuals is to maximize expected utility:

$$
\begin{aligned}
&b_t^{TT} \cdot u(P) \\
+b_t^{TB} \cdot \sqrt{\omega_t} \cdot u(P) + &b_t^{TB} \cdot (1 - \sqrt{\omega_t}) \cdot u(0) \\
+b_t^{BT} \cdot \sqrt{1-\omega_t} \cdot u(P) + &b_t^{BT} \cdot (1 - \sqrt{1-\omega_t}) \cdot u(0) \\
+&b_t^{BB} \cdot u(0)
\end{aligned} \tag{2}
$$

Taking first order conditions and setting the resulting equation equal to 0 yields:

$$
b_t^{TB} \cdot \frac{1}{2\sqrt{\omega_t}} \cdot [u(P) - u(0)] = b_t^{BT} \cdot \frac{1}{2\sqrt{1-\omega_t}} \cdot [u(P) - u(0)] \tag{3}
$$

This leads to the optimal weight,

$$
\omega_t^* = \frac{1}{1 + \left(\frac{b_t^{BT}}{b_t^{TB}}\right)^2}. \tag{4}
$$

Note that the optimal weight does not depend on the curvature of the utility function, $u(\cdot)$, and hence is independent of risk preferences. Unless there is certainty, extreme weights are never optimal. Intuitively, the optimal weight ω_t^* is increasing in b_t^{TB}, the belief that teammate 1 is in the top half and teammate 2 is in the bottom half, and is decreasing in b_t^{BT}, the belief that teammate 2 is in the top half and teammate 1 is in the bottom half.

Two observations are worth noting. First, given the functional form of expected utility, the optimum in Equation 4 is guaranteed to exist, and there is a unique solution for any beliefs except for the extreme case when $b_t^{TB} = b_t^{BT} = 0$.[15] Second, the optimal weight depends in opposite directions on the expected ability of teammate 1 and the expected ability of teammate 2. Thus, biases in beliefs regarding teammate 1 and 2 will be most costly when they are in opposing directions, for example, an upward bias for teammate 1 and a downward bias for teammate 2.[16]

3.3 Belief Updating

We first examine the Bayesian benchmark to study how beliefs evolve for the four states, and hence how beliefs about being in the top half evolve. Following the experiment, signals are

[15]Note that when $b_t^{TB} = 0$ and $b_t^{BT} > 0$, the unique optimal weight is $\omega_t^* = 0$. In the extreme case where both $b_t^{TB} = 0$ and $b_t^{BT} = 0$, payoffs are identical for every possible weight. Hence any weight is optimal. By the laws of probability $b_t^{TB} + b_t^{BT} \leq 1$.

[16]In period 0, this functional form generates the same self-defeating learning condition discussed in Heidhues et al. (2018). In our setup, the feedback that our individuals receive is independent of their weighting decisions, which precludes the type of self-defeating learning which they study. Heidhues et al. (2018) have a continuous state space for ability, while ours is binary. Thus, to be certain about ability and overconfident in our setting reduces to $b_0^1 = 1$. To see the result on self-defeating learning, note that one can rewrite Equation 4 in terms of priors about the ability of teammate 1 b_0^1 and teammate 2 b_0^2. Then one can see that expected utility is increasing in expected ability of teammate 1 and 2, b_0^1 and b_0^2 respectively, and the optimal weight ω^* is decreasing in the expected ability of teammate 2 b_0^2 and increasing in expected ability of teammate 1 b_0^1.

independent across time t and not perfectly informative about the states of the world (i.e. noisy). They are positive (p) with probability $\Phi_{A_1A_2} < 1$, otherwise they are negative (n). We denote them by $s_t = (p, n; \Phi_{A_1A_2})$. From now on we also make explicit the assumption that $1 > \Phi_{TT} > \Phi_{TB} = \Phi_{BT} > \Phi_{BB} = 1-\Phi_{TT} > 0$, in our experiment specifically $\Phi_{TT} = 0.9, \Phi_{TB} = \Phi_{BT} = 0.5, \Phi_{BB} = 0.1$.

A Bayesian will update beliefs about teammate 1 being in the top half given either positive (p) or negative (n) signals respectively as follows:[17]

$$\begin{aligned}
\left[b_{t+1}^{1,BAYES}|s_t = p\right] &= \frac{\Phi_{TT}b_t^{TT} + \Phi_{TB}b_t^{TB}}{\Phi_{TT}b_t^{TT} + \Phi_{TB}b_t^{TB} + \Phi_{BT}b_t^{BT} + \Phi_{BB}b_t^{BB}} \\
\left[b_{t+1}^{1,BAYES}|s_t = n\right] &= \frac{(1-\Phi_{TT})b_t^{TT} + (1-\Phi_{TB})b_t^{TB}}{(1-\Phi_{TT)}b_t^{TT} + (1-\Phi_{TB})b_t^{TB} + (1-\Phi_{BT})b_t^{BT} + (1-\Phi_{BB})b_t^{BB}}.
\end{aligned} \tag{5}$$

Analogously for teammate 2:

$$\begin{aligned}
\left[b_{t+1}^{2,BAYES}|s_t = p\right] &= \frac{\Phi_{TT}b_t^{TT} + \Phi_{BT}b_t^{BT}}{\Phi_{TT}b_t^{TT} + \Phi_{TB}b_t^{TB} + \Phi_{BT}b_t^{BT} + \Phi_{BB}b_t^{BB}} \\
\left[b_{t+1}^{2,BAYES}|s_t = n\right] &= \frac{(1-\Phi_{TT})b_t^{TT} + (1-\Phi_{BT})b_t^{BT}}{(1-\Phi_{TT)}b_t^{TT} + (1-\Phi_{TB})b_t^{TB} + (1-\Phi_{BT})b_t^{BT} + (1-\Phi_{BB})b_t^{BB}}.
\end{aligned} \tag{6}$$

3.4 Self-Serving Attribution Bias

In this section we present an updating framework which maintains the structure of Bayes' rule but allows for strategic mis-attribution of feedback across different sources. In our model, mis-attribution will correspond directly to mis-perceiving the likelihood of observing a given signal. That is, a positively biased attribution towards own performance will correspond to interpreting a signal (positive or negative) as being more indicative of high performance, compared to what the objective likelihood would suggest. In the Control treatment, since ego-utility is not at stake, we propose that there is no mis-attribution for teammate 1 and teammate 2, i.e. updating follows Bayes' rule.

In the following we focus on the case where the subject herself is teammate 1, corresponding to the Main treatment of the experiment. Thus, the driver of biased information processing comes from the benefits that individuals receive from inflating beliefs about their ability. We are agnostic over the precise source of these benefits, among the possibilities outlined in the introduction.

Following the literature, we assume that belief distortion is costly for two reasons: first, the material consequences which result from subsequent worse decision making, and second, the presence of direct mental costs of distorting beliefs. As is typical in these models, see also Brunnermeier and Parker (2005), we assume that these trade-offs occur at a subconscious level. If individuals were fully aware of their overconfidence, this would leave little scope for

[17] To derive this equation note (taking the case of a positive signal) that the probability of $s_t = p$ conditional on teammate 1 being in the top half is $\frac{\Phi_{TT}b_t^{TT}+\Phi_{TB}b_t^{TB}}{b_t^1}$. The probability of being in the top half is, b_t^1, and the perceived probability of receiving a signal $s_t = p$ is $\Phi_{TT}b_t^{TT} + \Phi_{TB}b_t^{TB} + \Phi_{BT}b_t^{BT} + \Phi_{BB}b_t^{BB}$.

the benefits of holding these biased beliefs in the first place. In this section we present a model of modified Bayesian updating which is not constrained to a biased interpretation of one dimension of uncertainty, but allows for flexible attribution across these different sources to arrive at optimal self-serving beliefs. The model's foundations are derived in Appendix A.

In our context, feedback depends on two dimensions of uncertainty: (1) own performance; and (2) performance of teammate 2 (the external fundamental). The model generates the clear prediction that attributions towards own performance will be positively biased, due to the assumed benefits of overconfidence. However, the model allows for either positive or negative attributions regarding the performance of teammate 2. The intuition for this result is that negative attributions towards one's teammate do increase self-serving beliefs (excess blame on the teammate reduces one's own responsibility by construction), a benefit, but also increase the financial costs, through more biased weighting choices.[18]

In the model, to arrive at self-serving beliefs we allow individuals to engage in distorted attributions when updating about their own performance or their teammate's performance. Starting from the Bayesian updating framework, we relax the model to include distortion parameters about own ability γ_s^1, and teammate ability γ_s^2 ($\gamma_s^i \in \mathbb{R}_+, i = 1, 2$), where $s \in \{p, n\}$ represents positive or negative signals. With regards to own performance, we assume the model of updating with self-serving attribution bias (AB) takes the following functional form for positive and negative signals respectively.

$$\left[b_{t+1}^{1,AB} | s_t = p\right] = \frac{\gamma_p^1 \gamma_p^2 \Phi_{TT} b_t^{TT} + \gamma_p^1 \Phi_{TB} b_t^{TB}}{\gamma_p^1 \gamma_p^2 \Phi_{TT} b_t^{TT} + \gamma_p^1 \Phi_{TB} b_t^{TB} + \gamma_p^2 \Phi_{BT} b_t^{BT} + \Phi_{BB} b_t^{BB}} \tag{7}$$

$$\left[b_{t+1}^{1,AB} | s_t = n\right] = \frac{\gamma_n^1 \gamma_n^2 (1-\Phi_{TT}) b_t^{TT} + \gamma_n^1 (1-\Phi_{TB}) b_t^{TB}}{\gamma_n^1 \gamma_n^2 (1-\Phi_{TT}) b_t^{TT} + \gamma_n^1 (1-\Phi_{TB}) b_t^{TB} + \gamma_n^2 (1-\Phi_{BT}) b_t^{BT} + (1-\Phi_{BB}) b_t^{BB}}$$

Regarding updating about the teammate:

$$\left[b_{t+1}^{2,AB} | s_t = p\right] = \frac{\gamma_p^1 \gamma_p^2 \Phi_{TT} b_t^{TT} + \gamma_p^2 \Phi_{BT} b_t^{BT}}{\gamma_p^1 \gamma_p^2 \Phi_{TT} b_t^{TT} + \gamma_p^1 \Phi_{TB} b_t^{TB} + \gamma_p^2 \Phi_{BT} b_t^{BT} + \Phi_{BB} b_t^{BB}} \tag{8}$$

$$\left[b_{t+1}^{2,AB} | s_t = n\right] = \frac{\gamma_n^1 \gamma_n^2 (1-\Phi_{TT}) b_t^{TT} + \gamma_n^2 (1-\Phi_{BT}) b_t^{BT}}{\gamma_n^1 \gamma_n^2 (1-\Phi_{TT}) b_t^{TT} + \gamma_n^1 (1-\Phi_{TB}) b_t^{TB} + \gamma_n^2 (1-\Phi_{BT}) b_t^{BT} + (1-\Phi_{BB}) b_t^{BB}}$$

These parameters have relatively straightforward interpretations. First, when $\gamma_s^1 = \gamma_s^2 = 1$, updating reduces to Bayesian. The larger γ_s^1 is, the greater are the positive attributions that the individual makes towards themselves, with an analogous relationship holding between γ_s^2 and the teammate. For example, a larger value of γ_s^1 increases the perceived likelihood that

[18]Our core model assumes that the mental costs of mis-attributions across the two sources are identical and independent. In our concluding discussion we argue how relaxing this assumption affects individuals' belief distortions about own and teammate's abilities.

the states TT and TB generated a signal s, the states of the world where own performance is in the top-half. Similarly, greater values of γ_s^2 increase the perceived likelihood that the states TT and BT generated a signal s. Our specification of the bias is thus reminiscent of the biased updating model of Gervais and Odean (2001).

Posterior beliefs, $b_{t+1}^{1,AB}$, are increasing in γ_s^1, but decreasing in γ_s^2; consequently self-serving bias implies that $\gamma_s^1 \geq 1$, see Appendix A. Regarding teammate 2, biased attributions necessarily do not exceed attributions about own performance, i.e. $\gamma_s^2 \leq \gamma_s^1$. However, γ_s^2 may be greater than, equal to, or less than one. On the one hand, as noted, posterior beliefs are greater for lower values of γ_s^2, hence we might expect the optimal $\gamma_s^2 < 1$. This is compatible with the psychology literature which suggests that one might expect that teammate 2 is a likely target of negative mis-attribution, i.e. blaming teammate 2 which leads to more pessimistic beliefs about their performance. On the other hand, a positive mis-attribution towards the teammate can mitigate the financial consequences of self-serving attributions in our experiment. The reason is that the optimal weight in the experiment becomes distorted, as derived in Appendix A:

$$\hat{\omega}_{t+1}^* = \frac{1}{1 + \left(\frac{\gamma_s^2 b_t^{BT}}{\gamma_s^1 b_t^{TB}}\right)^2}. \tag{9}$$

One can see that whenever $\gamma_s^1 \neq \gamma_s^2$ there is a distortion in the chosen weight relative to the Bayesian optimum. Thus while negative attributions towards teammate 2 ($\gamma_s^2 < 1$) do increase self-serving beliefs, this is ultimately costly in terms of financial penalties for submitting distorted weighting decisions.

The optimal $\gamma_s^1 \geq 1$ and $\gamma_s^2 \leq \gamma_s^1$ are such that $\left[b_{t+1}^{1,AB} | s_t = s\right] \geq \left[b_{t+1}^{1,BAYES} | s_t = s\right]$, i.e. posteriors about own performance are biased upwards. However, whether the biased posterior for teammate 2, $\left[b_{t+1}^{2,AB} | s_t = s\right]$, is smaller, equal, or larger than the Bayesian $\left[b_{t+1}^{2,BAYES} | s_t = s\right]$ depends on the value of γ_s^2.[19] Regardless of the direction, a key implication of the framework is that future decisions involving the external fundamental will result in additional negative penalties on optimal decision making.

Finally we note that we can examine the nested case of the model, where distortions only occur over one dimension of uncertainty, relating to own performance, as is the case for the papers cited in the introduction and discussed in Benjamin (2019). In this special case, $\gamma_s^2 = 1$. Because this is a restricted case, self-serving beliefs will be necessarily lower. The specific patterns of self-serving attribution bias are an empirical question, which we turn to next.

4 Hypotheses

In our theoretical model we assume that belief updating follows Bayes' rule in the Control treatment (Section 4). However, in order to allow for more flexibility and due to expected deviations from Bayes' rule, see Benjamin (2019), all of our hypotheses make comparisons

[19] If $\gamma_s^2 \leq 1$, then in our setting $\left[b_{t+1}^{2,AB} | s_t = s\right] \leq \left[b_{t+1}^{2,BAYES} | s_t = s\right]$, see Appendix A.

between the Main and Control treatments of the experiment. Only when relevant, we will refer to the Bayesian benchmark.

4.1 Belief Formation

While our main focus is on updating beliefs we also discuss belief formation and present hypotheses relating to overconfidence biases, which serve as a litmus test for whether subjects find the IQ task ego-relevant.

Our first hypothesis of interest concerns whether there is overconfidence in the Main treatment for teammate 1, relative to the control treatment benchmark. We test the following competing hypotheses, comparing initial beliefs across the two treatments. Let $b_0^{1,M}$ be the average initial ($t = 0$) belief about one's own probability of scoring in the top half, where the superscript M stands for Main treatment and 1 indicates that it is teammate 1. Similarly, $b_0^{1,C}$ refers to the initial belief for teammate 1 in the Control treatment, regarding a third party.

Hypothesis 1:

Initial beliefs are the same across Main and Control treatments.
($b_0^{1,M} = b_0^{1,C}$)

Hypothesis 1*:

Initial beliefs are larger in the Main than in the Control treatment.
($b_0^{1,M} > b_0^{1,C}$)

4.2 Belief Updating

Here we examine the implications of the model for the empirical framework, which follows Grether (1980) and Möbius et al. (2014); see Benjamin (2019) for additional references. Bayes' rule can be written in the following form, considering binary signals, s_t, for positive and negative signals respectively:

$$\frac{b_{t+1}^i}{1 - b_{t+1}^i} = \frac{b_t^i}{1 - b_t^i} \cdot LR_t^i(s) \tag{10}$$

where $LR_t^i(s)$ is the Bayesian likelihood ratio of observing signal $s_t = s \in \{p, n\}$ when updating beliefs about teammate i. For the sake of clarity, we take the perspective of updating beliefs about teammate 1; results for teammate 2 are derived similarly. From the model which includes potential attribution biases, the perceived likelihood of observing a positive signal conditional on teammate 1 being in the top half is:

$$\frac{\gamma_p^1 \gamma_p^2 0.9 b_t^{TT} + \gamma_p^1 0.5 b_t^{TB}}{b_t^{TT} + b_t^{TB}},$$

where $\gamma_p^1 = \gamma_p^2 = 1$ indicates the likelihood a Bayesian perceives. The perceived likelihood of observing a positive signal conditional on teammate 1 being in the bottom half is:

$$\frac{\gamma_p^2 0.5 b_t^{BT} + 0.1 b_t^{BB}}{b_t^{BT} + b_t^{BB}}$$

Recalling that $b_t^1 = b_t^{TT} + b_t^{TB}$, the perceived likelihood ratio, $\hat{LR}_t^1(p)$, is thus:

$$\hat{LR}_t^1(p) = \frac{\gamma_p^1 \gamma_p^2 0.9 b_t^{TT} + \gamma_p^1 0.5 b_t^{TB}}{\gamma_p^2 0.5 b_t^{BT} + 0.1 b_t^{BB}} \cdot \frac{1 - b_t^1}{b_t^1} \geq 1$$

Similarly, the perceived likelihood ratio, $\hat{LR}_t^1(n)$, is:[20]

$$\hat{LR}_t^1(n) = \frac{\gamma_n^1 \gamma_n^2 0.1 b_t^{TT} + \gamma_n^1 0.5 b_t^{TB}}{\gamma_n^2 0.5 b_t^{BT} + 0.9 b_t^{BB}} \cdot \frac{1 - b_t^1}{b_t^1} \leq 1$$

Note that the Bayesian likelihood ratios, $LR_t^i(s)$ are calculated by setting $\gamma_s^i = 1$.

Inserting the perceived likelihood ratio in Equation 10, taking natural logarithms of both sides, and adding an indicator function $I\{s_t = s\}$ for the type of signal observed,

$$logit(b_{t+1}^i) = logit(b_t^i) + I\{s_t = p\} \ln\left(\hat{LR}_t^i(p)\right) + I\{s_t = n\} \ln\left(\hat{LR}_t^i(n)\right). \tag{11}$$

The empirical model nests this Bayesian benchmark as follows,

$$logit(b_{j,t+1}^i) = \delta logit(b_{j,t}^i) + \beta_1 I(s_{j,t} = p) \ln\left(\hat{LR}_t^i(p)\right) + \beta_0 I(s_{j,t} = n) \ln\left(\hat{LR}_t^i(n)\right) + \epsilon_{j,t+1}. \tag{12}$$

δ captures the weight placed on the log prior odds ratio. β_0 and β_1 capture responsiveness to either negative or positive signals respectively. In the context of the experiment, $s_{j,t} = p$ corresponds to a positive signal, while $s_{j,t} = n$ corresponds to a negative signal. Since $I(s_{j,t} = n) + I(s_{j,t} = p) = 1$ there is no constant term. $\epsilon_{j,t+1}$ captures non-systematic errors, noting the use of j to identify the experimental subject.

Bayes' rule is a special case of this empirical model when $\delta = \beta_0 = \beta_1 = 1$, as well as $\gamma_s^i = 1$. $\delta^{1,M}$ will be used to describe the coefficient of δ for teammate 1 in the Main (M) treatment (i.e. the individual themselves), $\delta^{2,M}$ describes the coefficient of δ for teammate 2 in the Main treatment. Similarly for control (C), with analogous definitions for β_1 and β_0.

What are the implications of self-serving attribution bias for this framework? First note that $\hat{LR}_t^1(p) \geq LR_t^1(p)$ and $\hat{LR}_t^1(n) \geq LR_t^1(n)$. The intuition follows directly from the motivation for manipulating the γ_s^i in the first place – to arrive at self-serving beliefs.[21]

[20] We note that there is an implicit upper bound on γ_n^1 as this equation is ≤ 1. The reason is that we must assume that a negative signal is in fact perceived as negative information. If γ_n^1 were implausibly large, the interpretation of this would be that biased individuals actually perceive negative signals as indicating a greater likelihood of performing in the top half. Within the context of our deeper foundational model in Appendix A, we interpret this as a restriction on the shape of the mental costs of distorting γ_n^1.

[21] If any of these conditions were violated it would imply that signals are perceived as less indicative of being

Bayesian posteriors result in a weight of $\beta_1 = 1$ or $\beta_0 = 1$ on $LR_t^1(p)$ or $LR_t^1(n)$ respectively. For an individual suffering from attribution biases who perceives greater likelihood ratios, estimates of β_1 will be biased upwards for teammate 1, while estimates of β_0 will be biased downwards.[22] In other words, after either a positive or negative signal individuals will perceive the signal to be more indicative of being in the top than it really is. For teammate 2, the distortions could result in *over*-weighting or *under*-weighting of positive and negative signals. In the hypothesis, we refer to these modes of distortions as *positive bias* and *negative bias*, respectively. Since our theories of attribution bias do not alter predictions of δ, we remain agnostic over these values, and instead focus on the parameters β_0 and β_1.

Lastly, since there is no ego-utility at stake in the Control treatment, we do not expect that these individuals suffer from attribution biases that are driven by motives of ego-protection. They might, however, make some general, unsystematic mistakes in belief updating. This leads to the following competing hypotheses, comparing beliefs across the two treatments.[23]

Hypothesis 2:

Updating about one's self and teammate is the same across Main and Control treatments.
($\beta_1^{1,M} = \beta_1^{1,C};\ \beta_0^{1,M} = \beta_0^{1,C}$ *and* $\beta_1^{2,M} = \beta_1^{2,C};\ \beta_0^{2,M} = \beta_0^{2,C}$)

Hypothesis 2*:

Updating about one's self is self-serving: *individuals over-weight positive and under-weight negative signals about teammate 1 in Main compared to Control.*
($\beta_1^{1,M} > \beta_1^{1,C};\ \beta_0^{1,M} < \beta_0^{1,C}$)

And updating about teammate is biased:
Positive bias: *individuals over-weight positive and under-weight negative signals about teammate 2 in Main compared to Control.*
($\beta_1^{2,M} > \beta_1^{2,C};\ \beta_0^{2,M} < \beta_0^{2,C}$)
Or negative bias: *individuals under-weight positive and over-weight negative signals about teammate 2 in Main compared to Control.*
($\beta_1^{2,M} < \beta_1^{2,C};\ \beta_0^{2,M} > \beta_0^{2,C}$)

---

in the top than they really are. If this were the case then Bayesian updating would in fact give the individual higher utility (see also Appendix A).

[22] That β_1 is biased upwards is straightforward. Since $ln(\hat{LR}_t^1(p)) \geq 0$, a Bayesian response to in $\hat{LR}_t^1(p)$ will manifest itself as an over-response to the smaller unbiased $LR_t^1(p)$. β_0 is biased downwards because $ln(\hat{LR}_t^1(n)) \leq 0$ so a Bayesian response to $\hat{LR}_t^1(n)$ will manifest itself as an under-response to the smaller (more negative, i.e. larger in absolute value) $LR_t^1(n)$.

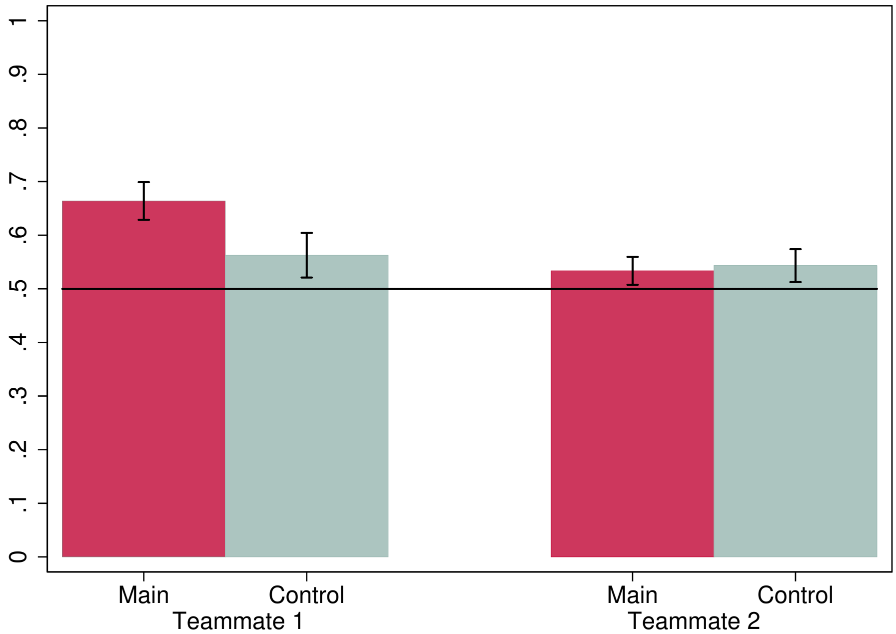
[23] In Hypothesis 2* we do not include the case of $\beta_1^{2,M} = \beta_1^{2,C}$, $\beta_0^{2,M} = \beta_0^{2,C}$, as with self-serving bias this only arises as a knife-edge (measure zero) case. In an earlier version of this paper we focused on initial predictions of self-serving mis-attributions at the expense of either the teammate or noise, but not both. These models are presented in the Online Appendix Section 8. While they generate stark predictions, neither is able to explain our results, in part due to their rigidity.

5 Results

5.1 Initial Beliefs

Figure 2 presents the first round beliefs in Main and Control treatments for both teammates. In the Main treatment, where individuals estimate beliefs about their own performance, the average reported belief about being in the top half is 66.4%, significantly different from 50% in a two-sided Wilcoxon signed rank test at the 1% level (p-value 0.0000).[24] In the Control treatment, where individuals estimate the performance of another, randomly selected individual in the position of teammate 1, the average reported belief is 56.3%. Intriguingly, this is also significantly different from 50% at the 1% level using a Wilcoxon signed rank test (p-value 0.0046). Similarly, the beliefs that teammate 2 scores in the top half are 53.4% and 54.3% in the Main and Control treatment, respectively. These beliefs are also significantly different from 50% (Wilcoxon signed rank tests p-values 0.0012 and 0.0017 respectively).

Figure 2: Prior Beliefs by Treatment



For teammate 1: Main, Belief about own performance; Control, Belief about other teammate 1's performance. For teammate 2: Belief about other teammate 2's performance. 95% Confidence intervals.

These results hence appear to present evidence for "overconfidence", according to the test of Benoît and Dubra (2011). However, as these beliefs do not involve estimation of one's own performance, we regard them as a general over-estimation that is not driven by differences in Main or Control, or in teammate 1 or teammate 2 framing: a Kruskal–Wallis test does not find a significant difference across performance beliefs about teammate 1 in Control and teammate 2 in Main and Control (p-value 0.2654). Also, there are no significant differences in

[24] Note that we use two-sided tests throughout the paper. Non-parametric tests are used as we reject normality in belief distributions, see Online Appendix Section 5.

initial beliefs about teammate 2 between the Main and Control treatment (Wilcoxon rank-sum p-value: 0.5723).

On the other hand, when we test Hypothesis 1 and compare initial beliefs about teammate 1 across the two treatments, Main (self) and Control (other), we can clearly reject equality of beliefs (Wilcoxon rank-sum test p-value: 0.0005). The results are thus in line with Hypothesis 1*. This provides robust evidence that what we are observing in the Main treatment does reflect true overconfidence. It further suggests that subjects find the IQ task ego-relevant.

Result 1: *Subjects in the Main treatment hold overconfident initial beliefs about their performance compared to the Control treatment. Initial beliefs about teammate 2 do not differ across treatments.*

Lastly, we also note that our hard-easy manipulation affects the initial beliefs. Individuals rate themselves in the top half with 72% probability when the test was easy, and with 62% when the test was hard (for more details, and a test of hard-easy effects on belief updating, see Online Appendix Section 2). While not our main focus, we also find evidence that men are more overconfident than women (further details, also concerning gender differences in belief updating are provided in Online Appendix Section 3).

5.2 Belief Updating

To study self-serving attribution bias discussed in Section 3 and to test the hypotheses from Section 4, we use Equation 12 in Section 4.2 for our primary empirical analysis. Later, in Section 5.2.2 we investigate updating biases taking a non-parametric approach, free of structural assumptions. This allows us to statistically distinguish posteriors in Main versus Control, accounting for differences in initial priors, utilizing a matching strategy. Moreover, we discuss individuals' willingness to pay (WTP) to be matched to a new teammate 2 in Section 5.3. For the interested reader we present an additional analysis of the resulting weights in Online Appendix Section 4, and examine the average evolution of beliefs in Online Appendix Section 5.

5.2.1 Structural Framework

Table 2 presents the main specification for belief updating about teammate 1 for the Main and Control treatments. Following previous literature on belief updating, we also include comparisons of the weighting of positive relative to negative signals (i.e. whether updating is asymmetric in the positive or negative direction). Our sample includes all updates from both waves, in Part 2 and 3. Samples excluding Part 3 are presented in Online Appendix Section 6, with similar results. We follow common sampling restrictions in the literature: excluding boundary observations and wrong direction updates. With two-dimensional uncertainty, we classify a wrong direction update as updating at least one belief in the wrong direction, without compensating by adjusting the other belief in the correct direction. More details are provided in Online Appendix Section 6.

Table 2: Updating Beliefs about Teammate 1

Regressor	(1) Main Treatment	(2) Control Treatment
δ	0.734***	0.751***
	(0.054)	(0.045)
β_1	0.573***	0.506***
	(0.071)	(0.075)
β_0	0.260***	0.507***
	(0.060)	(0.061)
P-Value ($\delta = 1$)	0.0000	0.0000
P-Value ($\beta_1 = 1$)	0.0000	0.0000
P-Value ($\beta_0 = 1$)	0.0000	0.0000
P-Value ($\beta_1 = \beta_0$)	0.0038	0.9906
R^2	0.56	0.60
Observations	863	829
P-Value [Chow-test] for δ (Regressions (1) and (2))		0.8089
P-Value [Chow-test] for β_1 (Regressions (1) and (2))		0.5152
P-Value [Chow-test] for β_0 (Regressions (1) and (2))		0.0040
P-Value [Chow-test] for $(\beta_1 - \beta_0)$ (Regressions (1) and (2))		0.0231

Analysis uses OLS regression. Difference is *significant from 1* at * 0.1; ** 0.05; *** 0.01. Robust standard errors clustered at individual level. R^2 corrected for no-constant. δ is the coefficient on the log prior odds ratio. β_1 and β_0 are coefficients on the log likelihood of observing positive and negative signals respectively. Constant omitted because of collinearity. Bayesian updating corresponds to $\delta = \beta_1 = \beta_0 = 1$. $\beta_1, \beta_0 < 1$ indicates conservative updating. $\beta_1 - \beta_0 > 0$ indicates positive asymmetric updating.

Updating is not Bayesian in either Main or Control. All coefficients in Table 2 are significantly different from the Bayesian prediction of 1, indicated by asterisks. Column 1 reveals that positive signals are given significantly more weight than negative signals when updating is about own performance ($\beta_1^{1,M} > \beta_0^{1,M}$, significant at the 1% level). No such asymmetry is observed in column 2, in the Control treatment, for updating about another's performance. Thus, Hypothesis 2 is rejected, updating is not the same across the Main and Control treatments.[25]

Notably $\beta_1^{1,M} > \beta_1^{1,C}$ and $\beta_0^{1,M} < \beta_0^{1,C}$. Subjects put a larger weight on positive signals and a smaller weight on negative signals when updating about teammate 1 in Main than in Control. The patterns appear consistent with the first part of Hypothesis 2*, concerning self-serving attribution bias in own belief updates. However, we only find a significant difference in response to negative, but not positive signals. Taken together, this results in $\beta_1^{1,M} - \beta_0^{1,M} > \beta_1^{1,C} - \beta_0^{1,C}$, i.e. a larger positive asymmetry in Main than in Control. We summarize our findings as follows:

Result 2: *When updating beliefs about one's self, subjects in the Main treatment display an under-responsiveness to negative signals compared to subjects from the Control treatment who update about other subjects.*

[25]We note that δ is significantly less than 1, though not different across Main and Control treatments. This is consistent with a large body of previous evidence, and indicative of base-rate neglect, see Benjamin (2019).

Table 3: Updating Beliefs about Teammate 2

Regressor	(1) Main Treatment	(2) Control Treatment
δ	0.770***	0.717***
	(0.048)	(0.050)
β_1	0.398***	0.491***
	(0.056)	(0.070)
β_0	0.248***	0.418***
	(0.043)	(0.061)
P-Value ($\delta = 1$)	0.0000	0.0000
P-Value ($\beta_1 = 1$)	0.0000	0.0000
P-Value ($\beta_0 = 1$)	0.0000	0.0000
P-Value ($\beta_1 = \beta_0$)	0.0358	0.3708
R^2	0.47	0.45
Observations	1016	916
P-Value [Chow-test] for δ (Regressions (1) and (2))		0.4408
P-Value [Chow-test] for β_1 (Regressions (1) and (2))		0.2977
P-Value [Chow-test] for β_0 (Regressions (1) and (2))		0.0235
P-Value [Chow-test] for $(\beta_1 - \beta_0)$ (Regressions (1) and (2))		0.4728

Analysis uses OLS regression. Difference is *significant from 1* at * 0.1; ** 0.05; *** 0.01. Robust standard errors clustered at individual level. R^2 corrected for no-constant. δ is the coefficient on the log prior odds ratio. β_1 and β_0 are coefficients on the log likelihood of observing positive and negative signals respectively. Constant omitted because of collinearity. Bayesian updating corresponds to $\delta = \beta_1 = \beta_0 = 1$. $\beta_1, \beta_0 < 1$ indicates conservative updating. $\beta_1 - \beta_0 > 0$ indicates positive asymmetric updating.

For a full picture of the self-serving patterns in attribution, we now examine updating about teammate 2. In our model of attribution bias, individuals either over-respond to positive signals and under-respond to negative signals or vice-versa, when updating about teammate 2 in Main compared to Control.

To identify which of these patterns are visible, Table 3 presents belief regressions for teammate 2 in Main (column 1) and Control (column 2) that are analogous to the ones in Table 2 for teammate 1. Interestingly, patterns are very similar, though less pronounced. In particular, $\beta_0^{2,M}$ and $\beta_0^{2,C}$ are significantly different at the 5% level – i.e. subjects under-weight negative feedback about their teammate when they are member of the team. Overall these results present even more evidence inconsistent with the hypothesis of equivalent updating across the Main and Control treatments (Hypothesis 2). More specifically, individuals appear to manipulate beliefs about their teammate to generate self-serving beliefs in a way that is largely in line with Hypothesis 2*, for the case of positive bias.

Result 3: *Just like for teammate 1, when updating beliefs about teammate 2, subjects in the Main treatment display an under-responsiveness to negative signals compared to subjects from the Control treatment.*

As noted earlier in Section 3 and detailed in Appendix A, some positively biased updating about teammate 2 can be optimal since it permits self-serving beliefs, while reducing the material costs of such beliefs, due to more moderate weighting between the two teammates. Interestingly, for positive signals, $\beta_1^{1,M}$ in Table 2 column 1 is significantly greater than $\beta_1^{2,M}$ in Table 3 column 1 (Chow test p-value 0.0062). For negative signals, the respective $\beta_0^{1,M}$ and $\beta_0^{2,M}$ coefficients do not differ significantly (Chow test p-value 0.8637). Taken together, the difference in asymmetry $(\beta_1^{1,M} - \beta_0^{1,M})$ versus $(\beta_1^{2,M} - \beta_0^{2,M})$ across the first columns in Tables 2 and 3 is significant at the 10% level (Chow test p-value 0.0963).[26] Hence, while we find positive asymmetry for both self and teammate 2, it is stronger when updating about one's self.

There are a few candidate alternative explanations for the observation of positively biased updating for both teammate 1 and teammate 2 in the Main treatment. We briefly discuss three more prominent ones here and address them in more detail in Online Appendix Section 7: first, that anchoring causes individuals to update similarly about teammate 2, second that positively biased updating for teammates is driven by an in-group bias, and third that subjects selectively discount or ignore negative signals. Briefly, we interpret the evidence as suggesting that these three explanations cannot explain the patterns in our data. First: raw absolute and percentage updates are not positively correlated across teammate 1 and 2 in our Main treatment, second: initial prior beliefs for teammate 2 are not statistically different across Main and Control, and third: positive asymmetry for one's self is statistically significantly stronger than for one's teammate and subjects in our Main treatment selectively ignore negative signals at equivalent rates to those in the Control treatment. Finally, we note that although we interpret our evidence as inconsistent with these explanations, the first two potential explanations could be incorporated into our underlying model, a point we return to in the concluding discussion.

5.2.2 Matching on Priors

After having shown that beliefs are updated differently in the Main versus Control treatments in a quasi-Bayesian framework, in this subsection we examine the extent to which updating differs across treatments without any reliance on the Bayesian benchmark. Specifically, we present a non-parametric analysis of updated beliefs, which utilizes a matching strategy that conditions the Main and Control subjects on their prior beliefs in round 1, and then compares their posteriors at the end of Part 2 after four rounds of feedback.[27] By matching on initial prior beliefs we are able to step away from the reliance on Bayes' rule, and instead ask the following question: given the same prior, do subjects arrive at different posteriors about their own abilities (Main treatment) versus the abilities of a randomly chosen teammate (Control treatment)? Beyond this, to ensure that these matched subjects face the same number of positive and negative signals, we force exact matching on the total number of negative signals received over the four rounds of feedback. Matching on both priors and the proportion of negative signals received summarizes all of the information that individuals have about the

[26]Moreover, this difference in the difference in asymmetry is also statistically significantly different from the difference in the difference in asymmetry in the Control treatment (Chow test p-value 0.0795).

[27]Since we are working with final posteriors, Part 3 is not comparable as it was not included in wave 1, and additionally involves some re-matching of teammates, invalidating these posteriors for this purpose.

teammates' abilities.[28]

Table 4 presents the results of this exercise reporting average treatment effects (ATE). The matching strategy reveals that individuals who are updating about their own performance (Main treatment) end up with posteriors that are 6.5 to 8.5 percentage points greater than those updating about the performance of a randomly chosen teammate 1, conditional on having the same priors and facing the *same proportion* of positive and negative signals. This provides strong evidence that information processing differs across the two treatments.

Table 4: Main vs Control: Belief Teammate 1 Top

	(1) 1 Neighbor	(2) 2 Neighbors
ATE	0.085***	0.065**
	(0.032)	(0.029)
Observations	372	372

Analysis uses nearest neighbor matching, with replacement when > 1 neighbor. Significantly different from zero at * 0.1; ** 0.05; *** 0.01. Abadie-Imbens Robust Standard Errors in parentheses. All matches received the exact same proportion of negative signals.

Table 5: Main vs Control: Belief Teammate 1 Top by Proportion of Negative Signals Received

	(1) 0/4 −	(2) 1/4 −	(3) 2/4 −	(4) 3/4 −	(5) 4/4 −
ATE	−0.015	0.104	0.139***	−0.025	0.185**
	(0.067)	(0.082)	(0.046)	(0.087)	(0.084)
Observations	73	68	99	60	72

Analysis uses nearest neighbor matching with 1 neighbor. Significantly different from zero at * 0.1; ** 0.05; *** 0.01. Abadie-Imbens Robust Standard Errors in parentheses. Each column restricts sample to specific proportion of negative signals received (out of 4 total signals).

Our structural analysis suggests this difference in updating is driven primarily by under-responsiveness to negative signals. To investigate this in our non-parametric framework, Table 5 presents matching estimates for each of the possible distributions of observed signals separately. Consistent with the structural framework, receiving 4 negative signals (0 positive) turns out to reveal the greatest difference between Main versus Control: subjects with the same initial priors end up an estimated 18.5 percentage points more confident when they are estimating their own performance. The only other significant effect is found for a balanced distribution of 2 positive and 2 negative signals.

[28]Priors of matched neighbors must be within 3 percentage points, i.e. a caliper of 0.03. The results are consistent for other calipers (available upon request).

Regarding the non-parametric estimates of the effect of differential updating about teammate 2 when one is a member of the team (Main treatment) versus not (Control), analogous regressions are presented in Tables 6 and 7. The estimates suggests that posterior beliefs about one's teammate are between 4.7 and 5.2 percentage points greater in Main relative to Control, however this is not statistically significant at conventional levels (respective p-values: 0.1294 and 0.1624). Examining the ATE estimates separately for different distributions of negative signals received, receiving all negative signals is associated with a large and significant effect. Individuals with the same priors about teammate 2 in Main and Control who receive only negative signals end up with posteriors about teammate 2 that are approximately 14 percentage points greater in Main relative to Control. Again, this supports our structural results.

Result 4: *In line with the findings from the structural framework, individuals who update about their own performance (Main treatment) end up with posteriors that are 6.5 to 8.5 percentage points greater than those who update about the performance of a randomly chosen teammate 1 (Control treatment). The bias is strongest for those who receive negative signals in all four feedback rounds. The treatment differences for updating about teammate 2 go into the same direction, but are smaller in magnitude and not statistically significant at conventional levels.*

Table 6: Main vs Control: Belief Teammate 2 Top

	(1) 1 Neighbor	(2) 2 Neighbors
ATE	0.052 (0.037)	0.047 (0.033)
Observations	374	374

Analysis uses nearest neighbor matching, with replacement when > 1 neighbor. Significantly different from zero at * 0.1; ** 0.05; *** 0.01. Abadie-Imbens Robust Standard Errors in parentheses. All matches received the exact same proportion of negative signals.

Table 7: Main vs Control: Belief Teammate 2 Top by Proportion of Negative Signals Received

	(1) 0/4 −	(2) 1/4 −	(3) 2/4 −	(4) 3/4 −	(5) 4/4 −
ATE	−0.014 (0.098)	0.077 (0.095)	0.032 (0.070)	−0.009 (0.096)	0.139** (0.063)
Observations	69	74	92	52	87

Analysis uses nearest neighbor matching with 1 neighbor. Significantly different from zero at * 0.1; ** 0.05; *** 0.01. Abadie-Imbens Robust Standard Errors in parentheses. Each column restricts sample to specific proportion of negative signals received (out of 4 total signals).

5.3 Willingness to Change Teammates

The result that self-serving motives lead to distorted interpretations of feedback regarding a teammate enables a new understanding on the persistence of overconfident beliefs. Beyond this, these resulting perceptions of one's teammate could influence future decision making. We now examine whether theses biases lead to further consequences in our experiment.

To do so, we provided our subjects with a surprise opportunity to change teammates. In wave 2 we measured the subjects' willingness to replace teammate 2 with a new (randomly selected) teammate, by submitting a willingness to pay (WTP) between 0 and 5€. Here our main interest is the extensive margin, i.e. the binary decision of whether a subject is willing to change teammates. While we also study the intensive margin in Appendix C, that analysis is confounded by the fact that the value of switching teammates depends also on beliefs about own performance.

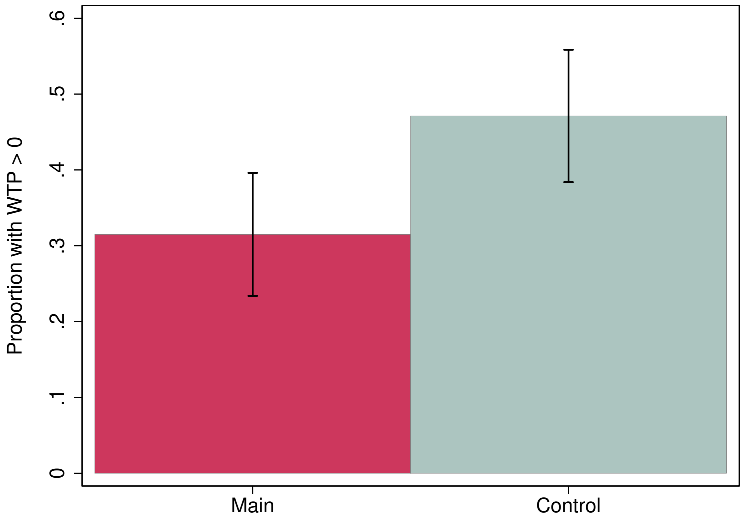
Given the patterns of biased updating we observe in our Main treatment, subjects end up with more positive performance beliefs about teammate 2. This lowers the proportion of subjects in Main who should be willing to pay to switch teammates, as Appendix C confirms given actual subject beliefs after four rounds of feedback. We also confirm this outcome in our WTP data. Figure 3 presents the proportion of subjects who submit a WTP strictly greater than zero, by Main and Control treatments. 31% of Main subjects and 47% of Control subjects were willing to pay to change teammates, a difference significant at the 5% level (Fisher's exact p-value 0.0207).

Result 5: *As a result of biased updating about teammate 2, subjects in the Main treatment are 34% less likely to want to change teammates than their Control counterparts.*

Note that this does not simply result from subject's more overconfident initial beliefs in the Main compared to the Control treatment. Before feedback, the proportion of those willing to switch teammates should be the same in both treatments. The reason is that before feedback, the decision to change teammates depends only on the belief about teammate 2's performance. Result 5 thus confirms that the biased updating patterns we observed translate into actual differences in future decision making. Moreover, it suggests that subjects are sufficiently confident about their reported beliefs that they act on them in a context which falls outside of the purview of the elicitation procedure.

In our further investigation of the intensive margin in Appendix C, we find that among those submitting a positive WTP, this WTP is smaller in the Main than Control treatment, though it is not significant at conventional levels (Wilcoxon rank-sum p-value 0.1321, $N = 89$). This finding is consistent with the model, as higher performance beliefs lead to a lower value of switching teammates, since the weight allows subjects to hedge against having a lower performing teammate.

Figure 3: Willingness to switch



Proportion of subjects who submitted strictly positive WTP to change teammate 2. Wave 2 only ($N = 231$). 95% confidence intervals shown.

6 Discussion

We believe that our model and empirical results present a novel way to navigate the different approaches to studying self-serving attribution bias across the literature of psychology and more recently economics. Our analysis offers a significant re-framing of the psychology literature – how much we credit or blame external factors shapes our beliefs about them, and consequently can alter our decisions. This subtle but powerful point means that if self-serving attributions are part-strategic (rather than naively triggered to defend our ego), economic payoffs (rather than purely psychological principles) can dictate the type of attribution to external factors. This could explain some of the mixed evidence on the strength and direction of attributions (Miller and Ross, 1975; Zuckerman, 1979).

Within economics, much of the literature has focused on the relatively narrow view of biased information processing as a one-dimensional phenomenon: self-serving attribution at the expense of "other factors" – which in empirical studies meant exclusively other idiosyncratic factors. Yet our model highlighted the one-dimensional setting as a specific nested case, where the formation of self-serving beliefs are most constrained. Broadening this focus, our experimental results suggest that when the information individuals receive depends on additional dimensions of uncertainty, they also end up distorting these dimensions to suit their desire to hold self-serving beliefs.

The result of these distortions in our experiment were that subjects ended up *positively* biased about their teammate's performance, which lowered the short term material costs, but reduced willingness to change teammates. As joining a different team provides a new, independent source of information, these short term distortions can slow down the learning process, providing a potential explanation for the persistence of overconfidence. Importantly, this contrasts with the results of Hestermann and Le Yaouanq (2020), who showed that with Bayesian updating, underconfidence, not overconfidence should persist in the long run.[29]

Yet we should not always expect positive bias, as the direction of the bias is context dependent, and varies based on both the material consequences as well as on the potentially varied cognitive costs of distortion. When should we expect positive versus negative attributions to other dimensions of uncertainty? Our theoretical framework emphasizes these critical interactions between additional dimensions of uncertainty and material and cognitive costs of distorted information processing. Our experiment focused on a particular team environment with a weighting decision, which enabled a type of material hedging of overconfidence, through positive attributions about one's teammate. However, more generally our framework provides a way forward for analyzing other contexts. Different contexts with alternative material cost structures would be expected to alter optimally biased attributions in different ways. Beyond this, the shape of cognitive costs of distortion across different dimensions will also play a key role in how attributions occur.

In our underlying model, we assumed that these costs of distortion were identical and independent across the two dimensions of uncertainty, which we believe is a natural starting point. Yet, could it be easier to distort beliefs about ourselves than others (or vice-versa)? Relatedly, there may also be differences in our ability to distort our perceptions of human versus artificially (computer) generated uncertainty. The end of Section 5.2.1 mentioned potential alternative explanations for our empirical findings. While we did not find in-group bias likely to explain our results, such a bias could manifest itself as further cognitive costs of negative distortions towards an in-group target.

Beyond this, one could consider a world where the costs of distortion across different sources of uncertainty are not independent. For example, does distortion in one dimension make distortion in another dimension more costly? Another potential alternative explanation we ruled out empirically was a simple form of anchoring, which might materialize as similar absolute or relative belief updating across the two teammates. However, there might be more complex forms of anchoring in information processing across multiple dimensions, which could be modeled as a cognitive cost of distorting these sources in opposing directions.

More broadly, our results present a way forward for thinking about how individuals select into or leave certain environments, to nurture their preferred worldview. Do people choose to work with others in anticipation of how they will rationalize good or bad outcomes? Do they choose environments in which the material costs of overconfidence are lower, or in which outcomes may be more easily attributed among various sources? These questions are critical for future research.

Apart from team settings, environments with multi-dimensional uncertainty are ubiquitous: a student receives a grade which may depend on the professor's teaching ability, or a trader realizes a return based on her portfolio and the underlying state of the economy. In the end, if self-serving belief formation motivates strategic behavior in how we choose our environments,

[29] In our experiment the opportunity to change teammates came as a surprise to subjects. To the extent that such opportunities can sometimes be predictable in the real world, we might expect this would limit the welfare consequences. We thank an anonymous referee for bringing this point to our attention.

and how we process information within those environments, we should not be surprised to find that for many individuals overconfidence could persist over the long run.

References

Azrieli, Yaron, Christopher P Chambers, and Paul J Healy, "Incentives in Experiments: A Theoretical Analysis," *Journal of Political Economy*, 3 2018.

Baldiga, Katherine, "Gender differences in willingness to guess," *Management Science*, 2014.

Barber, Brad M. and Terrance Odean, "Online Investors: Do the Slow Die First?," *Review of Financial Studies*, 1 2002, *15* (2), 455–488.

Barron, Kai, "Belief updating: does the 'good-news, bad-news' asymmetry extend to purely financial domains?," *Experimental Economics*, 4 2020, (October).

Becker, Gordon M., Morris H. Degroot, and Jacob Marschak, "Measuring utility by a single-response sequential method," *Behavioral Science*, 1964, *9* (3), 226–232.

Bénabou, Roland, "The Economics of Motivated Beliefs," *Revue d'économie politique*, 2015, *125* (5), 665.

_ **and J. Tirole**, "Self-Confidence and Personal Motivation," *The Quarterly Journal of Economics*, 8 2002, *117* (3), 871–915.

_ **and Jean Tirole**, "Over My Dead Body: Bargaining and the Price of Dignity," *American Economic Review*, 2009, *99* (2), 459–465.

_ **and** _ , "Identity, morals, and taboos: Beliefs as assets," *Quarterly Journal of Economics*, 2011, *126* (2), 805–855.

Benjamin, Daniel J., *Errors in probabilistic reasoning and judgment biases*, Vol. 2, Elsevier B.V., 2019.

Benoît, Jean-Pierre and Juan Dubra, "Apparent Overconfidence," *Econometrica*, 2011, *79* (5), 1591–1625.

Benoît, Jean Pierre, Juan Dubra, and Don A. Moore, "Does the better-than-average effect show that people are overconfident?: Two experiments," *Journal of the European Economic Association*, 2015, *13* (2), 293–329.

Bracha, Anat and Donald J. Brown, "Affective decision making: A theory of optimism bias," *Games and Economic Behavior*, 5 2012, *75* (1), 67–80.

Brunnermeier, Markus K and Jonathan A Parker, "Optimal Expectations," *American Economic Review*, 2005, *95* (4), 1092–1118.

Burks, S. V., J. P. Carpenter, L. Goette, and a. Rustichini, "Overconfidence and Social Signalling," *The Review of Economic Studies*, 1 2013, *80* (3), 949–983.

Buser, Thomas, Leonie Gerhards, and Joël van der Weele, "Responsiveness to feedback as a personal trait," *Journal of Risk and Uncertainty*, 4 2018, *56* (2), 165–192.

Camerer, Colin and Dan Lovallo, "Overconfidence and Excess Entry: An Experimental Approach," *American Economic Review*, 3 1999, *89* (1), 306–318.

Campbell, W. Keith and Constantine Sedikides, "Self-Threat Magnifies the Self-Serving Bias: A Meta-Analytic Integration," *Review of General Psychology*, 3 1999, *3* (1), 23–43.

Charness, Gary and Chetan Dave, "Confirmation bias with motivated beliefs," *Games and Economic Behavior*, 2017, *104*, 1–23.

Coutts, Alexander, "Good news and bad news are still news: experimental evidence on belief updating," *Experimental Economics*, 6 2019, *22* (2), 369–395.

_ , "Testing models of belief bias: An experiment," *Games and Economic Behavior*, 1 2019, *113*, 549–565.

Dunning, David, *Self-Insight: Roadblocks and Detours on the Path to Knowing Thyself*, New York: Psychology Press, 2005.

Eberlein, Marion, Sandra Ludwig, and Julia Nafziger, "The Effects of Feedback on Self-Assessment," *Bulletin of Economic Research*, 4 2011, *63* (2), 177–199.

Eil, David and Justin M Rao, "The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself," *American Economic Journal: Microeconomics*, 5 2011, *3* (2), 114–138.

Engelmann, Jan, Maël Lebreton, Peter Schwardmann, Joel J. van der Weele, and Li-Ang Chang, "Anticipatory Anxiety and Wishful Thinking," *SSRN Electronic Journal*, 2019.

Erkal, Nisvan, Lata Gangadharan, and Boon Han Koh, "By chance or by choice? Biased attribution of others' outcomes," *Working Paper*, 2019.

Ertac, Seda, "Does self-relevance affect information processing? Experimental evidence on the response to performance and non-performance feedback," *Journal of Economic Behavior & Organization*, 12 2011, *80* (3), 532–545.

_ **and Balazs Szentes**, "The Effect of Information on Gender Differences in Competitiveness: Experimental Evidence," *mimeo*, 2011.

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

Gervais, Simon and Terrance Odean, "Learning to Be Overconfident," *Review of Financial Studies*, 1 2001, *14* (1), 1–27.

Goette, Lorenz and Marta Kozakiewicz, "Experimental Evidence on Misguided Learning," *Working Paper*, 2020.

Golman, Russell, David Hagmann, and George Loewenstein, "Information avoidance," *Journal of Economic Literature*, 2017, *55* (1), 96–135.

Gotthard-Real, Alexander, "Desirability and information processing: An experimental study," *Economics Letters*, 2017, *152*, 96–99.

Grether, David M., "Bayes Rule as a Descriptive Model: The Representativeness Heuristic," *The Quarterly Journal of Economics*, 11 1980, *95* (3), 537.

Grossman, Zachary and David Owens, "An unlucky feeling: Overconfidence and noisy feedback," *Journal of Economic Behavior & Organization*, 11 2012, *84* (2), 510–524.

Hastorf, Albert H., David J. Schneider, and Judith Polefka, *Person perception*, Reading, Massachusetts: Addison-Wesley Publishing Company, 1970.

Heider, Fritz, "Social perception and phenomenal causality," *Psychological Review*, 1944, *51* (6), 358–374.

_ , *The psychology of interpersonal relations*, Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1958.

Heidhues, Paul, Botond Kőszegi, and Philipp Strack, "Unrealistic Expectations and Misguided Learning," *Econometrica*, 2018, *86* (4), 1159–1214.

Hestermann, Nina and Yves Le Yaouanq, "Experimentation with Self-Serving Attribution Biases," *American Economic Journal: Microeconomics*, 2020.

Holt, Charles and Angela M. Smith, "An update on Bayesian updating," *Journal of Economic Behavior & Organization*, 2 2009, *69* (2), 125–134.

Hossain, Tanjim and Ryo Okui, "The Binarized Scoring Rule," *The Review of Economic Studies*, 1 2013, *80* (3), 984–1001.

Karni, Edi, "A Mechanism for Eliciting Probabilities," *Econometrica*, 2009, *77* (2), 603–606.

Kelley, Harold H., "The processes of causal attribution.," *American Psychologist*, 1973, *28* (2), 107–128.

_ **and John L. Michela**, "Attribution Theory and Research," *Annual Review of Psychology*, 1 1980, *31* (1), 457–501.

Larrick, Richard P., Katherine A. Burson, and Jack B. Soll, "Social comparison and confidence: When thinking you're better than average predicts overconfidence (and when it does not)," *Organizational Behavior and Human Decision Processes*, 1 2007, *102* (1), 76–94.

Lassiter, G Daniel, Andrew L Geers, Patrick J Munhall, Robert J Ploutz-snyder, and David L Breitenbecher, "Illusory Causation: Why It Occurs," *Psychological science*, 2002, *13* (4), 299–306.

Machina, Mark J, ""Expected Utility" Analysis without the Independence Axiom," *Econometrica*, 1982, *50* (2), 277–323.

Mezulis, Amy H., Lyn Y. Abramson, Janet S. Hyde, and Benjamin L. Hankin, "Is There a Universal Positivity Bias in Attributions? A Meta-Analytic Review of Individual, Developmental, and Cultural Differences in the Self-Serving Attributional Bias.," *Psychological Bulletin*, 2004, *130* (5), 711–747.

Mijovic-Prelec, Danica and Drazen Prelec, "Self-deception as self-signalling: A model and experimental evidence," *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2010, *365* (1538), 227–240.

Miller, Dale T. and Michael Ross, "Self-serving biases in the attribution of causality: Fact or fiction?," *Psychological Bulletin*, 1975, *82* (2), 213–225.

Möbius, Markus, Muriel Niederle, Paul Niehaus, and Tanya Rosenblat, "Managing Self-Confidence," *mimeo*, 2014, pp. 1–43.

Moore, Don A. and Deborah A. Small, "Error and bias in comparative judgment: On being both better and worse than we think we are.," *Journal of Personality and Social Psychology*, 2007, *92* (6), 972–989.

Oster, Emily, Ira Shoulson, and E. Ray Dorsey, "Optimal Expectations and Limited Medical Testing: Evidence from Huntington Disease," *American Economic Review*, 4 2013, *103* (2), 804–830.

Prelec, Drazen and Ronit Bodner, "Self-signaling and self-control," *Time and Decision: Economic and Psychological Perspectives on Intertemporal Choice*, 2003, pp. 277–298.

Pryor, John B. and Mitchel Kriss, "The cognitive dynamics of salience in the attribution process," *Journal of Personality and Social Psychology*, 1977, *35* (1), 49–55.

Pulford, Briony D. and Andrew M. Colman, "Overconfidence: Feedback and item difficulty effects," *Personality and Individual Differences*, 7 1997, *23* (1), 125–133.

Schwardmann, Peter and Joël van der Weele, "Deception and self-deception," *Nature Human Behaviour*, 10 2019, *3* (10), 1055–1061.

Tetlock, Philip E. and Ariel Levi, "Attribution bias: On the inconclusiveness of the cognition-motivation debate," *Journal of Experimental Social Psychology*, 1982, *18* (1), 68–88.

Tversky, Amos and Daniel Kahneman, "Availability: A heuristic for judging frequency and probability," *Cognitive Psychology*, 9 1973, *5* (2), 207–232.

Weiner, Bernard, "Attribution Theory," in "A Companion to the Philosophy of Action," Oxford, UK: Wiley-Blackwell, 7 2010, pp. 366–373.

Wozniak, David, William T. Harbaugh, and Ulrich Mayr, "The Menstrual Cycle and Performance Feedback Alter Gender Differences in Competitive Choices," *Journal of Labor Economics*, 1 2014, *32* (1), 161–198.

Zimmermann, Florian, "The Dynamics of Motivated Beliefs," *American Economic Review*, 2019.

Zuckerman, Miron, "Attribution of success and failure revisited, or: The motivational bias is alive and well in attribution theory," *Journal of Personality*, 6 1979, *47* (2), 245–287.

Appendix

A Model of Optimal Information Distortion

In this section we provide a micro-foundation for self-serving attribution biases. Specifically we follow Brunnermeier and Parker (2005) by assuming that individuals engage in a subconscious optimization problem which selects the optimal belief distortion parameter $\gamma_s^i \in \mathbb{R}_+$ at the moment the individual processes new information, trading off the benefits from overconfidence against the costs. While updating beliefs over time is a dynamic problem, we assume a static model of updating. We do this to avoid the additional complexity involved in a dynamic model of optimally biased updating, but also, our focus here is on the short-run. Unlike Brunnermeier and Parker (2005) we relax the assumption of Bayesian updating, and assume that this optimization occurs directly over the updating process, through parameters γ_s^i rather than beliefs b_{t+1}^1. The updating process is precisely that outlined in Equations 7 and 8.

We introduce the possibility that individuals receive direct utility over the belief that they are in the top half, through a linear function $\alpha \cdot b_{t+1}^1$.[30] $\alpha \in [0, \infty)$ indicates the extent to which the individual benefits from holding overconfident beliefs. This can be thought of as a reduced form interpretation of the benefits to overconfidence, for example direct hedonic utility benefits, signalling to others, or benefits from motivation. Importantly, we assume that individuals do not derive any benefit from beliefs about others' ability, nor do they derive direct benefit from beliefs about the four states TT, TB, BT, BB. Of course, since $b_{t+1}^1 = b_{t+1}^{TT} + b_{t+1}^{TB}$, indirectly they can benefit from these beliefs.

We follow the literature and assume that a subconscious process trades off these benefits from overconfidence against the costs, which we posit to be material costs from inefficient decision making as well as mental costs of distorting the updating process. In the experiment, these material costs are the lower expected probability of earning $P =$ €10. Following Bracha and Brown (2012), we assume a mental cost function $J(\gamma_s^i, 1)$ that is convex and strictly increasing in $|\gamma_s^i - 1|$, i.e. is minimized at the Bayesian information processing parameter $\gamma_s^i = 1$.[31] We will further assume that the mental costs of distorting γ_s^1 and γ_s^2 are separable.

In the following we denote $\hat{b}_{t+1}^1$ as potentially biased beliefs, with b_{t+1}^1 referring to the posteriors that would arise following Bayes rule.[32] We first note that if subjects hold biased beliefs, they will submit a distorted weight in the experiment, $\hat{\omega}_{t+1}^*$, which generates material costs from foregone expected income. Critically, the optimal weight depends on beliefs about two states, $\hat{b}_{t+1}^{TB}$ and $\hat{b}_{t+1}^{BT}$. Given the form of the bias for updating about own ability, this will

[30] We choose this for simplicity, though our results would hold for both concave belief value functions, as well convex belief value functions – as long as the mental cost function was sufficiently convex to dissuade extreme beliefs.

[31] Following Bracha and Brown (2012) we further assume that $\lim_{\gamma_s^i \to \{\infty\}} J'(\gamma_s^i, 1) \to \infty$. Intuitively, absent monetary incentives the model would always predict extreme overconfidence, which seems implausible. Justifications for such a cost function are discussed in Bracha and Brown (2012). Finally, experimental evidence suggests that such mental costs are necessary if one wishes to take models of belief distortion seriously (Engelmann et al., 2019; Coutts, 2019b).

[32] In the main text we take subjective beliefs as given, and so do not follow this notation for simplicity.

imply an over-weighting of the likelihood of state TB by γ_s^1, and an over- or under-weighting of the likelihood of state BT by γ_s^2.

Under this formulation we present again the resulting biased posterior beliefs for teammate 1 and 2, as shown in Equations 7 and 8. We show the case for a positive signal, noting that the results are unchanged by replacing $\Phi_{A_1A_2}$ by the negative signal equivalent $1-\Phi_{A_1A_2}$.

$$\begin{aligned}\left[\hat{b}_{t+1}^1|s_t=p\right] &= \frac{\gamma_p^1\gamma_p^2\Phi_{TT}b_t^{TT}+\gamma_p^1\Phi_{TB}b_t^{TB}}{\gamma_p^1\gamma_p^2\Phi_{TT}b_t^{TT}+\gamma_p^1\Phi_{TB}b_t^{TB}+\gamma_p^2\Phi_{BT}b_t^{BT}+\Phi_{BB}b_t^{BB}}\\ \left[\hat{b}_{t+1}^2|s_t=p\right] &= \frac{\gamma_p^1\gamma_p^2\Phi_{TT}b_t^{TT}+\gamma_p^2\Phi_{BT}b_t^{BT}}{\gamma_p^1\gamma_p^2\Phi_{TT}b_t^{TT}+\gamma_p^1\Phi_{TB}b_t^{TB}+\gamma_p^2\Phi_{BT}b_t^{BT}+\Phi_{BB}b_t^{BB}}.\end{aligned}$$

Evidently, own beliefs should be strictly increasing in γ_p^1 for interior beliefs. To see this is the case, define $x_1=\gamma_p^1\gamma_p^2\Phi_{TT}b_t^{TT}+\gamma_p^1\Phi_{TB}b_t^{TB}$ and $x_2=\gamma_p^2\Phi_{BT}b_t^{BT}+\Phi_{BB}b_t^{BB}$. Then $\left[\hat{b}_{t+1}^1|s_t=p\right]=\frac{1}{1+\frac{x_2}{x_1}}$. Taking the derivative with respect to γ_p^1:

$$\frac{\partial\left[\hat{b}_{t+1}^1|s_t=p\right]}{\partial\gamma_p^1}=\frac{1}{\left(1+\frac{x_2}{x_1}\right)^2}\cdot\frac{x_2}{x_1^2}\cdot\left(\gamma_p^2\Phi_{TT}b_t^{TT}+\Phi_{TB}b_t^{TB}\right)>0.$$

Taking the second derivative, and letting $\bar{x}_1=\gamma_p^2\Phi_{TT}b_t^{TT}+\Phi_{TB}b_t^{TB}$:

$$\begin{aligned}\frac{\partial^2\left[\hat{b}_{t+1}^1|s_t=p\right]}{\partial^2\gamma_p^1} &= \frac{2}{\left(1+\frac{x_2}{x_1}\right)^3}\cdot\left(\frac{x_2}{x_1^2}\right)^2\cdot(\bar{x}_1)^2-\frac{2}{\left(1+\frac{x_2}{x_1}\right)^2}\cdot\frac{x_2}{x_1^3}\cdot(\bar{x}_1)^2\\ &= \frac{2x_2\left(\bar{x}_1\right)^2}{\left(1+\frac{x_2}{x_1}\right)^3\cdot x_1^4}\cdot\left(x_2-x_1\cdot\left(1+\frac{x_2}{x_1}\right)\right)<0.\end{aligned}$$

Thus own beliefs are increasing and concave in γ_p^1 (and γ_n^1, as the above are true for arbitrary $\Phi_{A_1A_2}$). We next examine how own beliefs are affected by γ_s^2. In our context they should be decreasing in γ_s^2.

Taking the derivative with respect to γ_p^2:

$$
\begin{aligned}
\frac{\partial\left[\hat{b}_{t+1}^1|s_t=p\right]}{\partial\gamma_p^2} &= \frac{1}{\left(1+\frac{x_2}{x_1}\right)^2}\cdot\frac{x_2}{x_1^2}\cdot\left(\gamma_p^1\Phi_{TT}b_t^{TT}\right) - \frac{1}{\left(1+\frac{x_2}{x_1}\right)^2}\cdot\frac{1}{x_1}\cdot\left(\Phi_{BT}b_t^{BT}\right)\\
&= \frac{1}{x_1^2\left(1+\frac{x_2}{x_1}\right)^2}\cdot\left(x_2\cdot\gamma_p^1\Phi_{TT}b_t^{TT} - x_1\cdot\Phi_{BT}b_t^{BT}\right)\\
&= \frac{1}{x_1^2\left(1+\frac{x_2}{x_1}\right)^2}\cdot\left((\gamma_p^2\Phi_{BT}b_t^{BT}+\Phi_{BB}b_t^{BB})\cdot\gamma_p^1\Phi_{TT}b_t^{TT} - (\gamma_p^1\gamma_p^2\Phi_{TT}b_t^{TT}+\gamma_p^1\Phi_{TB}b_t^{TB})\cdot\Phi_{BT}b_t^{BT}\right)\\
&= \frac{\gamma_p^1}{x_1^2\left(1+\frac{x_2}{x_1}\right)^2}\cdot\left(\Phi_{TT}b_t^{TT}\cdot\Phi_{BB}b_t^{BB} - \Phi_{TB}b_t^{TB}\cdot\Phi_{BT}b_t^{BT}\right) < 0
\end{aligned}
$$

Given our specification of the signal structure $\Phi_{A_1A_2}$, $\Theta = \Phi_{TT}b_t^{TT}\cdot\Phi_{BB}b_t^{BB} - \Phi_{TB}b_t^{TB}\cdot\Phi_{BT}b_t^{BT} < 0$, as detailed in Section B. Hence $\frac{\partial\left[\hat{b}_{t+1}^1|s_t=p\right]}{\partial\gamma_p^2} < 0$, and similarly for γ_n^2.

Regarding the second derivative, it is positive, recalling that $\Theta < 0$:

$$
\begin{aligned}
\frac{\partial^2\left[\hat{b}_{t+1}^1|s_t=p\right]}{\partial^2\gamma_p^2} &= \frac{2\gamma_p^1\cdot\Theta}{x_1^2\left(1+\frac{x_2}{x_1}\right)^3}\cdot\left(\frac{x_2}{x_1^2}\cdot\left(\gamma_p^1\Phi_{TT}b_t^{TT}\right)-\frac{1}{x_1}\cdot\left(\Phi_{BT}b_t^{BT}\right)\right) - \frac{2\gamma_p^1\cdot\Theta}{x_1^3\left(1+\frac{x_2}{x_1}\right)^2}\cdot\gamma_p^1\Phi_{TT}b_t^{TT}\\
&= \frac{2(\gamma_p^1)^2\cdot\Theta}{x_1^4\left(1+\frac{x_2}{x_1}\right)^3}\cdot(\Theta) - \frac{2\gamma_p^1\cdot\Theta}{x_1^3\left(1+\frac{x_2}{x_1}\right)^2}\cdot\gamma_p^1\Phi_{TT}b_t^{TT} > 0.
\end{aligned}
$$

Thus own beliefs are a decreasing and convex function of γ_p^1 (and γ_n^1, noting that $\Phi_{TT} = 1-\Phi_{BB}$ and $\Phi_{TB} = \Phi_{BT}$). Finally we note that by symmetry, all of these results apply analogously to beliefs about teammate 2 performance, $\hat{b}_{t+1}^2$. That is, they are increasing in γ_s^2 and decreasing in γ_s^1.

Given the impact of the distortion parameters γ_s^i on own beliefs, we can turn to the impact of these parameters on other elements of the decision problem. The resulting (biased) optimal weight is $\hat{\omega}_{t+1}^*$. From Equation 4, setting $\Phi_{BT} = \Phi_{TB} = 0.5$, we have[33]

$$
\hat{\omega}_{t+1}^* = \frac{1}{1+\left(\frac{\gamma_s^2 b_t^{BT}}{\gamma_s^1 b_t^{TB}}\right)^2} \tag{13}
$$

[33] We note that, given the biased updating process, this is simplified from the following equation (analogously for a negative signal): $\frac{\hat{b}_{t+1}^{BT}}{\hat{b}_{t+1}^{TB}} = \frac{\frac{\gamma_p^2\Phi_{BT}b_t^{BT}}{\gamma_p^1\gamma_p^2\Phi_{TT}b_t^{TT}+\gamma_p^1\Phi_{TB}b_t^{TB}+\gamma_p^2\Phi_{BT}b_t^{BT}+\Phi_{BB}b_t^{BB}}}{\frac{\gamma_p^1\Phi_{TB}b_t^{TB}}{\gamma_p^1\gamma_p^2\Phi_{TT}b_t^{TT}+\gamma_p^1\Phi_{TB}b_t^{TB}+\gamma_p^2\Phi_{BT}b_t^{BT}+\Phi_{BB}b_t^{BB}}} = \frac{\gamma_p^2\Phi_{BT}b_t^{BT}}{\gamma_p^1\Phi_{TB}b_t^{TB}}$.

This leads to the following optimization problem, taking into account the mental cost functions:

$$\max_{\{\gamma_s\}} \Bigg\{ \alpha \cdot \hat{b}^1_{t+1} + b^{TT}_{t+1} \cdot u(P) + b^{TB}_{t+1} \cdot \sqrt{\hat{\omega}^*_{t+1}} \cdot u(P) + b^{TB}_{t+1} \cdot (1 - \sqrt{\hat{\omega}^*_{t+1}}) \cdot u(0) \tag{14}$$
$$+ b^{BT}_{t+1} \cdot \sqrt{1 - \hat{\omega}^*_{t+1}} \cdot u(P) + b^{BT}_{t+1} \cdot (1 - \sqrt{1 - \hat{\omega}^*_{t+1}}) \cdot u(0) + b^{BB}_{t+1} \cdot u(0)$$
$$- J(\gamma^1_s, 1) - J(\gamma^2_s, 1) \Bigg\}$$

There are three important forces at work here. The first term involves the belief utility benefits from increasing γ^1_s and decreasing γ^2_s. The middle terms present the financial payoffs, which are maximized when $\gamma^1_s = \gamma^2_s$, resulting in an unbiased weight. The final two terms are mental costs, which are minimized when $\gamma^i_s = 1$, i.e. updating is Bayesian.

By the properties of the mental cost function $J(\gamma^i_s, 1)$, extreme values of γ^i_s are never optimal, and thus we restrict our attention to an interior solution. We also will restrict our focus to solutions with $\gamma^1_s \geq 1$, without loss of generality to the paper's predictions.[34] Substituting biased beliefs and weights into the maximization, and substituting the values of Φ from the experiment, the first order condition with respect to γ^1_s is (where $u(P) - u(0) = \Delta u$):

$$\alpha \cdot \frac{\partial \left[\hat{b}^1_{t+1} | s_t\right]}{\partial \gamma^1_s} + \frac{\gamma^2_s \cdot \left(b^{TB}_{t+1} \cdot b^{BT}_{t+1}\right)^2 \cdot \Delta u}{\left(\left(\gamma^2_s b^{BT}_{t+1}\right)^2 + \left(\gamma^1_s b^{TB}_{t+1}\right)^2\right)^{\frac{3}{2}}} \cdot (\gamma^2_s - \gamma^1_s) - J'(\gamma^1_s, 1) \tag{15}$$

The first order condition with respect to γ^2_s is:

$$\alpha \cdot \frac{\partial \left[\hat{b}^1_{t+1} | s_t\right]}{\partial \gamma^2_s} + \frac{\gamma^1_s \cdot \left(b^{TB}_{t+1} \cdot b^{BT}_{t+1}\right)^2 \cdot \Delta u}{\left(\left(\gamma^2_s b^{BT}_{t+1}\right)^2 + \left(\gamma^1_s b^{TB}_{t+1}\right)^2\right)^{\frac{3}{2}}} \cdot (\gamma^1_s - \gamma^2_s) - J'(\gamma^2_s, 1) \tag{16}$$

Result 1: When $\alpha = 0$ there will be no belief distortion.

This result derives directly from setting the two FOCs equal to zero. When $\alpha = 0$ the unique optimal solution is to set $\gamma^1_s = \gamma^2_s = 1$.

Result 2: $\gamma^1_s \geq \gamma^2_s$.

This result derives from the second FOC. By contradiction, if $\gamma^1_s < \gamma^2_s$, the equation setting the FOC equal to zero cannot be satisfied.

If $\alpha = 0$, the optimal $\gamma^1_s = \gamma^2_s = 1$. When $\alpha > 0$, $\gamma^1_s > 1$, while the optimal γ^2_s may be less than, equal to, or greater than 1, though $\gamma^s_2 < \gamma^s_1$. The reason why γ^2_s is not unambiguously smaller than one is that there is a benefit to updating in a biased way about teammate 2, which counter-balances the biased updating about teammate 1, leading to a closer to optimal

[34] Note that self-serving beliefs can arise from setting $\gamma^1_s > 1$ or $\gamma^2_s < 1$. Regarding the latter case, while unlikely in our setting, it does not preclude that $\gamma^1_s < 1$. As the distortions of both parameters must lead to upwardly biased posteriors about own performance to be optimal, all of the results in the main paper are unaffected. In our context it is also sufficient to include a condition such as $\gamma^2_s \geq \frac{\gamma^1_s}{2}$, or $\gamma^2_s \geq \frac{1}{2}$ to rule out $\gamma^1_s < 1$.

weighting decision.

When $\alpha = 0$ updating is Bayesian for both teammates. When $\alpha > 0$ the resulting biased updating leads to inflated posteriors about own performance, while posteriors about the teammate's performance may be inflated or deflated. A sufficient condition for posteriors about the teammate's performance to be lower than Bayesian is $\gamma_s^2 < 1$, since $\frac{\partial\left[\hat{b}_{t+1}^2|s_t=s\right]}{\partial\gamma_s^2} > 0$ and $\frac{\partial\left[\hat{b}_{t+1}^2|s_t=s\right]}{\partial\gamma_s^1} < 0$. By continuity, for any $\gamma_s^1 > 1$, there exists $1 < \gamma_s^2 < \gamma_s^1$ such that posteriors are greater than Bayesian, since posteriors are lower than Bayesian for $\gamma_s^2 = 1$ and greater than Bayesian for $\gamma_s^2 = \gamma_s^1$.

B Deriving the condition for $\Theta < 0$

B.1 Theoretical Result

In this section we show that starting from any non-degenerate prior beliefs and assuming that individuals update according to our model of self-serving attribution bias,

$$\begin{aligned}\Theta &= \Phi_{TT}b_t^{TT} \cdot \Phi_{BB}b_t^{BB} - \Phi_{TB}b_t^{TB} \cdot \Phi_{BT}b_t^{BT} \\ &= (1-\Phi_{TT})b_t^{TT} \cdot (1-\Phi_{BB})b_t^{BB} - (1-\Phi_{TB})b_t^{TB} \cdot (1-\Phi_{BT})b_t^{BT} < 0.\end{aligned}$$

In particular, we show that this condition will hold whenever $\Phi_{TT} \cdot \Phi_{BB} - \Phi_{TB} \cdot \Phi_{BT} < 0$. This is satisfied in our experiment as $0.9 \cdot 0.1 - 0.5 \cdot 0.5 = -0.16 < 0$.

Denote prior beliefs by b_0^1, b_0^2. In the first round the performances of both teammates are independent, hence $b_0^{TT} = b_0^1 \cdot b_0^2$, $b_0^{TB} = b_0^1 \cdot (1 - b_0^2)$, and so on.

The expression of interest in the first round is thus:

$$\begin{aligned}&\Phi_{TT}(b_0^1 \cdot b_0^2) \cdot \Phi_{BB}((1-b_0^1)\cdot(1-b_0^2)) - \Phi_{TB}(b_0^1\cdot(1-b_0^2))\cdot\Phi_{BT}((1-b_0^1)\cdot b_0^2) \\ &\qquad = (b_0^1 \cdot b_0^2)((1-b_0^1)\cdot(1-b_0^2))\cdot[\Phi_{TT}\cdot\Phi_{BB} - \Phi_{TB}\cdot\Phi_{BT}] \qquad (17)\end{aligned}$$

Thus, this expression will be negative, whenever $\Phi_{TT} \cdot \Phi_{BB} - \Phi_{TB} \cdot \Phi_{BT} < 0$.

We now consider the next round of updating, after a positive signal is received. We show the case for state TT, but the derivation is analogous for the other three states.

$$[b_1^{TT}|s_t = p] = \frac{\gamma_p^1\gamma_p^2\Phi_{TT}\cdot b_0^{TT}}{\gamma_p^1\gamma_p^2\Phi_{TT}\cdot b_0^{TT} + \gamma_p^1\Phi_{TB}\cdot b_0^{TB} + \gamma_p^2\Phi_{BT}\cdot b_0^{BT} + \Phi_{BB}\cdot b_0^{BB}}$$

We note that the denominator of beliefs for all four states will be identical. Denote it by $\mathcal{D}_1 = \gamma_p^1\gamma_p^2\Phi_{TT}\cdot b_0^{TT} + \gamma_p^1\Phi_{TB}\cdot b_0^{TB} + \gamma_p^2\Phi_{BT}\cdot b_0^{BT} + \Phi_{BB}\cdot b_0^{BB}$. We now substitute these expressions for the four states back into the initial expression of interest, Equation 17:

$$\frac{1}{\mathcal{D}_1}\left(\Phi_{TT}^2\gamma_p^1\gamma_p^2 b_0^{TT}\Phi_{BB}^2 b_0^{BB} - \Phi_{TB}^2\gamma_p^1 b_0^{TB}\cdot\Phi_{BT}^2\gamma_p^2 b_0^{BT}\right)$$

We now note that this is simply an iteration of Equation 17. As such it reduces to:

$$= \frac{\gamma_p^1 \gamma_p^2}{\mathcal{D}_1} \left((b_0^1 \cdot b_0^2)((1 - b_0^1) \cdot (1 - b_0^2)) \cdot [(\Phi_{TT} \cdot \Phi_{BB})^2 - (\Phi_{TB} \cdot \Phi_{BT})^2] \right) < 0$$

We continue this inductive process once more:

$$[b_2^{TT} | s_t = p] = \frac{\gamma_p^1 \gamma_p^2 \Phi_{TT} \cdot b_1^{TT}}{\gamma_p^1 \gamma_p^2 \Phi_{TT} \cdot b_1^{TT} + \gamma_p^1 \Phi_{TB} \cdot b_1^{TB} + \gamma_p^2 \Phi_{BT} \cdot b_1^{BT} + \Phi_{BB} \cdot b_1^{BB}}$$

Where we denote $\mathcal{D}_2 = \gamma_p^1 \gamma_p^2 \Phi_{TT} \cdot b_1^{TT} + \gamma_p^1 \Phi_{TB} \cdot b_1^{TB} + \gamma_p^2 \Phi_{BT} \cdot b_1^{BT} + \Phi_{BB} \cdot b_1^{BB}$ and so hence:

$$\begin{aligned} [b_2^{TT} | s_t = p] &= \frac{\gamma_p^1 \gamma_p^2 \Phi_{TT} \cdot \frac{\gamma_p^1 \gamma_p^2 \Phi_{TT}}{\mathcal{D}_1} b_0^{TT}}{\mathcal{D}_2} \\ &= \frac{\left(\gamma_p^1 \gamma_p^2 \Phi_{TT}\right)^2 \cdot b_0^{TT}}{\mathcal{D}_1 \cdot \mathcal{D}_2} \end{aligned}$$

Thus we arrive at the third term:

$$= \frac{(\gamma_p^1 \gamma_p^2)^2}{\mathcal{D}_2 \cdot \mathcal{D}_1} \left((b_0^1 \cdot b_0^2)((1 - b_0^1) \cdot (1 - b_0^2)) \cdot [(\Phi_{TT} \cdot \Phi_{BB})^3 - (\Phi_{TB} \cdot \Phi_{BT})^3] \right) < 0$$

Following this process, assume the k^{th} posterior is given by:

$$[b_k^{TT} | s_t = p] = \frac{\left(\gamma_p^1 \gamma_p^2 \Phi_{TT}\right)^k \cdot b_0^{TT}}{\mathcal{D}_1 \cdots \mathcal{D}_k}$$

Then the $k + 1^{th}$ posterior:

$$[b_{k+1}^{TT} | s_t = p] = \frac{\gamma_p^1 \gamma_p^2 \Phi_{TT} \cdot b_k^{TT}}{\gamma_p^1 \gamma_p^2 \Phi_{TT} \cdot b_k^{TT} + \gamma_p^1 \Phi_{TB} \cdot b_k^{TB} + \gamma_p^2 \Phi_{BT} \cdot b_k^{BT} + \Phi_{BB} \cdot b_k^{BB}}$$

In particular, the $k + 1^{th}$ term of this inductive process is:

$$= \frac{(\gamma_p^1 \gamma_p^2)^k}{\mathcal{D}_1 \cdots \mathcal{D}_{k+1}} \left((b_0^1 \cdot b_0^2)((1 - b_0^1) \cdot (1 - b_0^2)) \cdot [(\Phi_{TT} \cdot \Phi_{BB})^{k+1} - (\Phi_{TB} \cdot \Phi_{BT})^{k+1}] \right) < 0$$

We note that given $\Phi^{TT} \cdot \Phi^{BB} = 0.09$ and $\Phi^{TB} \cdot \Phi^{BT} = 0.25$, this expression is strictly negative for all positive integers k.

B.2 Empirical Result

Without making any assumptions on the updating process, we can also simply examine the value of the expression: $\Phi_{TT}b_t^{TT} \cdot \Phi_{BB}b_t^{BB} - \Phi_{TB}b_t^{TB} \cdot \Phi_{BT}b_t^{BT}$, given actual beliefs in the experiment, and check whether it is less than or equal to 0. In fact in only 2% of cases is this expression positive.

C WTP to Switch Teammates

In wave 2 we provided subjects with the opportunity to be randomly re-matched to a new teammate 2, using the BDM mechanism. Subjects i could bid $x_i \in €[0, 5]$, where €5 is the risk-neutral maximum value of switching.[35] After submitting their bid, the computer randomly generated a price, $p \in [0, 1]$ using a continuous distribution. Whenever $x_i > p$ they would pay the price p out of their earnings, and be matched with a new teammate. If $x_i \leq p$ they would not pay anything, and stay matched with the same teammate.

Given the reported beliefs of subjects we are able to calculate whether it would be optimal for them to switch teammates, assuming risk neutrality. Before receiving feedback, this decision depends entirely on the belief about teammate 2. If subjects believe their teammate is in the top half with probability less than 50% they should pay to switch, otherwise they should not be willing to pay any positive amount.[36]

Since initial beliefs about teammate 2 are not statistically different across Main and Control treatments, we would predict that the number of subjects willing to pay a positive amount to switch teammates will be the same across both groups. Figure C.1 confirms this is the case given prior beliefs in Main and Control (Round 1). This figure plots the theoretically optimal proportion of subjects which should opt to switch teammates.

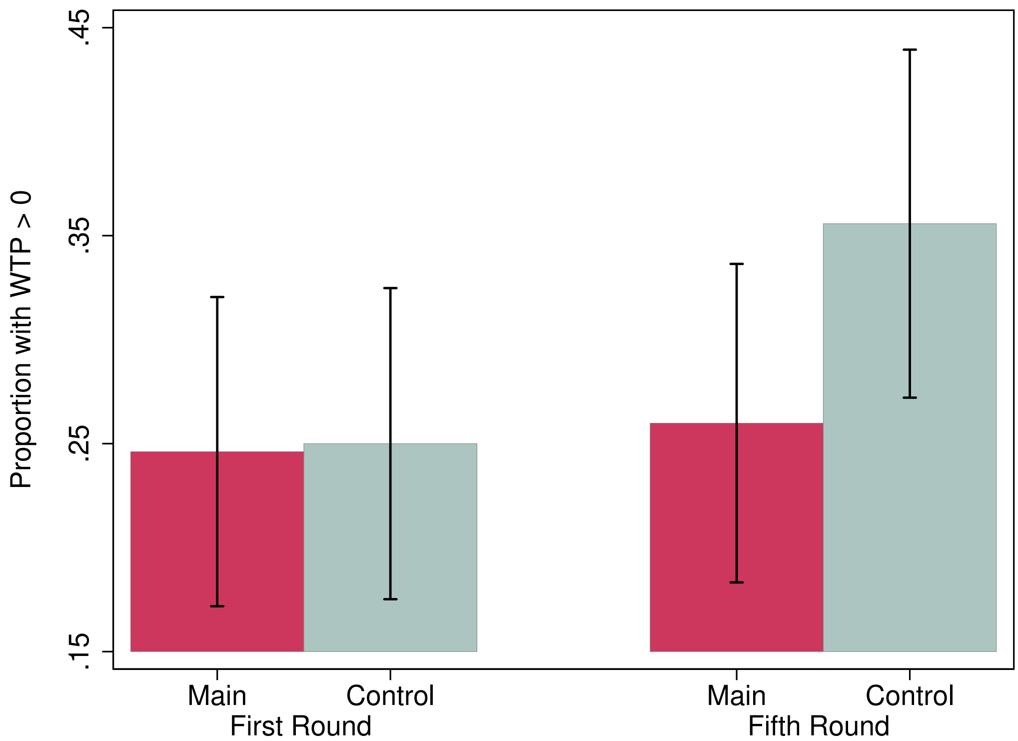
While initial prior beliefs are such that there are no differences across Main and Control treatments, beliefs after four rounds of feedback (Round 5) are such that in fact a higher proportion of individuals in Control should be willing to switch teammates. This is because in Control, subjects update in a symmetric way about their teammate, and end up with more moderate beliefs.[37] In Main, because of the positive bias in updating about the teammate, there is no corresponding increase in the proportion that should switch teammates. As was shown in Figure 3, this is indeed the case for actual subject decisions.

[35] Note that the worst outcome for subjects is when both teammates are in the bottom half, where they will earn €0 with certainty. If one is in the top half, they can select ω accordingly to ensure a high probability of earning €10. Since there is a 50% probability a randomly selected person is in the top half, the expected value of being matched with them is €5.

[36] One exception is if they believe with probability 1 that they themselves are in the top half, since they can choose a weight of $\omega = 1$ and mitigate any effect of a bad teammate. Note also that the *price* one is willing to pay is decreasing in beliefs about own performance. Higher performers are better able to hedge using their own performance, through choosing the optimal weight.

[37] In fact, since beliefs are initially slightly inflated about teammate 2, they end up with more pessimistic (but accurate) beliefs in Control.

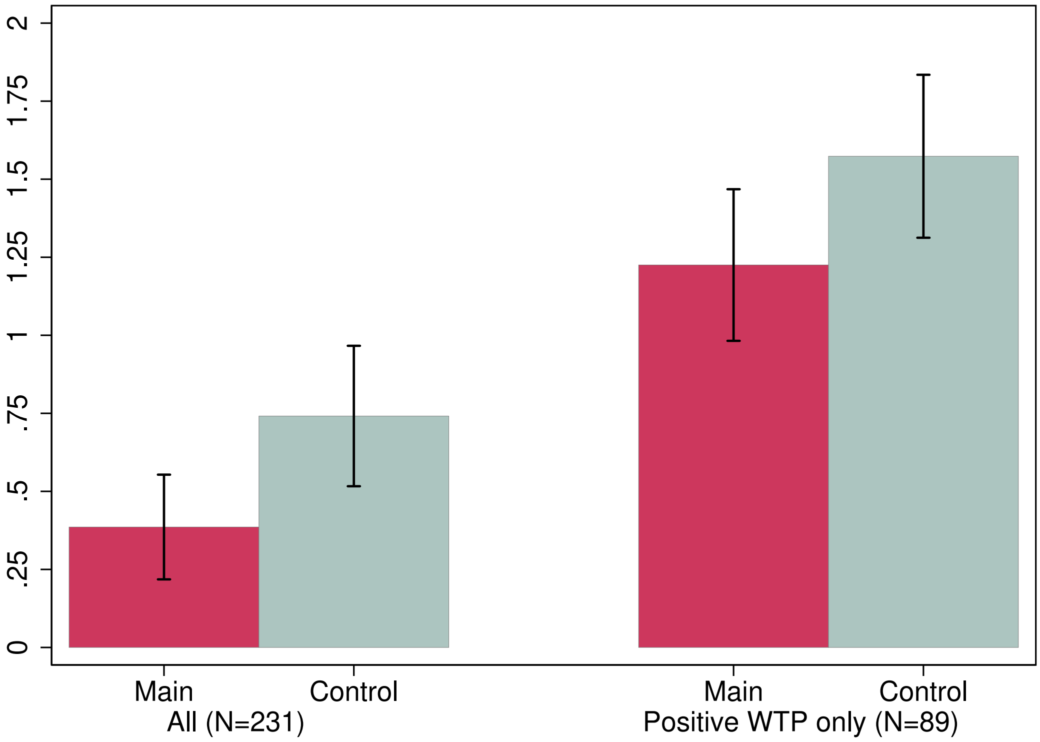
Figure C.1: (Calculated) Optimal Proportion Willing to Switch



Given subject beliefs, this shows the proportion of subjects that would (hypothetically) gain from switching teammates. 95% confidence intervals shown.

Figure C.2 presents the actual values of WTP submitted. The average WTP in Main is €0.39, while in Control it is €0.74, significantly different at the 1% level (Wilcoxon rank-sum p-value 0.006). Restricting the sample only to positive WTP, the Wilcoxon rank-sum p-value is 0.132, $N = 89$. Thus while there is lower WTP among this restricted sample in Main treatment relative to Control, this can be accounted for by the more overconfident beliefs in Main, for which there is less material benefit to having a new teammate.

Figure C.2: Willingness to pay



WTP (in Euro) of subjects to switch teammate 2. Left side includes all data, right side includes only positive values of WTP. Wave 2 only. 95% confidence intervals shown.