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

Alexander Coutts Leonie Gerhards Zahra Murad

February 16 2020

Abstract

We investigate how overconfidence persists in the face of objective feedback which depends on two dimensions of uncertainty. Self-serving attribution biases exist when individuals arrive at overconfident beliefs through biases in how they process information. Yet how individuals manipulate information from their environment to arrive at these beliefs is not known. We present a modified Bayesian framework to study attribution biases and to distinguish whether self-serving information processing results in biased perceptions of other fundamental states of the world. In an experiment where individuals receive noisy performance feedback that also depends on a teammate, we identify precise patterns in attribution. We find that individuals do update in a biased, self-serving way. Moreover, we find that in nurturing these self-serving beliefs, they also end up with positively biased beliefs about their teammate. Such beliefs can discourage individuals from changing their environment, impeding learning about ability. We confirm in our experiment that individuals are less likely to change teammates, highlighting an important mechanism which could help explain the persistence of overconfidence.

*__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 (email: 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 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, SHUFE, THEEM, and WZB. We gratefully acknowledge financial support from the Hamburgische Wissenschaftliche Stiftung and the University of Hamburg.

1 Introduction

Overconfidence has been shown to be a persistent bias in human decision making, and has been linked to financial decision making (Barber and Odean, 2001), CEO investment decisions (Malmendier and Tate, 2005), as well as career choice (Köszegi, 2006). The persistence of overconfidence is especially puzzling when considering that individuals receive informative feedback about their ability in many contexts. In this paper we study how overconfidence persists through biased information processing about self-relevant information, specifically when this information comes bundled with an additional fundamental source of uncertainty.[1] For example, consider a student who receives a grade for group work, an employee who receives a bonus based on her team's performance, or a trader who realizes a return based on her portfolio and the underlying state of the economy.

A large literature in psychology is dedicated to the study of self-serving attribution bias: an over-attribution of past successes to internal factors such as ability, relative to failures which are attributed to external factors (Mezulis et al., 2004). For example, the student above would be biased if he takes credit for high grades, but blames colleagues or bad luck for low grades. The underlying motivation for such behavior is often traced back to Freudian principles: the pleasures associated with success and the pains associated with failure (Weiner and Graham, 1999). Such motivated cognition thus can enhance pleasure and/or reduce pain through biased attribution patterns.[2]

Our focus in this paper is to explore the dynamics of such self-serving attributions when noisy feedback about internal qualities such as ability comes bundled with an external fundamental source of uncertainty.[3] While self-serving attributions will lead to overconfident beliefs, it is not known whether attribution biases will also impact assessments of other states of the world, namely the external fundamental. One prominent consequence appears if these assessments are biased upwards, lowering the expected returns to changing environments (i.e the fundamental), and hence ultimately reducing opportunities to learn about one's true intrinsic qualities.

A key contribution of our paper is to test these dynamics empirically in a controlled experiment. In our Main treatment, a two-person team's output depends on the ability of both members, measured through an IQ-style test. Individual payoffs from the team's output depend on both members' abilities as well as on the weight that an individual places on his or her own ability relative to the teammate's ability (the external fundamental).[4] Individuals receive noisy

[1] We abstract away from other channels involving assessment of past or future information, such as biased memory or selective information acquisition. Considering the past, hindsight-bias or biased memory theories, see Fischhoff (1975) and Bénabou and Tirole (2002), could lead to overconfidence if individuals recall information in a biased way; Zimmermann (2019) in fact finds evidence of asymmetric recall of feedback. Regarding the future, individuals may selectively sample information, choosing only sources of information that are likely to nurture overconfidence, e.g. Eliaz and Spiegler (2006).

[2] There could also be self-motivational or signalling motives for ego-enhancement or protection, see Bénabou and Tirole (2002) and our later discussion.

[3] Following Heidhues et al. (2018) and Hestermann and Le Yaouanq (2019) we refer to these fundamental quantities as stable dimensions of uncertainty, in contrast with idiosyncratic noise.

[4] This weight can be thought of as an effort delegation decision within an organizational context. Similarly,

aggregate feedback, and can attribute the feedback to both their own and their teammate's ability. The updating problem is then one of joint inference; however the feedback from these two sources cannot be disentangled. To properly assess whether individuals exhibit self-serving attributions, we compare subjects' resulting beliefs to those of a fully powered Control treatment, which is identical but removes the ego-relevance of the feedback. Subjects in this control group are matched with another two-person team and observe feedback about this team, that is unrelated to their own ability.

Our first main result is that relative to the Control (and relative to a Bayesian benchmark), individuals in our Main experiment do engage in self-serving attributions. Using a structural analysis we find a high degree of positive asymmetry in updating about own performance: significantly under-weighting negative to positive feedback when updating beliefs. These effects are confirmed in a non-parametric matching strategy which matches on initial priors. After receiving four rounds of feedback, subjects in our Main experiment end up 8.4 percentage points more confident about their performance than comparable Control subjects. By matching on the sequence of signals received, the result is strongest for those receiving mostly negative signals.

Our second main result is that these self-serving attributions also affect beliefs about the teammate. We find that in our Main experiment, subjects also end up positively biased about their teammates. Using the same strategy which matches on prior beliefs, after receiving feedback, subjects end up 5.2 percentage points more optimistic about their teammate's performance in the Main experiment, relative to what subjects in the Control experiment believe about their teammates. Similar to updating about own performance, our structural estimations suggest these updating patterns are driven by under-weighting of negative feedback. As a result, we show that when given a surprise opportunity to change teammates, individuals in our Main experiment are 34% less likely to be willing to pay to change teammates than their Control counterparts, who switch at optimal levels. Combined these results show that the mechanics which precipitate self-serving beliefs also affect beliefs about other states of the world, leading to subsequently biased decision making.

Our findings are in line with a micro-founded quasi-Bayesian model of self-serving attribution bias, and our theoretical contribution is to show that the observed behavior is indeed optimal. In the experiment, subjects' expected payoffs depend on how they allocate the weight between their own and their teammate's performance. Positive attributions about own ability will cause an over-weighting of own performance relative to one's teammate. Yet positive bias towards one's teammate creates a countervailing, downward bias on the weight. Thus, the patterns we observe are consistent with subjects savoring the ego-benefits of holding overconfident beliefs, without suffering the totality of the financial consequences.

There are many examples of such dynamics in real world group settings. Consider an overconfident individual who must decide how much group work to delegate to the other members. If she believes that the others are average or poor performers, she will expect that to achieve a good outcome, she will need to do the brunt of the work. Yet if she can subconsciously

in an example of a non-team context, the weight could represent a trader's implicit decision of how much effort to put into a portfolio choice in a given market.

convince herself that the other members are also high performers, she can allocate this work more equally among all members. In doing so, she can avoid the subsequent effort costs which would result from overconfidence.

Our results generate new insights in relation to an emerging literature on learning with two dimensions of uncertainty. Motivated reasoning can spill over to distort beliefs about non-ego relevant states of the world. This result has a number of important implications. A first order effect is that individuals who end up biased about other states of the world will subsequently make sub optimal decisions. A student who ends up positively biased about their group members will be less likely to change to a potentially better group. A trader who ends up positively biased towards market fundamentals, may inefficiently remain active in the market or a CEO who holds unjustifiably high beliefs about a certain business unit, might be unwilling to divest this unit.

Yet there are also 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. Such a result goes counter to rational models of updating with two dimensions of uncertainty which predict that only under-confidence will persist in the long run (Hestermann and Le Yaouanq, 2019). The intuition for these dynamics in rational models is that initially overconfident agents 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 significantly more prevalent than underconfidence (Dunning, 2005).

Overall our results document how attribution biases can distort other dimensions of uncertainty, leading to clear consequences for decision making. As most feedback in the real world comes bundled with more than one dimension of uncertainty, the implications of our results are far-reaching. The rest of the paper proceeds as follows. After a literature review, we outline our experimental context and design. This is followed by our theory, which focuses on self-serving attributions with an additional source of uncertainty. We finally describe our predictions, followed by results, and conclude with a short discussion.

2 Related Literature

Our study links multiple strands of literature in economics and psychology, namely: those on overconfidence, attribution biases, and belief updating. Behavior consistent with overconfidence about ability has been documented in numerous settings, such as driving (Svenson, 1981), financial trading (Barber and Odean, 2001), as well as in a number of lab experiments concerning tests of academic ability. Benoît and Dubra (2011) noted that rational behavior may generate overconfident-appearing data. Yet even accounting for this, many studies have found evidence consistent with overconfidence, see Benoît et al. (2015), and the discussion contained therein.

To understand the presence and persistence of overconfidence, a broad literature has emerged

within economics on motivated cognition, which explores the motivations for holding self-serving beliefs. The benefits to overconfidence may arise from (i) direct utility from holding overconfident beliefs (Möbius et al., 2014; Brunnermeier and Parker, 2005) for example arising from self-esteem or ego-protection, (ii) benefits to personal motivation or self-signalling (Bénabou and Tirole, 2002, 2009, 2011), or (iii) strategic signalling motives/persuasion of others (Burks et al., 2013; Schwardmann and Van der Weele, 2018).[5]

A key progression of this literature was to specify and empirically test models of belief updating in ego-relevant settings (Buser et al., 2018; Coutts, 2019a; Eil and Rao, 2011; Ertac, 2011; Grossman and Owens, 2012; Möbius et al., 2014; Schwardmann and Van der Weele, 2018).[6] In one of these earlier studies, Möbius et al. (2014) present a theory which provides a common motivation for this line of research: a model of asymmetric updating bias that arises from a world where individuals derive direct utility from believing they have high ability, à la Brunnermeier and Parker (2005).[7]

These aforementioned studies have examined updating with one ego-relevant dimension of uncertainty, which can potentially capture some elements of self-serving attribution biases. However the focus on one-dimensional uncertainty precludes the comprehensive study of how self-serving biases may arise or how they may alter perceptions more broadly. In contrast, our two-dimensional setting contributes on two fronts. First, we are uniquely able to examine whether attributions spill over to affect other external fundamentals. And second, moving to two-dimensional uncertainty could alter the scope for self-serving beliefs, potentially increasing them, through altering the constraints that individuals face from distorting beliefs.

The premise that two or more dimensions of uncertainty could alter self-serving biases can be traced to a long-standing literature in social psychology. The study of attribution bias has its origins in the writings of Fritz Heider. Heider (1944, 1958) described the innate human desire to explain behaviors and outcomes, noting that people tend to attribute outcomes to more salient sources such as other individuals, rather than luck, with clear parallels to availability bias of Tversky and Kahneman (1973). It follows that the presence of salient factors could potentially enable stronger attributions.[8]

[5]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) positive presentation of self to others, and (iii) belief in effective control; see Kelley and Michela (1980) and Tetlock and Levi (1982).

[6]Evidence of asymmetric information processing is mixed, see Benjamin (2019). 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 settings, see Barron (2017), Burks et al. (2013), Eberlein et al. (2011), Erkal et al. (2019), Pulford and Colman (1997), Ertac and Szentes (2011), and Wozniak et al. (2014).

[7]Related theory and evidence in finance on self-serving attribution bias can be found in Daniel et al. (1998) and Gervais and Odean (2001).

[8]Some related evidence of this can be found in Pryor and Kriss (1977) with further discussion in Lassiter et al. (2002). 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). Attributions were empirically tested by asking subjects to express responsibility for outcomes among listed internal vs external factors (Miller and Ross, 1975; Mezulis et al., 2004). See Pekrun and Marsh (2018) for a more detailed discussion of some empirical concerns of this literature. As Silvia and Duval (2001) note, some concepts such as "luck" are not

While, to our knowledge, ours is the first empirical study of self-serving attribution biases with two-dimensional uncertainty within economics, the topic has recently been studied by Heidhues et al. (2018) and Hestermann and Le Yaouanq (2019).[9] Both model the theoretical long run consequences of confidence biases for decision making with two dimensions of uncertainty: ability and another external fundamental, assuming Bayesian updating. In contrast our focus is on short term updating biases and their consequences for future decision making. However, it is worth discussing the overlap with each paper in turn. When possible we discuss their theory within our context of teams.

Hestermann and Le Yaouanq (2019) 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. Initially overconfident individuals rationally attribute successes as reflective of their ability, while they attribute failures as reflective of the fundamental. There are asymmetric dynamic consequences of initial biases in confidence: overconfident individuals end up being dissatisfied with their environment (and hence quit "too early"), while initially underconfident individuals are more likely to be satisfied with the environments they find themselves in, and hence may remain "stuck".[10]

In contrast, Heidhues et al. (2018) assume that individuals believe with certainty that their ability is higher than it really is, and remain matched to a constant underlying fundamental.[11] They demonstrate that under certain conditions, an overconfident individual will perceive poor outcomes as reflecting poor performance by another teammate rather than herself. In response, they show that individual decision making can lead to a cycle of self-defeating learning and poor outcomes, which the agent increasingly attributes to her teammate.

Our setup is a variation of both these models, though with the crucial difference that we study non-Bayesian information processing due to self-serving attribution bias. Like Heidhues et al. (2018), our framework involves a delegation-type decision between two teammates. However, in our environment we shut-down the feedback mechanism from this decision, which precludes the type of self-defeating learning they study. In our setup, these dynamics can only occur through the channel of biased inference, not through the link between weighting decisions and outcomes.

straightforward to interpret outside of a quantitative framework.

[9] A related theoretical paper is Deimen and Wirtz (2019), who examine the optimal strategy of an agent who faces dual uncertainty about own ability and the returns to effort. Differing from our paper, their focus is on the optimal effort decision to facilitate learning.

[10] Our experimental setup relates to their theory, as our feedback structure is a particular case of their setup, where there is neither complementarity nor substitutability between teammates' abilities.

[11] Regarding the overconfidence assumption, they take steps to show how it can be relaxed, by considering a form of biased updating and showing that this 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. This differs in both scope and consequence from our theory.

3 Experimental Design

3.1 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.[12]

Depending on the wave, the experimental session comprised two or three parts, as summarized in Table 1 (full experimental instructions are presented in the Online Appendix Section 8). Below we describe the components of the experiment in the framework of the Main treatment in more detail. 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 in emphasizing 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.[13] Subjects were unaware of these differences and were incentivized the same 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.[14] 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

[12]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.

[13]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 estimates of their relative performance on easy rather than hard tasks.

[14]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 effect in our experiment.

Table 1: Experimental Flow

Part 1	• IQ task (10 minutes) • Piece rate paid: wrong answers penalized
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

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 provided some limited information about the teammate's performance, generating variation in 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 each other.

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.

3.2 Weighting Decision and Belief Elicitation

Subjects were informed that their *individual* financial rewards 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 test. 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).[15] Assuming subjects can form subjective beliefs, as long as they strictly

[15]If subjects choose to enter different weights from those suggested, we are no longer able to claim incentive

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).

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. Without additional assumptions, see Section 4.2, 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.

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 8 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.[16] Strictly speaking, when faced with the 2×2 set which corresponds to each 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 again 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

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.

[16]Independence fails to hold after feedback, which create 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 half.

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.

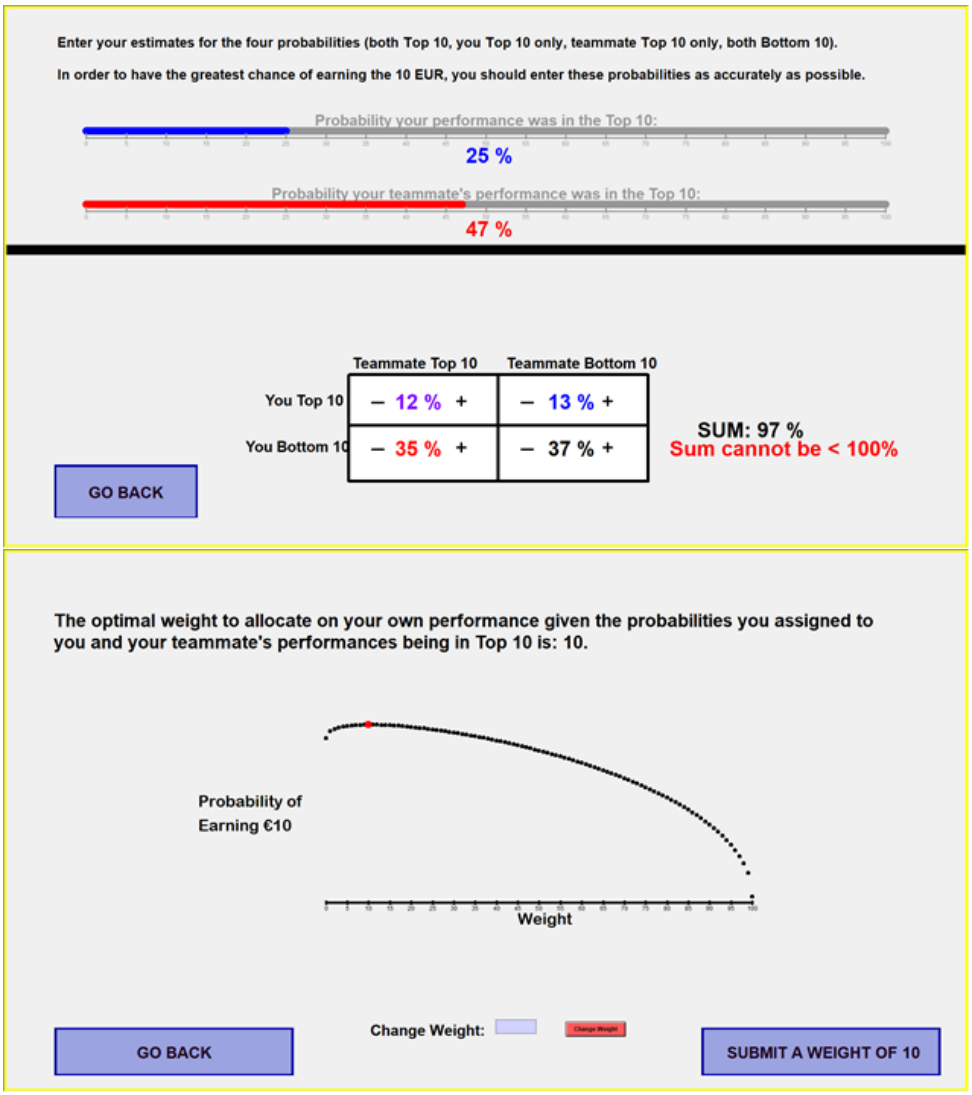


Figure 1: Screenshot of the mapping from chosen weight to probability of winning €10 which was calculated for every subject, conditional on the beliefs they entered.

3.3 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 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% 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.[17] 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.

3.4 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 switch their teammate 2 for Part 3, i.e. be randomly re-matched with a new teammate 2. 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.

[17] 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.

4 Theory

4.1 Preliminaries

We first setup the theoretical framework which follows from the experimental design. An individual faces an uncertain 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 following lottery. $(P, 0; \sqrt{\omega_t})$ is the lottery that pays P with probability $\sqrt{\omega_t}$ and 0 otherwise.

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

4.2 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. Denote beliefs about the four states at time t by $b_t^{A_1A_2}$. Thus, agents 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. Thus, biases in beliefs regarding teammate 1 and 2 will be most costly when they are in opposing directions, e.g. an upward bias for teammate 1 and a downward bias for teammate 2.

We note a few things. 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 $b_t^{TB} = b_t^{BT} = 0$.[18] The optimal weight depends in opposite directions on the expected ability of teammate 1 and the expected ability of teammate 2.[19]

4.3 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 independent across time t and positive (p) with probability $\Phi_{A_1 A_2}$, otherwise they are negative

[18] 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 $b_t^{TB} + b_t^{BT} \leq 1$.

[19] 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 agents 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.

(n). We denote them by $s_t = (p, n; \Phi_{A_1 A_2})$. From now on we also make explicit the assumption that: $1 > \Phi_{TT} > \Phi_{TB} = \Phi_{BT} > \Phi_{BB} = 1 - \Phi_{TT} > 0$, 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:[20]

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

Analogously for teammate 2:

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

4.4 Self-Serving Attribution Bias

In this section we present an updating framework which maintains the structure of Bayes' rule but allows for mis-attribution of feedback across different sources. In the Control treatment, since ego-utility is not at stake, we propose that belief formation for teammate 1 and teammate 2 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 Section 2.

We assume that belief distortion is costly for two reasons: first, the financial consequences which result from subsequent worse decision making, and second, the presence of direct mental costs of distorting beliefs, as in Bracha and Brown (2012).[21] In this section we present a model of modified Bayesian updating which allows for flexible attribution across the different sources of uncertainty. The model's foundations are derived in Appendix A, resulting from the trade-off between the benefits and costs of self-serving attributions, following our experimental design.

In our context, feedback depends on two sources of uncertainty: (1) own performance; and (2) performance of teammate 2 (the external fundamental). There is also the further

[20] 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 $\Phi_{TT} b_t^{TT} + \Phi_{TB} b_t^{TB} + \Phi_{BT} b_t^{BT} + \Phi_{BB} b_t^{BB}$.

[21] Engelmann et al. (2019) present experimental evidence consistent with the existence of such mental costs, while Coutts (2019b) also shows empirically that patterns in belief distortion cannot be rationalized without such mental costs. 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 the benefits of holding these biased beliefs in the first place.

element of noise. Noise is present since signals are not perfectly informative about the states of the world, i.e. $\Phi_{A_1,A_2} \in (0,1)$. The resulting theory allows for us to solve for the optimal form of self-serving attributions across these different sources. The theory 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, a benefit, but also increase the financial costs, through more biased weighting choices.

As an alternative to this main model, we first present a myopic, constrained model of biased attribution, which imposes the restriction of unbiased attribution towards the teammate. In this myopic model, self-serving beliefs can only come at the expense of noise – as a result, individuals end up unbiased about their teammate's performance. This type of updating behavior would be consistent with the existing empirical literature in economics on studying ego-relevant belief updating with one-dimensional uncertainty. We view this as an important special case, as it follows from this case that self-serving attribution biases will not lead to biased beliefs about other fundamentals, unlike the more general model, which we discuss immediately after.[22]

To arrive at self-serving beliefs we assume that the agent can engage in distorted attributions when updating about the two sources of uncertainty. Starting from the Bayesian updating framework, we relax the model to include distortion parameters about own ability γ_s^1, where $s \in \{p, n\}$ represents positive or negative signals. With regards to own performance, we assume the general model of updating with myopic attribution bias (MAB) takes the following functional form for positive and negative signals respectively.

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

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

These parameters have relatively straightforward interpretations. First, when $\gamma_s^1 = 1$, updating reduces to Bayesian. The larger γ_s^1 is, the greater are the positive attributions that the agent makes towards themselves. For example, a larger value of γ_s^1 increases the perceived likelihood that the states TT and TB generated a signal s, the states of the world where own performance is in the top-half. Our specification of the bias is thus similar to the biased updating model of Gervais and Odean (2001).

[22] In an earlier version of this paper we focused on initial predictions of self-serving mis-attributions at the expense of either factor (2) the teammate or (3) noise, but not both. These models are presented in the Online Appendix Section 7. While they generate stark predictions, neither is able to explain our results, in part due to their rigidity.

In our model, $\gamma_s^1 \geq 1$, whereas Bayesian updating prescribes that $\gamma_s^1 = 1$. The former implies that $\left[b_{t+1}^{1,MAB}|s_t = s\right] \geq \left[b_{t+1}^{1,BAYES}|s_t = s\right]$. Agents engage in positively biased attributions regarding their own performance, a result which follows from the foundations built in Appendix A. Regarding updating about teammate 2, agents are Bayesian, i.e. $\gamma_s^1 = 1$. These updating patterns are myopic in the sense that subjects bias their inferences about the strength of the signal across two states when it comes to updating about own ability, but are unbiased about these same two states when it comes to their teammate's ability. Thus biased updating results in upward biased beliefs about self, but unbiased beliefs about one's teammate.

We now turn to the more general model. In this general model, we allow the agent to engage in distorted attributions when updating about both sources of uncertainty. Namely, we relax the model to include additional distortion parameters γ_s^2. These parameters have an analogous interpretation: larger values of γ_s^2 increase the perceived likelihood that the states TT and BT generated a signal s. With regards to own performance, we assume the general model of updating with attribution bias (AB) takes the following functional form for positive and negative signals respectively.

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

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

Regarding updating about the teammate:

$$\left[b_{t+1}^{2,AB}|s_t = p\right] = \frac{\gamma_p^1\gamma_p^2\Phi_{TT}b_t^{TT} + \gamma_p^2\Phi_{BT}b_t^{BT}}{\gamma_p^1\gamma_p^2\Phi_{TT}b_t^{TT} + \gamma_p^1\Phi_{TB}b_t^{TB} + \gamma_p^2\Phi_{BT}b_t^{BT} + \Phi_{BB}b_t^{BB}} \tag{9}$$

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

Posterior beliefs, $b_{t+1}^{1,AB}$, are increasing in γ_s^1, but decreasing in γ_s^2; consequently $\gamma_s^1 \geq 1$, 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 or less than one. On the one hand, the psychology literature suggests that one might expect that teammate 2 is a likely target of negative mis-attribution, i.e. attribution biases will lead to more pessimistic beliefs about the performance of teammate 2, $\gamma_s^2 < 1$. The logic has parallels to availability bias, (Tversky and Kahneman, 1973), due to the salience of the teammate as a stable factor to be blamed. 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:

$$\hat{\omega}^*_{t+1} = \frac{1}{1 + \left(\frac{\gamma^2_s b^{BT}_t}{\gamma^1_s b^{TB}_t}\right)^2}. \tag{10}$$

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

The optimal $\gamma^1_s \geq 1$ and $\gamma^2_s \leq \gamma^1_s$ are selected such that $\left[b^{1,AB}_{t+1}|s_t = s\right] \geq \left[b^{1,BAYES}_{t+1}|s_t = s\right]$, i.e. posteriors about own performance are biased upwards. However, whether the biased posterior for teammate 2, $\left[b^{2,AB}_{t+1}|s_t = s\right]$, is smaller, equal, or larger than the Bayesian $\left[b^{2,BAYES}_{t+1}|s_t = s\right]$ depends on the value of γ^2_s.[23] 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.

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

5.1 Belief Formation

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

Our first null hypothesis of interest concerns whether there is overconfidence in the Main treatment for teammate 1. Let $b^{1,M}_0$ 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, $b^{1,C}_0$ refers to the initial belief for teammate 1 in the Control treatment, regarding a third party.

Hypothesis 1:

$$b^{1,M}_0 = b^{1,C}_0,$$

[23] If $\gamma^2_s \leq 1$, then in our setting $\left[b^{2,AB}_{t+1}|s_t = s\right] \leq \left[b^{2,BAYES}_{t+1}|s_t = s\right]$, see Appendix A.

5.2 Belief Updating

Here we examine the implications of the theory 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, $s_t = s \in \{p, n\}$, for positive and negative signals respectively:

$$\frac{b^i_{t+1}}{1 - b^i_{t+1}} = \frac{b^i_t}{1 - b^i_t} \cdot LR^i_t(s) \tag{11}$$

where $LR^i_t(s)$ 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 focus this discussion from the perspective of updating beliefs about teammate 1; results for teammate 2 are derived similarly. We present the estimation strategy for our main model of self-serving attribution bias, as MAB can be represented as a specific case when $\gamma^2_s = 1$, and updating is Bayesian for teammate 2. From the theory which includes potential attribution biases, the perceived likelihood of observing a positive signal conditional on teammate 1 being in the top half is:

$$\frac{\gamma^1_p \gamma^2_p 0.9 b^{TT}_t + \gamma^1_p 0.5 b^{TB}_t}{b^{TT}_t + b^{TB}_t},$$

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

$$\frac{\gamma^2_p 0.5 b^{BT}_t + 0.1 b^{BB}_t}{b^{BT}_t + b^{BB}_t}$$

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

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

Similarly, the perceived likelihood ratio, $\hat{LR}^1_t(n)$, is:[24]

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

Note that the Bayesian likelihood ratios, $LR^i_t(s)$ are calculated by setting $\gamma^i_s = 1$.

Taking natural logarithms of both sides of Equation 11, using the perceived likelihood ratio,

[24] We note that there is an implicit upper bound on γ^1_n 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^1_n 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^1_n.

and an indicator function, $I\{s_t = s\}$, for the type of signal observed,

$$logit(b^i_{t+1}) = logit(b^i_t) + I\{s_t = p\} \ln\left(\hat{LR}^i_t(p)\right) + I\{s_t = n\} \ln\left(\hat{LR}^i_t(n)\right). \tag{12}$$

The empirical model nests this Bayesian benchmark as follows,

$$logit(b^i_{j,t+1}) = \delta logit(b^i_{j,t}) + \beta_1 I(s_{j,t} = p) \ln\left(\hat{LR}^i_t(p)\right) + \beta_0 I(s_{j,t} = n) \ln\left(\hat{LR}^i_t(n)\right) + \epsilon_{j,t+1}. \tag{13}$$

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 model when $\delta = \beta_0 = \beta_1 = 1$, as well as $\gamma^i_s = 1$. $\delta^{1,M}$ will be used to describe the coefficient of δ for teammate 1 in the main (M) sessions (i.e. the individual themselves), $\delta^{2,M}$ describes the coefficient of δ for teammate 2 in the main sessions. 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 $\hat{LR}^1_t(p) \geq LR^1_t(p)$ and $\hat{LR}^1_t(n) \geq LR^1_t(n)$. The intuition follows directly from the motivation for manipulating the γ^i_s in the first place – to arrive at self-serving beliefs.[25]

Bayesian posteriors result in a weight of $\beta_1 = 1$ or $\beta_0 = 1$ on $LR^1_t(p)$ or $LR^1_t(n)$ 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.[26] In other words, after a positive signal individuals will perceive the signal to be more indicative of being in the top than it really is. After a negative signal they will perceive the signal to be less indicative about being in the bottom. For teammate 2, the distortions could result in either positive or negative asymmetry. In the myopic model (MAB) updating about teammate 2 will be perfectly symmetric. 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

[25] If any of these conditions were violated it would imply that signals are perceived as less indicative of being in the top than they really are. If this were the case then Bayesian updating would in fact give the agent higher utility (see also Appendix A). One potential concern with the structural framework may arise if asymmetry could be generated by mean-zero "errors" in updating. Online Appendix Section 2 presents simulated updating data showing this is not the case.

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

to the following competing hypotheses.

Hypothesis 2:

Updating 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 3:

Updating is self-serving:
$\beta_1^{1,M} > \beta_1^{1,C};\ \beta_0^{1,M} < \beta_0^{1,C}$
Updating about teammate is unbiased (myopic model):
$\beta_1^{2,M} = \beta_1^{2,C};\ \beta_0^{2,M} = \beta_0^{2,C}$
Updating about teammate is biased (general model):
$\beta_1^{2,M} > \beta_1^{2,C};\ \beta_0^{2,M} < \beta_0^{2,C}$ *OR* $\beta_1^{2,M} < \beta_1^{2,C};\ \beta_0^{2,M} > \beta_0^{2,C}$

6 Results

6.1 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.[27] In the Control treatment, where individuals estimated the performance of another, randomly selected individual in the position of teammate 1, the average reported belief was 56.3%. Intriguingly, this is also significantly different from 50% at the 1% level using a Wilcoxon signed rank test. Similarly, the belief that teammate 2 scores in the top half are 53.4% and 54.3% in the Main and Control treatment, respectively. Also these beliefs are significantly different from 50% (Wilcoxon signed rank tests p-values 0.001 and 0.002 respectively). These results hence appear to present evidence for "overconfidence", according to the test of Benoît and Dubra (2011).

Note however that the latter beliefs do not reflect overconfidence in the traditional sense, as they do not involve estimation of one's own performance. As we do not find any evidence of differential assessments across the three other-subject teammate conditions (Kruskal–Wallis test p-value 0.2654)[28], this phenomenon appears to be a general over-estimation, that is not driven by differences in Main or Control, or in teammate 1 or teammate 2 framing. On the other hand, when we test Hypothesis 1 and compare initial beliefs across the two treatments, Main (self) and Control (other), we can clearly reject equality of beliefs (Wilcoxon rank-sum test p-value: 0.0005). This provides robust evidence that what we are observing in the Main

[27] Note that unless clearly stated otherwise, we use two-sided tests.

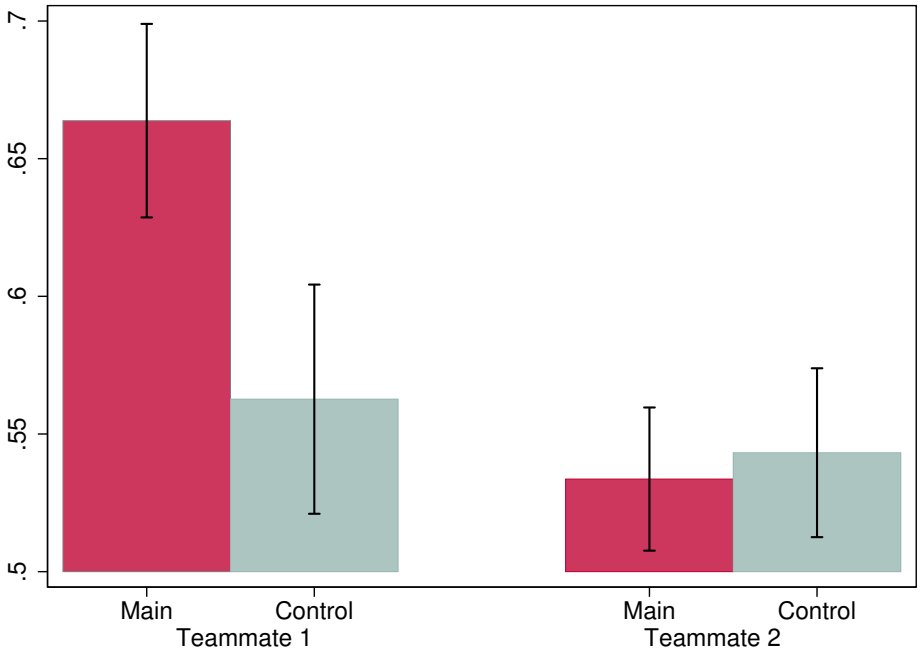
[28] Moreover, we do not find any differences in initial beliefs about teammate 2 between the Main and Control treatment (Wilcoxon rank-sum p-value: 0.5723).

treatment does reflect true overconfidence. It further suggests that subjects do find the IQ task ego-relevant.

Finding 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.*

We also note that our hard-easy manipulation was successful. 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 3). 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 4).

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.

6.2 Belief Updating

To study the self-serving attribution bias we discuss in Section 4, we use the structural model presented in Section 5.2 for our primary empirical analysis. Later, in Section 6.2.2 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 6.2.3. We present an additional analysis of the resulting weights in Appendix C, examine the (evolution of) posterior beliefs in more

detail in Appendix D and substantiate the analysis of individuals' WTP to switch teammate 2 in with additional tests in Appendix E.

6.2.1 Structural Framework

Table 2 presents the main specification for belief updating about teammate 1 for the Main and Control treatments. Our sample includes all updates from both waves, in Part 2 and 3.[29]

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 (positive asymmetry) when updating is about one's own performance. The positive asymmetry observed is significant at the 1% level. No asymmetry is observed in column 2, in the Control treatment, for updating about another individual's performance.

Thus, Hypothesis 2 is rejected, as updating is not the same across the Main and Control treatments. Notably $\beta_1^{1,M} - \beta_0^{1,M} > \beta_1^{1,C} - \beta_0^{1,C}$, indicating that individuals exhibit more

[29]Samples excluding Part 3 are presented in Online Appendix Section 5, 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 5.

(positive) asymmetric updating in Main relative to Control (significant at the 5% level). This seems to be particularly driven by subjects responding less to negative signals in the Main compared to the Control treatment ($\beta_0^{1,M} < \beta_0^{1,C}$). Overall these results are partially consistent with the first part of Hypothesis 3, concerning self-serving attribution bias in own belief updates: namely we find significant differences in response to negative, but not positive, signals.

Finding 2: *When updating beliefs about their own performance, subjects in the Main treatment display a stronger positive asymmetry than subjects from the Control treatment who update about another subject's performance. This asymmetry is driven by under-responsiveness to negative signals.*

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 bias identified in Table 2, we now examine updating about teammate 2. In our general model of attribution bias, updating about teammate 2 may be positively or negatively asymmetric. Either of these biases would impose negative consequences on future decision making involving teammate 2. Alternatively, we also specified that updating could be myopic, i.e. that individuals mis-attribute feedback about their own performance, relegating the difference to noise, but not to their teammate.

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 fact there is evidence of positive asymmetry for teammate 2 in the Main treatment, significant at the 5% level. In particular, one can reject that the coefficient β_0 is the same across the Main and Control treatment at the 5% level. Subjects under-weight negative feedback about their teammate when they are member of the team. The treatment difference in asymmetric updating between Main and Control is not statistically significant, though. Overall these results present even more evidence inconsistent with the hypothesis of equivalent updating across the Main and Control treatments (Hypothesis 2). More specifically, these patterns indicate biased updating about teammate 2, a rejection of the myopic model of attribution bias (found within Hypothesis 3).

Finding 3: *Just like for teammate 1, subjects display positive asymmetry when updating about teammate 2 in the Main, but not in the Control treatment. These patterns are driven by under-responsiveness to negative feedback. The general model of self-serving attribution bias is consistent with these findings.*

In line with the general model, individuals appear to manipulate beliefs about their teammate in the process of generating self-serving beliefs. As noted earlier in Section 4 and detailed in Appendix A, some positive asymmetry about teammate 2 can be optimal since it permits self-serving beliefs, while reducing the financial costs of such beliefs, due to more moderate weighting between the two teammates. Importantly, the positive asymmetry in updating about teammate 2 was predicted to be smaller than the positive asymmetry in updating about one's self. Indeed, for positive signals, β_1 in Table 2 column 1 is significantly greater than β_1 in Table 3 column 1 (Chow test p-value 0.0062). Thus individuals appear to be more positively asymmetric when updating about themselves versus for another subject. We can also note that when comparing the difference in asymmetry $(\beta_1 - \beta_0)$ across the first column in Tables 2 and 3 the difference is significant at the 10% level (Chow test p-value 0.0963).

There are a few candidate alternative explanations for the observation that updating is positively asymmetric for both teammate 1 and teammate 2 in the Main treatment. We discuss three more prominent ones here: first, that anchoring causes individuals to update similarly about teammate 2, second that subjects selectively ignore negative signals overall, and third that that teammate asymmetry is driven by an in-group bias. We find evidence 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, subjects in our Main treatment respond to negative signals at equivalent rates to those in the Control treatment, and third, prior beliefs for teammate 2 are not statistically different across Main and Control. We address these alternative explanations in more detail in Online Appendix Section 6.

6.2.2 Matching on Priors

While the previous evidence showed that beliefs are updated differently in the Main versus Control relative to the Bayesian benchmark, it is also important to examine the extent to which updating differs across the Main and Control without relying on the Bayesian benchmark or a quasi-Bayesian framework. In this subsection we present a non-parametric analysis of updated beliefs, which utilizes a matching strategy that matches 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.[30] By matching on prior beliefs we are able to step away from the reliance on Bayes' rule, and instead ask the following question: given the same priors, 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.[31]

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 frequency* of positive and negative signals. This indicates 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 distribution of signals.

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

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

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

	(1) 0 −	(2) 1 −	(3) 2 −	(4) 3 −	(5) 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 distribution of negative signals received (out of 4 total signals).

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

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 estimated ATE is between 4.7 and 5.2 percentage points greater posterior belief about one's teammate in Main relative to Control, however this is not statistically significant at conventional levels (respective p-values: 0.1294 and 0.1624). Of note is that when examining separately the ATE estimates 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 14 percentage points greater in Main relative to Control. Again, this supports our structural results.

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 distribution of signals.

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

	(1) 0 −	(2) 1 −	(3) 2 −	(4) 3 −	(5) 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 distribution of negative signals received (out of 4 total signals).

Finding 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.*

6.2.3 Willingness to Change Teammates

As highlighted by our earlier motivation, self-serving beliefs can severely bias future decision making. Given the evidence that individuals bias their beliefs about their teammates in order to nurture self-serving beliefs, it is important to examine whether theses biases lead to further consequences in our setup.

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. In Appendix E we additionally investigate treatment differences in the actual value of subjects' WTP (intensive margin). In a nutshell, 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 ($N = 89$). This finding is consistent with the theory, 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.

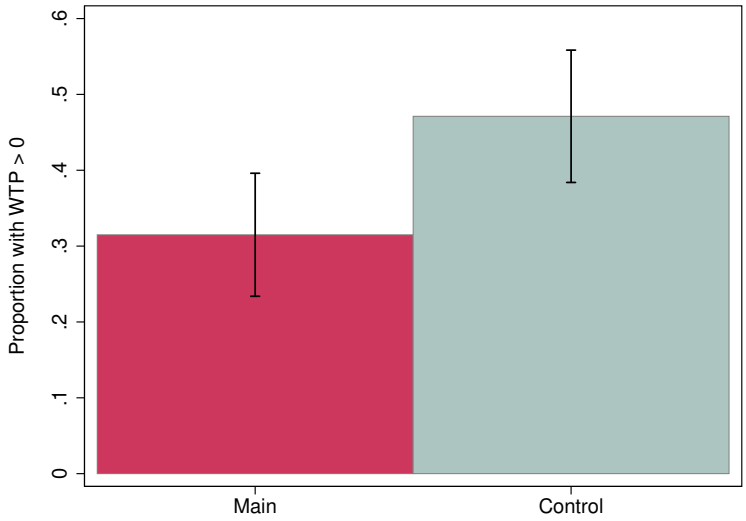
Turning to the extensive margin, given the patterns of biased updating we observe in our Main treatment, subjects end up with more positive performance beliefs about teammate 2. This lowers the proportion of subjects in Main who should be willing to pay to switch teammates, as Appendix E confirms given actual subject beliefs after 4 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 (Ranksum p-value 0.0155).

Finding 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 prior beliefs in the Main compared to the Control treatment. Ex-ante, 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 the performance of teammate 2. Initial beliefs about own performance only affect the value of one's WTP, not whether it is positive or not, see Appendix E.

Finding 5 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.

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.

7 Conclusion

How does overconfidence persist in the face of feedback? Psychologists have proposed and tested theories of self-serving attribution bias, which posit that individuals will be more likely to attribute positive feedback to internal qualities about themselves, and negative feedback to external factors. Yet little is known about these patterns of attribution and their consequences, namely, whether they result in biased beliefs about these external factors.

We addressed this question by examining a micro-founded theory of self-serving attribution bias, and placed it within a quasi-Bayesian updating framework where individuals face two dimensions of uncertainty. In the context of a naturally framed lab experiment with two person teams, we examined how individuals attributed feedback about an IQ test between themselves and their teammate.

We find significant evidence of overconfidence and subsequent biased patterns in belief updating when one is a member of the team, our Main treatment. Individuals update in a positively biased asymmetric way about their own performance. However, their updating about their teammate follows similar patterns, over-weighting positive relative to negative feedback. As a result of these self-serving attributions, individuals end up biased both about own performance but also about their teammate's performance. Notably, in our Control treatment, in which subjects' own performance was not relevant, individuals update symmetrically when receiving both positive and negative feedback.

Our structural results are consistent with additional non-parametric tests. After matching individuals in our Main and Control groups on the value of the prior and the sequence of signals observed, those who are updating beliefs about their own performance end up significantly more confident about their ability than those updating about another person. This effect is strongest for those receiving all negative signals. A similar effect when receiving all negative signals is

that they end up more optimistic about their teammate's performance as well.

Our model of attribution bias is able to account for these findings. As a result of the updating patterns, the submitted weight is more moderate, and the resulting losses from overconfidence are lower than they would otherwise be, absent no bias or negative asymmetry in updating about teammate 2. As such, material losses from overconfidence in the experiment are mitigated. Importantly, our model also provides an explanation for why we observe strong positive asymmetry in updating, while some other studies have not – the possibility of biased updating for another dimension of uncertainty permits even stronger self-serving beliefs, by providing a tool which in certain contexts can mitigate the material costs of overconfidence.

Our results and theoretical insights provide important contributions to our understanding of belief updating with two-dimensional uncertainty. Specifically, in contexts where individuals interact repeatedly in a fixed environment with another source of uncertainty, positive asymmetry about this source could actually mitigate the immediate consequences of overconfidence. Thus, while Heidhues et al. (2018) show negative consequences from self-defeating learning, our results suggest that being additionally biased about the fundamental can lower these costs.

On the other hand, in the real world individuals may have ample opportunities to change their environment. Our results suggest that overconfident individuals may be more likely to stay in lower quality environments, due to positively biased beliefs about the unknown fundamental. This is borne out in our data: individuals in our Main treatment are significantly less likely to demand to change environments. This contrasts with the results of Hestermann and Le Yaouanq (2019), who showed that with Bayesian updating, underconfidence, *not overconfidence* should persist in the long run. Given that overconfidence appears to be a robust phenomenon across a number of different settings, our theory and results can reconcile this puzzle.

More broadly, the fact that most real world informational environments are complex and involve more than one dimension of uncertainty, our findings suggest important new insights about how information is processed in such settings. A key takeaway is that biases which enable self-serving beliefs do not exist in a vacuum, and can lead to distorted perceptions about the broader world.

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, B M and T Odean, "Boys will be boys: Gender, overconfidence, and common stock investment," *Quarterly Journal of Economics*, 2001, *116* (1), 261–292.

Barron, Kai, "Belief updating: Does the 'good-news, bad-news' asymmetry extend to purely financial domains?," *WZB Discussion Paper*, 2017, (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 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.

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.

Daniel, Kent, David Hirshleifer, and Avanidhar Subrahmanyam, "Investor Psychology and Security Market Under- and Overreactions," *The Journal of Finance*, 12 1998, *53* (6), 1839–1885.

Deimen, Inga and Julia Wirtz, "Control, Cost, and Confidence: Explaining Perseverance in the Face of Failure," *SSRN Electronic Journal*, 2019, pp. 1–27.

Dunning, David, *Self-Insight: Roadblocks and Detours on the Path to Knowing Thyself* 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.

Eliaz, Kfir and Ran Spiegler, "Can anticipatory feelings explain anomalous choices of information sources?," *Games and Economic Behavior*, 7 2006, *56* (1), 87–104.

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.

Fischhoff, Baruch, "Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty.," *Journal of Experimental Psychology: Human Perception and Performance*, 1975, *1* (3), 288–299.

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

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.

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

Heider, Frtiz, *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," 2019.

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.

Köszegi, Botond, "Ego Utility, Overconfidence, and Task Choice," *Journal of the European Economic Association*, 6 2006, *4* (4), 673–707.

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.

Malmendier, Ulrike and Geoffrey Tate, "Does Overconfidence Affect Corporate Investment? CEO Overconfidence Measures Revisited," *European Financial Management*, 11 2005, *11* (5), 649–659.

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.

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, M M, Muriel Niederle, Paul Niehaus, and Tanya S 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.

Pekrun, Reinhard and Herbert W. Marsh, "Weiner's attribution theory: Indispensable—but is it immune to crisis?," *Motivation Science*, 2018, *4* (1), 19–20.

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 Joel Van der Weele, "Deception and Self-Deception," 2018.

Silvia, Paul J. and T. Shelley Duval, "Predicting the Interpersonal Targets of Self-Serving Attributions," *Journal of Experimental Social Psychology*, 2001, *37* (4), 333–340.

Svenson, Ola, "Are we all less risky and more skillful than our fellow drivers?," *Acta Psychologica*, 2 1981, *47* (2), 143–148.

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.

_ **and Sandra Graham**, "Attribution in personality psychology.," in "Handbook of personality: Theory and research, 2nd ed.," New York, NY, US: Guilford Press, 1999, pp. 605–628.

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.

Appendix

A Model of Optimal Information Distortion

In this section we provide a micro-foundation for self-serving attribution bias. Specifically we follow Brunnermeier and Parker (2005) by assuming that agents 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.[32] 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 8 and 9.

We introduce the possibility that individuals receive direct utility over the belief that they are in the top half, through a linear function $\alpha \cdot b_{t+1}^1$.[33] $\alpha \in [0, \infty)$ 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, e.g. 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, strictly increasing in $|\gamma_s^i - 1|$, and is minimized at the Bayesian information processing parameter $\gamma_s^i = 1$.[34]

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.[35] 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

[32] Long run models of belief distortion are studied by Heidhues et al. (2018) and Möbius et al. (2014).

[33] 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.

[34] 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. Justification 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, see Coutts (2019b).

[35] 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 now derive the resulting biased posterior beliefs for teammate 1 and 2. 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 $1 - \Phi_{A_1A_2}$.

$$\begin{aligned}\left[\hat{b}_{t+1}^1 | s_t = p\right] &= \frac{\gamma_p^1\gamma_p^2\Phi_{TT}b_t^{TT} + \gamma_p^1\Phi_{TB}b_t^{TB}}{\gamma_p^1\gamma_p^2\Phi_{TT}b_t^{TT} + \gamma_p^1\Phi_{TB}b_t^{TB} + \gamma_p^2\Phi_{BT}b_t^{BT} + \Phi_{BB}b_t^{BB}} \\ \left[\hat{b}_{t+1}^2 | s_t = p\right] &= \frac{\gamma_p^1\gamma_p^2\Phi_{TT}b_t^{TT} + \gamma_p^2\Phi_{BT}b_t^{BT}}{\gamma_p^1\gamma_p^2\Phi_{TT}b_t^{TT} + \gamma_p^1\Phi_{TB}b_t^{TB} + \gamma_p^2\Phi_{BT}b_t^{BT} + \Phi_{BB}b_t^{BB}}.\end{aligned} \tag{14}$$

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. Intuitively 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[36]

$$\hat{\omega}_{t+1}^*=\frac{1}{1+\left(\frac{\gamma_s^2 b_t^{BT}}{\gamma_s^1 b_t^{TB}}\right)^2} \tag{15}$$

[36]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:

$$\max_{\{\gamma_s\}} \Big\{ \alpha \cdot \hat{b}^1_{t+1} + b^{TT}_{t+1} \cdot u(P) + b^{TB}_{t+1} \cdot \sqrt{\hat{\omega}^*_{t+1}} \cdot u(P) + b^{TB}_{t+1} \cdot (1 - \sqrt{\hat{\omega}^*_{t+1}}) \cdot u(0) \tag{16}$$
$$+ b^{BT}_{t+1} \cdot \sqrt{1 - \hat{\omega}^*_{t+1}} \cdot u(P) + b^{BT}_{t+1} \cdot (1 - \sqrt{1 - \hat{\omega}^*_{t+1}}) \cdot u(0) + b^{BB}_{t+1} \cdot u(0)$$
$$- J(\gamma^1_s, 1) - J(\gamma^2_s, 1) \Big\}$$

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.[37] 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 $u(P) - u(0) = \Delta u$):

$$\alpha \cdot \frac{\partial \left[\hat{b}^1_{t+1} | s_t\right]}{\partial \gamma^1_s} + \frac{\gamma^2_s \cdot \left(b^{TB}_{t+1} \cdot b^{BT}_{t+1}\right)^2 \cdot \Delta u}{\left(\left(\gamma^2_s b^{BT}_{t+1}\right)^2 + \left(\gamma^1_s b^{TB}_{t+1}\right)^2\right)^{\frac{3}{2}}} \cdot (\gamma^2_s - \gamma^1_s) - J'(\gamma^1_s, 1) \tag{17}$$

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

$$\alpha \cdot \frac{\partial \left[\hat{b}^1_{t+1} | s_t\right]}{\partial \gamma^2_s} + \frac{\gamma^1_s \cdot \left(b^{TB}_{t+1} \cdot b^{BT}_{t+1}\right)^2 \cdot \Delta u}{\left(\left(\gamma^2_s b^{BT}_{t+1}\right)^2 + \left(\gamma^1_s b^{TB}_{t+1}\right)^2\right)^{\frac{3}{2}}} \cdot (\gamma^1_s - \gamma^2_s) - J'(\gamma^2_s, 1) \tag{18}$$

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

[37] 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 $\gamma_s^2 = \gamma_s^1$.

A.0.1 Example

Here we denote an example taking on a specific functional form to illustrate the properties mentioned above. In particular we assume that $U(P) - u(0) = 10$, and $J(\gamma, 1) = \frac{(1-\gamma)^2}{10}$.

Taking beliefs as $b_t^{A_1A_2} = 0.25$ for all states, the optimal $\gamma_p^1 = 1.494$, while $\gamma_p^2 = 1.344$. Total utility is given by 4.093.[38]

For comparison we consider the analogous myopic context where the subconscious process can only bias γ_p^1 for self (not for the teammate). In this case the optimal $\gamma_p^1 = 1.244$. Total utility in this case is given by 4.063.[39]

In the case of updating in a biased way about both teammates, the subconscious process becomes more biased about own performance, in order to gain from the benefits from overconfidence, whilst managing to lower the material costs.

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 performance 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.

[38] 0.599 utility from overconfident beliefs, 3.531 expected utility from material income, -0.036 dis-utility from mental costs.

[39] 0.554 utility from overconfident beliefs, 3.515 expected utility from material income, -0.006 dis-utility from mental costs.

The expression of interest in the first round is thus:

$$\begin{aligned}\Phi_{TT}(b_0^1 \cdot b_0^2) \cdot \Phi_{BB}((1-b_0^1) \cdot (1-b_0^2)) - \Phi_{TB}(b_0^1 \cdot (1-b_0^2)) \cdot \Phi_{BT}((1-b_0^1) \cdot b_0^2) \\ = (b_0^1 \cdot b_0^2)((1-b_0^1) \cdot (1-b_0^2)) \cdot [\Phi_{TT} \cdot \Phi_{BB} - \Phi_{TB} \cdot \Phi_{BT}] \qquad (19)\end{aligned}$$

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 19:

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

We now note that this is simply an iteration of Equation 19. 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:

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

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 Chosen Weights

While the primary focus of the empirical analysis is on determinants of beliefs and belief updating, it is informative to investigate how beliefs and updating affect subject's weighting decisions. Recall that individuals had to choose a weight from 0 to 1, with 0 representing all of the weight on teammate 2, and 1 representing all of the weight on teammate 1. Here we evaluate optimal weights relative to the Bayesian prediction: the weight chosen should be invariant to feedback. While feedback will impact beliefs, it does so proportionately for both teammates, leaving the weight unchanged. That is, after controlling for the initial weight, neither positive nor negative feedback should alter the submitted weight.

Table C.1 shows regressions which examine impacts of subject characteristics and the main treatment on weighting decisions. The Bayesian prediction is that the initial weight in round one should have a coefficient of one, and all other coefficients should be zero. From the table one can see that this is not the case. While the initial weight is positive and significant, it is less than one. What is more interesting is that against the Bayesian predictions, positive feedback has a statistically significant effect on the weight chosen, in columns (1) and (2).

Additionally, there is some evidence that being a member of the team, i.e. our Main treatment, has a statistically significant effect on the chosen weight.

Yet, as columns (3) and (4) show, the positive effect of both a positive signal and the Main treatment are coming from the interaction between the two. In particular, this interaction increases the weight by 6.4 percentage points. This is about an 11% increase on the average weight chosen. Thus, when individuals are part of the team, when receiving a positive signal they increase the weight on their own performance by 6.4 percentage points, despite the Bayesian benchmark being to not alter the weight.

The result that there is some limited evidence of a larger weight after positive signals is consistent with the results on asymmetric updating. Since subjects were also positively biased in updating about their teammate, this creates an overall moderating effect: the positive bias for both teammates works to cancel out, producing a more moderate weight report. A slight effect for positive signals is consistent with the slight over-weighting of positive signals for self relative to teammate 2, while for negative signals there was no significant difference in the structural framework. In the Control treatment the responsiveness to feedback was balanced across both teammate 1 and 2, and for both positive and negative feedback. This is consistent with the results in Table C.1.[40]

[40]Finally, 7% of observations involved the submission of a different weight than what was recommended by z-tree. The average difference from the optimal recommended by z-tree was 0.056 (recalling that $\omega \in [0, 1]$). However there are no systematic differences in submitting a different weight behavior by treatment.

Table C.1: Submitted Weight on teammate 1

	(1)	(2)	(3)	(4)
Initial Weight	0.600***	0.515***	0.518***	0.473***
	(0.033)	(0.042)	(0.042)	(0.044)
+ Signal	5.435***	5.364***	2.367	0.521
	(1.458)	(1.435)	(2.136)	(2.081)
Main Treatment	3.540	4.267*	1.210	0.854
	(2.320)	(2.273)	(2.843)	(2.726)
+ Signal × Main Treatment			5.982**	6.435**
			(2.833)	(2.738)
Female	2.767	2.223	2.480	2.244
	(2.271)	(2.194)	(2.179)	(2.143)
Age	−0.387	−0.409*	−0.410*	−0.381
	(0.237)	(0.241)	(0.241)	(0.236)
# Attempted by teammate 1		2.976***	2.965***	1.675**
		(0.626)	(0.626)	(0.687)
# Attempted by teammate 2		−1.272**	−1.276**	−1.675***
		(0.558)	(0.555)	(0.528)
Score of teammate 1 on IQ Test				0.608***
				(0.158)
Round Fixed Effects	✓	✓	✓	✓
R^2	0.38	0.40	0.40	0.42
Observations	2595	2595	2595	2595

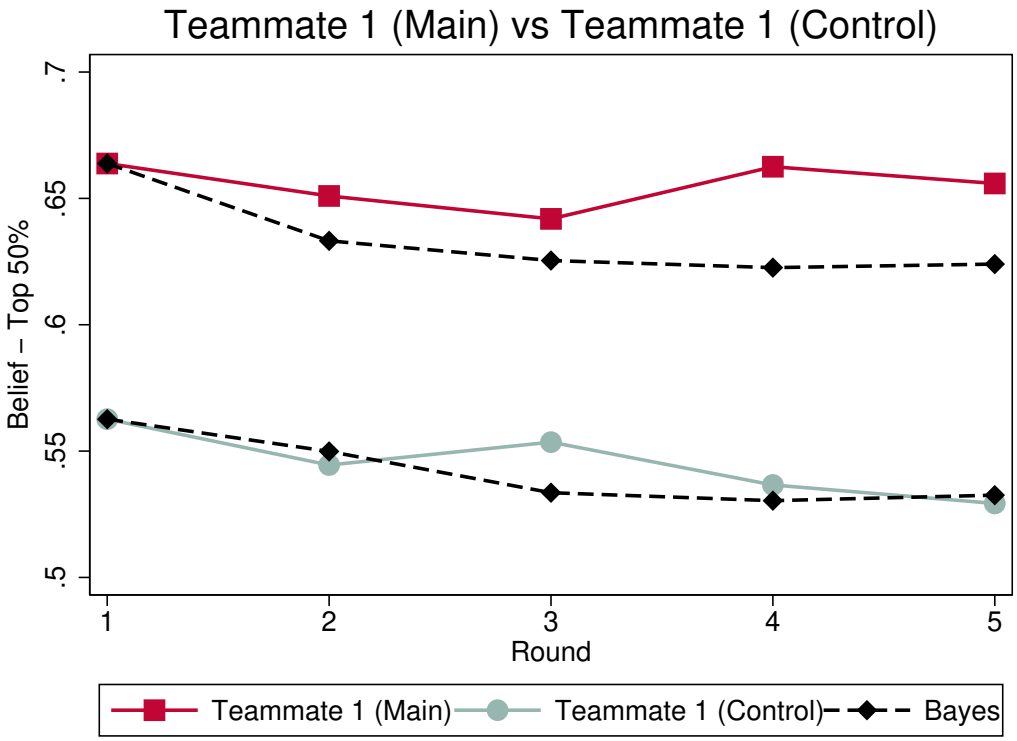
Analysis uses OLS regression. Difference is significant from 0 at * 0.1; ** 0.05; *** 0.01. Robust standard errors clustered at individual level.

D Examining Posterior Beliefs

Figures D.1 and D.2 examine the evolution of beliefs in response to feedback for teammate 1 and 2 respectively, starting from the first prior, before receiving any feedback. While posterior beliefs about one's self (Main, teammate 1) are significantly greater than beliefs about teammate 1 in the Control, this is in large part driven by differences in prior beliefs due to overconfidence. In both figures one can see a pattern that posterior beliefs in the final round deviate further from the Bayesian prediction in Main compared to Control, both for teammate 1 and 2.

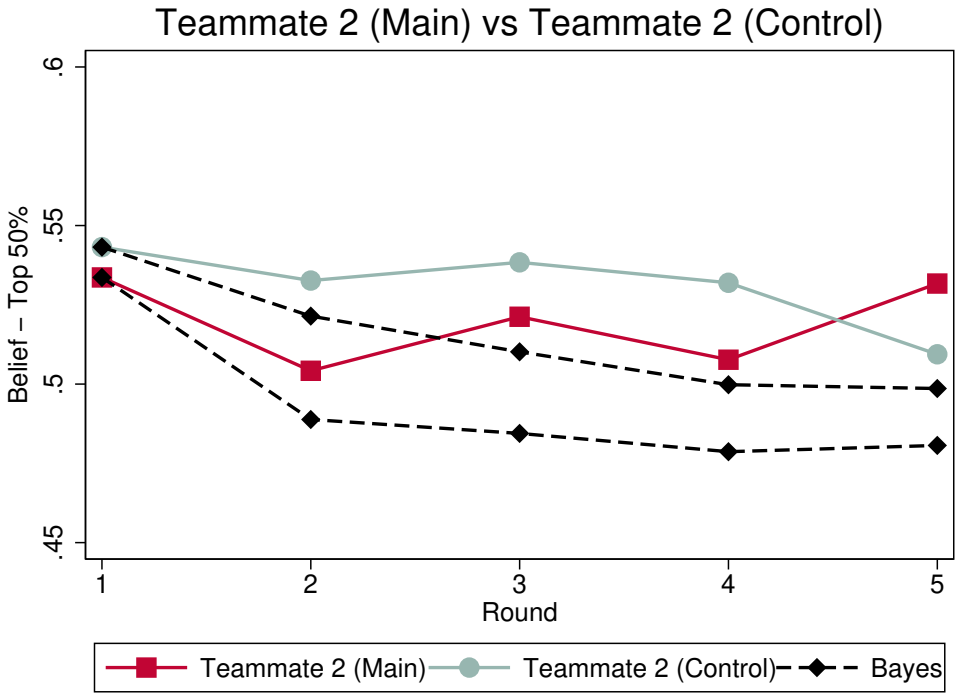
Figure D.3 examines this more closely, presenting the difference between reported posteriors and the Bayesian prediction given subjects' initial priors, after four rounds of feedback. This corresponds to round 5 in the two figures above. While this does present evidence that positive deviations are more pronounced in the Main treatment, we also note that the difference between the deviations in Main and Control are not significantly different at conventional levels.

Figure D.1: Evolution of Beliefs: teammate 1



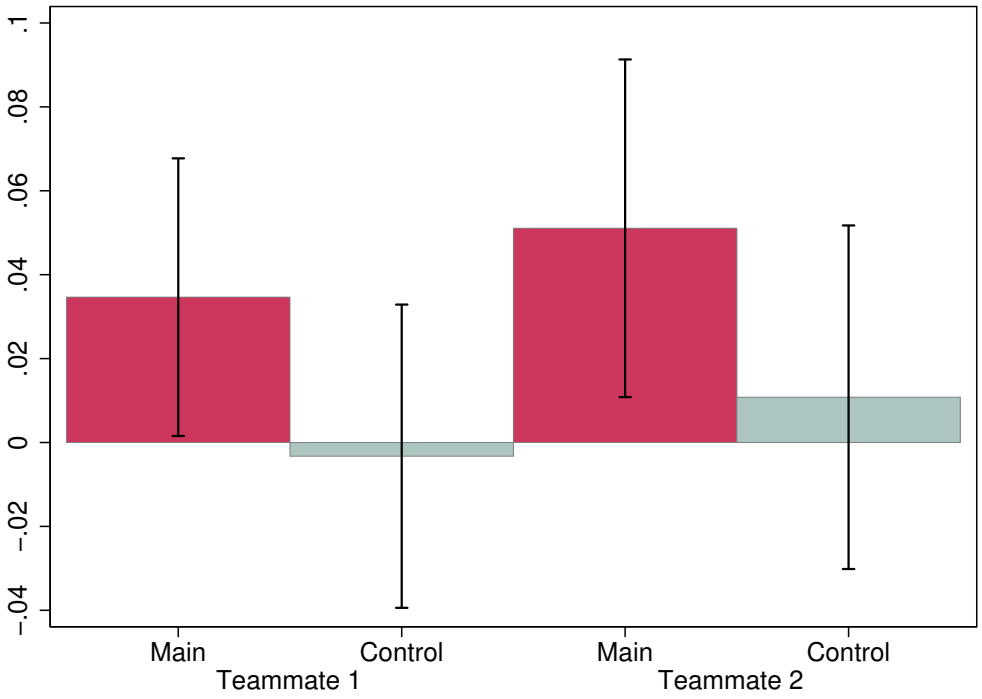
Evolution of beliefs about teammate 1 starting from prior beliefs with 4 round of feedback. Bayesian benchmark is calculated from subject's first prior, then evolves given actual signals observed. Standard error bars omitted for clarify (error bars are always overlapping with bayesian predictions).

Figure D.2: Evolution of Beliefs: teammate 2



Evolution of beliefs about teammate 2 starting from prior beliefs with 4 round of feedback. Bayesian benchmark is calculated from subject's first prior, then evolves given actual signals observed. Standard error bars omitted for clarify (error bars are always overlapping with bayesian predictions).

Figure D.3: Raw Deviation of Posterior Beliefs from Bayesian Benchmark



Plot of the difference between Posterior beliefs and Bayesian beliefs after 4 rounds of feedback. Bayesian beliefs are calculated using subject priors before any feedback.

E 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.[41] 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 change teammates, assuming risk neutrality. Before receiving feedback, this decision depends entirely on the belief about teammate 2. If a subject believes their teammate is in the top half with probability less than 50% they should pay to change, otherwise they should not be willing to pay any positive amount.[42]

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 change teammates will be the same across both groups. Figure E.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 change teammates.

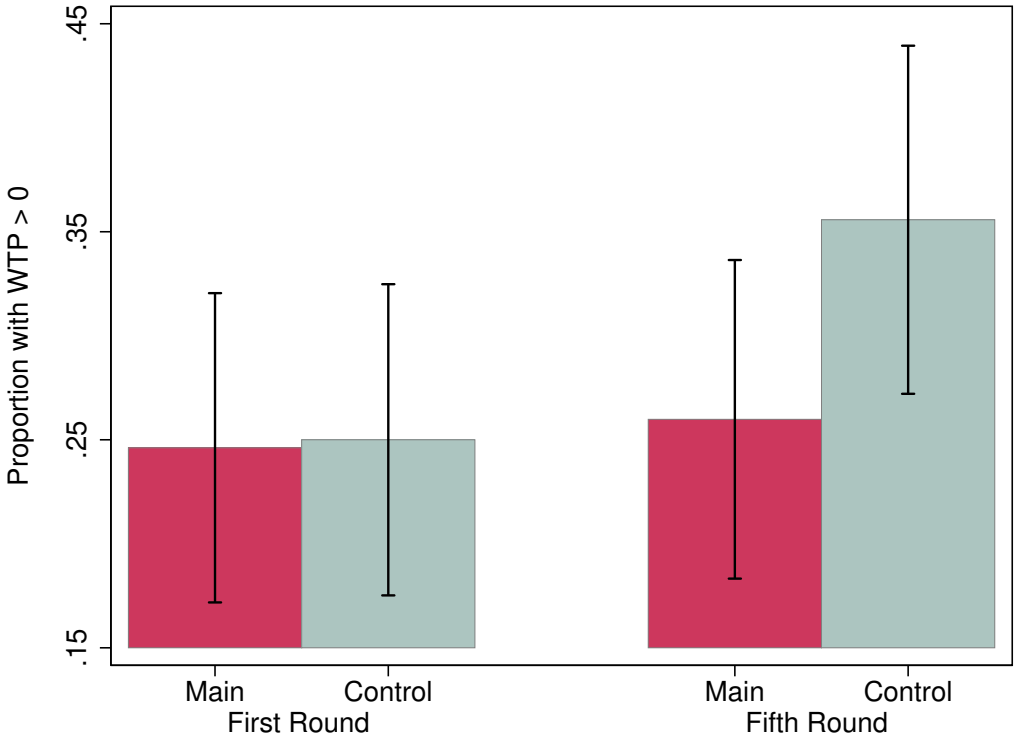
While prior beliefs are such that there are no differences across Main and Control treatments, beliefs after 4 rounds of feedback (Round 5) are such that in fact a higher proportion of individuals in Control should be willing to switch teammates. This is because in Control, subjects update in a symmetric way about their teammate, and end up with more moderate beliefs.[43] In Main, because of the asymmetry 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.

[41] 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.

[42] 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.

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

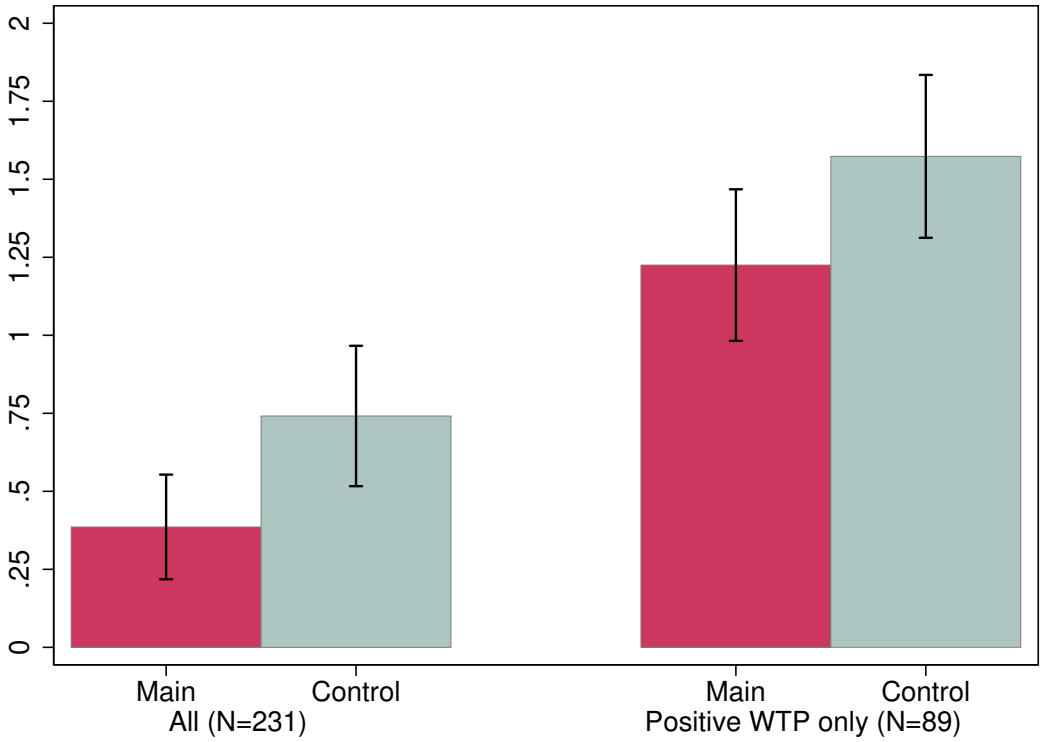
Figure E.1: (Calculated) Optimal Proportion Willing to Switch



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

Figure E.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 (Ranksum p-value 0.006). Restricting the sample only to positive WTP, the Ranksum 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 E.2: Willingness to pay



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