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

Alexander Coutts Leonie Gerhards Zahra Murad

December 2023

Abstract

People often receive feedback influenced by external factors, yet little is known about how this affects self-serving biases. Our theoretical model explores how multi-dimensional uncertainty allows additional degrees of freedom for self-serving bias. In our Primary experiment, feedback combining an *individual's ability* and a *teammate's ability* leads to biased belief updating. However, in a Follow-up with a random fundamental replacing the teammate, unbiased updating occurs. A Validation experiment shows belief distortion is greater when outcomes originate from human actions. Overall, our experiments highlight how multi-dimensional environments can enable self-serving biases.

Keywords: Motivated beliefs; multi-dimensional; belief updating; overconfidence
JEL Codes: D9, C91, D81, D83, D84

*Coutts: Schulich School of Business, York University, 4700 Keele Street, Toronto, Ontario, Canada (email: acoutts@schulich.yorku.ca); Gerhards: King's Business School, King's College London, Bush House, London, WC2B 4BG, United Kingdom (email: leonie.gerhards@kcl.ac.uk); Murad: Accounting, Economics and Finance, The University of Portsmouth, Portsmouth, PO1 2UP, United Kingdom & UNEC Cognitive Economics Center, Azerbaijan State Economics University, Baku (email: zahra.murad@port.ac.uk).

†We are very grateful for useful comments from Kai Barron, Thomas Buser, Tingting Ding, Boon Han Koh, Yves Le Yaouanq, Robin Lumsdaine, Cesar Mantilla, Luis Santos Pinto, Giorgia Romagnoli, Adam Sanjurjo, Marcello Sartarelli, Peter Schwardmann, Sebastian Schweighofer-Kodritsch, Séverine Toussaert, Joël van der Weele, and Georg Weizsäcker, as well as helpful comments from seminar and conference participants at University of Alicante, University of Amsterdam, Bayesian Crowd Conference, briq Workshop on Beliefs, CEA Banff, ECBE San Diego, ESA Berlin, HEC Lausanne, IMEBESS Utrecht, King's College London, Lisbon Game Theory Meetings, LMU Munich, M-BEES, NASMES Seattle, NYU CESS, NYU Shanghai, University of Portsmouth, RWTH Aachen, Schulich School of Business, SHUFE, THEEM, TRIBE Copenhagen, and WZB. We thank two anonymous reviewers for helpful comments on earlier drafts of the manuscript. We gratefully acknowledge financial support from the Hamburgische Wissenschaftliche Stiftung, the Genderförderfonds of the University of Hamburg, the Graduate School of Economics and Social Sciences of the University of Hamburg and the Research Project Fund at the University of Portsmouth. Ethical approval for the lab experiments was granted by the Faculty of Business, Economics and Social Sciences at the University of Hamburg (29 June 2017); ethical approval for the online experiment was awarded by King's College London (reference number: MRA-21/22-33480, 22 July 2022).

1 Introduction

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

The existing approach to understanding the formation of self-serving beliefs has been to focus on one dimension of relevance that an individual cares about (e.g., ability), and study the trade-offs that lead to distortion of that specific dimension. For example, previous work has focused on the material costs of holding biased beliefs (Brunnermeier and Parker, 2005), as well as cognitive constraints to self-deception (Bénabou and Tirole, 2002; Bracha and Brown, 2012; Engelmann et al., 2019). Yet, in many real world settings, information comes bundled with other sources of uncertainty, such as a teammate's ability or market fundamentals. Following the existing approach, in these rich environments with multi-dimensional uncertainty, individuals would process information about the dimension of relevance in a self-serving way, but would otherwise update their beliefs using Bayes' rule for any other dimensions of uncertainty (Heidhues et al., 2018; Hestermann and Le Yaouanq, 2021).

In this paper we move beyond this one-dimensional paradigm, and allow for the possibility that individuals can manipulate other features of their environment to arrive at self-serving beliefs. In our theoretical framework, belief-updating about each dimension of uncertainty can be distorted, subject to a context-dependent cognitive cost, consistent with evidence suggesting that belief distortion varies across environments (Engelmann et al., 2019).[3] Optimally distorted beliefs then arise as the result of the trade-off between the hedonic benefits from motivated beliefs, against the material and cognitive costs. The key insight is that the ability to distort other dimensions presents an additional degree of freedom which can enable greater levels of self-serving beliefs.

Consider an example which mirrors the environment we study in this paper, where an individual receives feedback that depends on their own performance and a teammate's performance.

[1] Specifically, benefits may arise from: (i) direct utility from holding overconfident beliefs for example arising from self-esteem or ego-protection (Möbius et al., 2022; Brunnermeier and Parker, 2005), (ii) benefits to personal motivation or self-signalling (Bénabou and Tirole, 2002, 2009, 2011), or (iii) strategic signalling motives and persuasion of others (Burks et al., 2013; Schwardmann and van der Weele, 2019; Schwardmann et al., 2022). These three explanations have long been a part of the core motivation for attribution theory of social psychology, corresponding to (i) self-enhancement/protection (ii) belief in effective control, and (iii) positive presentation of self to others; see Kelley and Michela (1980) and Tetlock and Levi (1982).

[2] While we focus on biases in information processing, there is evidence for other self-serving strategies such as avoiding negative information (e.g., for health (Oster et al., 2013); see Golman et al. (2017) for a broader review), or biased recall (Zimmermann (2020)).

[3] Engelmann et al. (2019) find that more ambiguous environments permit greater belief distortions in the context of wishful thinking regarding future electric shocks. More broadly, cognitive costs of belief distortion would be expected to vary based on how costly it is to employ mental strategies to justify desired beliefs in different contexts (Bracha and Brown, 2012; Kunda, 1990).

Under the existing one-dimensional approach, individuals would only be assumed to process information in a biased way regarding their own performance – they would not strategically distort beliefs about their teammate. In contrast, we assume that individuals have the additional degree of freedom to manipulate their information processing about the teammate's performance. Intuitively, having two levers instead of one expands the potential for nurturing greater levels of self-serving beliefs.

While our theoretical framework shows how multi-dimensional uncertainty can facilitate self-serving bias, it also highlights the importance of context to understand whether and how belief distortions will appear. The first insight pertains to the *direction of distortions* about other dimensions. Specifically, we show that context-specific costs and benefits, i.e. material incentives, can influence whether beliefs for these other dimensions are distorted in a positive or negative direction. This finding hints at an underlying complexity that may alter common interpretations of self-serving attribution bias (Hastorf et al., 1970). Typically, the conventional view is that individuals tend to attribute negative feedback to external factors while taking credit for positive feedback.[4] These negative distortions can be beneficial as they amplify self-serving beliefs. However, our framework emphasises an important point: depending on context-specific incentives, optimal distortions about the other dimensions can also be positive.

The second insight is that the nature of the additional dimension of uncertainty will matter for the *magnitude of distortions*. If individuals find it more costly to distort beliefs about some dimensions, then we should expect a lower extent of self-serving distortions in these contexts. To better understand self-serving belief distortion with multi-dimensional uncertainty, and the importance of context, we conducted a series of experiments. In what follows, we refer to these experiments as Primary, Follow-up, and Validation.

In the Primary lab experiment, participants take an IQ-style test, and are anonymously paired with a teammate who took the same test. The team's output depends on the performance of both teammates. Participants receive noisy aggregate feedback, and can attribute the feedback to both their own and the teammate's performances. The updating problem is then one of joint inference; however, the feedback from these two sources cannot be disentangled. Based on the reported beliefs about the two performances, the computer automatically calculates and recommends a weight which optimally balances their teammate's performance (lower weight) and their own performance (higher weight). Accurate beliefs are incentivised because

[4]The study of self-serving attribution biases within psychology has naturally focused on environments with multi-dimensional uncertainty. While the overall evidence suggests significant evidence in favour of the existence of self-serving attribution biases (Mezulis et al., 2004), the resulting studies of attribution were focused on general principles rather than tractable models, discussed in Kelley (1973) and Weiner (2010). Moreover, the study of self-serving biases in psychology is often framed as one of trade-offs for managing blame in order to maintain desirable beliefs (Campbell and Sedikides, 1999). Outside of the self-serving realm, attribution biases have been studied more generally, such as examining whether attributions are biased towards more salient sources such as other individuals (Heider, 1944, 1958; Pryor and Kriss, 1977; Lassiter et al., 2002), with parallels to availability bias (Tversky and Kahneman, 1973); other work has focused on mis-attribution due to mood (Schwarz and Clore, 1983). Relatedly, outcome bias occurs when the randomness or luck inherent to outcomes disproportionately influence judgements of decision quality (Baron and Hershey, 1988). With multi-dimensional uncertainty, while the distinction between different sources of uncertainty is central for the predictions of attribution bias, it does not appear relevant for the theory of outcome bias. Brownback and Kuhn (2019) study outcome bias in the context of a principal-agent framework, and provide a useful review of this literature within economics.

this resulting weight determines their material payoffs in the experiment.

In spite of these incentives for accuracy, relative to a control in which we remove ego-relevance, we find that individuals distort beliefs not only about themselves, but also about their teammate. Own belief distortions are self-serving, as expected, which in isolation, would imply an upward biased weight. However, we find that individuals update their beliefs in a positively-biased way towards the teammate. As suggested by the theoretical framework, the motives for such a positive bias towards the teammate can be identified through the incentives in the experiment – they counteract the immediate negative material consequences of self-serving beliefs by lowering the weight. However, these positive distortions towards the teammate do have subsequent consequences. In the experiment we find that individuals are significantly less likely to change teammates compared to a control, when given a surprise opportunity.[5]

Beyond showcasing the role for strategic belief distortion as a tool to mitigate financial consequences and enable self-serving beliefs, our theory highlights the critical role of the uncertainty source itself. Even under the same material incentives, the nature of the source of uncertainty can matter as well. Context-specific cognitive costs – illustrated by factors like the ease of distorting beliefs about a teammate – can influence the magnitude of distortions, thereby underscoring their importance in shaping self-serving biases.

In our Follow-up lab experiment, we explored precisely this theme, by repeating core elements of the Primary experiment but replace the human teammate with a random fundamental source of uncertainty. Our results from this experiment are notably distinct: we observe no systematic bias in belief updating, neither for individuals' beliefs about own abilities nor for their beliefs about the non-human teammates. In terms of our theoretical framework, this suggests that cognitive costs are greater in the non-human Follow-up. Under this rationale, the limited distortion about this non-human teammate precludes its use as an additional degree of freedom, thereby limiting own belief distortion.

We explored this conjecture with an additional online Validation experiment which examined the extent of belief distortion across two between-subjects treatments: a pairing with either a (i) human, or a (ii) random fundamental. Our findings reveal a significantly higher likelihood of distortion in the human treatment. These results corroborate the presumption that, holding material incentives fixed, cognitive costs of distorting beliefs can vary across different contexts. Overall, this series of findings allows for a new understanding of how self-serving beliefs may be nurtured in environments with multi-dimensional uncertainty. They suggest that some environments, such as those involving other people, may be more amenable to belief distortions about others, which in turn provide the additional degrees of freedom that can be used to enable self-serving beliefs.

The remainder of the paper unfolds as follows. In the upcoming section, we provide a comprehensive overview of our experimental context and the design of the Primary experi-

[5] While our Primary experiment showcases how self-serving biases can be enabled through positive distortions in an environment with multi-dimensional uncertainty, our theory makes the broader point that there exist other environments where negative distortions would be beneficial. For instance, an overconfident investor might attribute poor performance to an unfavourable market fundamental and consequently exercise more caution.

ment. Following that, we delve into our theoretical framework, which focuses on self-serving attributions with an additional source of uncertainty. Subsequently, we present our predictions, followed by the results, including those from the Follow-up and Validation experiments. Finally, we conclude with a comprehensive discussion.

2 Experimental Design

2.1 Overview

The Primary experiment sessions were conducted in-person at the WiSo experimental laboratory at the University of Hamburg, using z-tree (Fischbacher, 2007). A total of 426 student participants participated in 17 sessions, across two waves in the 2017-18 academic year. The main distinction between these waves was that in the second wave we included an additional part in which individuals could switch teammates.[6] The Primary experiment comprised of Main and Control treatments. In the Main treatment (226 participants), participants were paired in teams of two. The Control treatment (200 participants) served as a benchmark where participants, positioned as a third-party, made equivalent decisions about the performance of another two-person team. Table 1 summarises the structure of the experiment, full experimental instructions are presented in Online Appendix Section 10. We defer discussion of the Follow-up and Validation experiments to Section 6.

2.1.1 Main Treatment

We first describe the Main treatment, with the Control treatment presented in the following subsection. At the beginning of the experiment we provided participants 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 participants had 10 minutes to complete a trivia and logic test consisting of 15 questions. The instructions stated: "Questions similar to these are often used to measure a person's general intelligence (IQ). Your task is to answer as many of these questions correctly as possible." Our priority was to emphasise the importance of the test to participants, so that they would care about their ranking. Our intention was not to actually measure their IQ. Participants were assigned either a hard or easy version of the test, randomised at the session level.[7]

[6] Wave 1 had 192 participants, while wave 2 had 234; recruitment aimed for gender balance, 52% reported to be women. Experimental sessions in the first wave lasted approximately 1 hour, in which participants received an average payment of €14. The second wave was for the most part identical to the first but, in addition to the option of changing teammates, had a slight difference in the belief elicitation as detailed in Online Appendix Section 10. Experimental sessions in wave 2 lasted approximately 1.5 hours in which participants earned on average €19. 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 and the final questionnaire. Due to a small glitch, some participants inadvertently skipped entering beliefs, which leaves us with 3155 out of 3170 observations.

[7] The motive for including two test versions was to explore whether, independently of our theoretical model, there would be a hard-easy effect (Larrick et al., 2007; Moore and Small, 2007) in information processing. We do not find this to be the case, see also Section 5.1.

Table 1: Experimental Flow

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

Each correct answer would earn 2.5 points while an incorrect answer would be penalised by 1 point. Unanswered questions did not affect the final score. These incentives ensured that the attempted number of questions (which we use in later parts of the experiment) would carry some informational value.[8] Participants could not score below zero and were paid €0.10 per point earned in Part 1 at the very end of the experiment. At this stage no feedback on performance was given.

At the beginning of Part 2, participants were paired into teams of two that remained constant throughout this part. Participants' individual performances on the test from Part 1 jointly defined their "team performance" in Part 2. We neither provided participants with any information about their teammates' identity nor about their teammates' actual test scores. Participants 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,

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

generating variation in initial prior beliefs.

We designed the team formation protocol such that both teammates' test scores were compared to the same randomly selected group of 19 other test scores from the experimental session. Each participant 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 participants believe that they and their teammate score in the top half of performances. Participants neither learned their absolute score nor whether they themselves or their teammate belonged to the top or bottom half until the end of the experiment. Not comparing teammates' scores to each other, but to the same comparison group, ensured that the teammates' individual rankings were independent of the other's score.

2.1.2 Control Treatment

It was also critical for us to conduct a fully powered comparison group as a control. To this end, randomised across sessions, we varied whether participants 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 participants' beliefs and subsequent earnings depended on participants' own performance, while in the Control treatment own test performance was not relevant.

In Control, at the beginning of Part 2 each participant was assigned to a team consisting of two randomly selected other participants (the teammates) from the same session. Participants 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 participants in the Control treatment had nearly identical information about all decision-relevant variables as the participants in the Main treatment. As a result, by comparing reported beliefs across the Main and Control treatments, we can better isolate biases driven by reasons of ego-protection and abstract from other sources of belief updating biases.[9]

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.

[9]While access to information about the questions and answers is the same, we note that participants could have private information about their ability which affect their judgement. We also do not wish to claim that beliefs about performance will be the same. For example, participants may view their own attempted answers as likely to be correct, more-so than when viewing others' attempted answers. On aggregate we believe such patterns would be consistent with overconfident beliefs. We thank two anonymous referees for raising these issues.

2.2 Weighting Decision and Belief Elicitation

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

The weighting decision and its direct relationship with earnings provided participants with a transparent monetary incentive to truthfully report their beliefs about the probabilities of the two teammates scoring in the top half of performances on the IQ task. Based on participants' reported beliefs, the computer then calculated the optimal weight and recommended how much to weight teammate 1's performance relative to teammate 2's performance, using graphical tools and an explanation of which weight would give them the highest expected payoffs (see Figure 1). Thus, the weights are aimed at providing a natural framing to a team decision making context to elicit beliefs in an incentive compatible way. We will not focus on the analysis of the weights in the main text of the paper and delegate it to the Online Appendix Section 4, since the weights are a secondary measure, and less informative than beliefs.

Assuming participants can form subjective beliefs, as long as they strictly prefer a higher probability of earning €10, it is in their best interest to truthfully report those beliefs. This procedure is thus novel in its indirect implementation, but shares similar incentive compatibility properties of other elicitation procedures such as matching probabilities (Holt and Smith, 2009; Karni, 2009), or the binarised scoring rule (Hossain and Okui, 2013).[10] 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).

Participants 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 participants at this point of the experiment), the participant would earn an amount of €10 for sure. Analogously, if both of the teammates were ranked in the bottom half, the participant would earn an amount of €0 for sure. If, however, one teammate was ranked in the top half and the other was ranked in the bottom half, a participant'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. These payoffs can be linked to many contexts, e.g., allocating work among team members of potentially different abilities.

For each elicitation, participants entered beliefs for the probability that teammate 1 scored

[10] This presumes that individuals follow the recommended weight. Initially we surmised that some individuals might prefer a biased weighting decision (akin to a type of "illusion of control" bias), and as a result we chose to give participants the flexibility to override the recommendation. Reassuringly, only 7% of weights did not correspond to the recommended 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 participants can arrive at the optimal weight, but intentionally report different combinations of beliefs to deceive the experimenter. We do not find this likely.

in the top half, and the probability that teammate 2 scored in the top half. Intuitively, higher probabilities assigned to teammate 1 (teammate 2) would increase (decrease) the calculated weight, and participants were able to move backwards if they preferred to alter beliefs. Regarding the optimal weight calculation, this requires knowledge of the probabilities of the two payoff relevant states: whether teammate 1 ranks in the top half and teammate 2 in the bottom half, and vice-versa, see Section 3.2. In wave 1 we derived these probabilities, assuming independence between beliefs about the teammates scoring in the top half. In wave 2 we further elicited beliefs for all four possible states as shown in Figure 1. More detail about these procedures and their potential impact can be found in Online Appendix Section 1.

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

2.3 Feedback

Once their weight was submitted, participants 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% probability and a Green Check with 10% probability.

Note that the feedback received from the Team Evaluator was (i) derived from the actual performance of the teammates in Part 1, (ii) independent across feedback rounds, and (iii) depended neither on the beliefs reported by participants nor on the previous weights submitted. This ensured that participants did not have incentives to "experiment" with their chosen beliefs and weights to learn more about their rankings. It also precludes self-defeating learning as, for instance, studied by Heidhues et al. (2018).

After receiving the Team Evaluator's feedback, participants 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, participants 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, participants 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.[11] Before the start of Part 2, participants 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. Participants were only allowed to start Part 2 of the experiment and enter their first belief when the experimenter had checked that the answers provided were correct.

2.4 Part 3: Willingness to Pay to Change Teammate 2

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

Part 3 otherwise was identical to Part 2. We elicited WTP using the BDM mechanism of Becker et al. (1964). The mechanism asked participants 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 participants 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.

[11] 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 participants are not aware of its structure until completing Part 2.

3 Theoretical Framework

3.1 Preliminaries

We first setup our framework which follows from the experimental design. An individual faces an environment with two sources of uncertainty: (i) the ability of teammate 1 (own ability in Main) and (ii) the ability of teammate 2 (though we use the term teammate 2, note that this can refer to any source of uncertainty). 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). Similarly, Teammate 2's unknown ability is given by $A_2 \in \{B, T\}$, which corresponds to whether teammate 2 is in the bottom half or top half of performances. 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 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 optimise by considering the payoffs of each period, which are determined according to the lottery $(P, 0; \sqrt{\omega_t})$ that pays P with probability $\sqrt{\omega_t}$ and 0 otherwise.

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

3.2 Optimal Weight

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

The optimisation problem of individuals is to maximise expected utility:

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

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

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

This leads to the optimal weight,

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

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

Two observations are worth noting. First, given the functional form of expected utility, the optimum in Equation 4 is guaranteed to exist, and there is a unique solution for any beliefs except for the extreme case when $b_t^{TB} = b_t^{BT} = 0$.[12] Second, the optimal weight depends in opposite directions on the expected ability of teammate 1 and the expected ability of teammate 2. Thus, biases in beliefs regarding teammate 1 and 2 will be most costly when they are in opposing directions, for example, an upward bias for teammate 1 and a downward bias for teammate 2.[13]

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

[13]In period 0, this functional form initially generates the same self-defeating learning condition discussed in Heidhues et al. (2018). However, as previously noted, the feedback received by our participants is not influenced by their weighting decisions, which prevents the occurrence of the self-defeating learning 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.

3.3 Belief Updating

We first examine the Bayesian benchmark to study how beliefs evolve for the four states, and hence how beliefs about being in the top half evolve. Following the experiment, signals are independent across time t and not perfectly informative about the states of the world (i.e. noisy). They are positive (p) with probability $\Phi_{A_1A_2}$, otherwise they are negative (n). We denote them by $s_t = (p, n; \Phi_{A_1A_2})$. From now on we also make explicit the assumption that $1 > \Phi_{TT} > \Phi_{TB} = \Phi_{BT} > \Phi_{BB} = 1 - \Phi_{TT} > 0$, in our experiment specifically $\Phi_{TT} = 0.9, \Phi_{TB} = \Phi_{BT} = 0.5, \Phi_{BB} = 0.1$.

A Bayesian will update beliefs about teammate 1 being in the top half given either positive (p) or negative (n) signals respectively as follows:[14]

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

Analogously for teammate 2:

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

3.4 Self-Serving Attribution Bias

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

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

Following the literature and our discussion in the introduction, we assume that belief distortion is costly for two reasons: first, the material consequences which result from subsequent

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

worse decision making, and second, the presence of mental or cognitive costs of distorting beliefs. As is typical in these models (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. In this section we present a model of modified Bayesian updating which moves beyond the existing literature. Specifically, in our model, updating is not constrained to a biased interpretation of just one dimension of uncertainty, but allows for flexible attribution across these multiple dimensions of uncertainty to arrive at optimal self-serving beliefs.

Here we present a brief overview; the model's foundations are derived in Appendix A. Individuals derive utility from beliefs about their ability. To reap these benefits from overconfidence, individuals update according to a variation of Bayesian updating that is optimally distorted across *two* dimensions. First, the perceived likelihood of signals being generated by $A_1 = T$ (i.e. teammate 1 being in the top half) is distorted by a term γ_s^1 (s can refer to either (p)ositive or (n)egative signals). Second, and analogously, the perceived likelihood of signals being generated by $A_2 = T$ is distorted by a term γ_s^2. While Bayes' rule corresponds to $\gamma_s^i = 1$, larger values increase the perception that the relevant state generated a particular signal, with the opposite for smaller values. Hence $\gamma_s^1 > 1$ would lead an individual to believe a signal s is more likely to occur when the state is $A_1 = T$, while $\gamma_s^2 < 1$ would lead them to believe the signal s is more likely when the state is $A_2 = B$. Each dimension of distortion entails its own cognitive cost, i.e. how difficult it is for individuals to distort their information processing about that dimension – contrary to the underlying reality (Bracha and Brown, 2012). The resulting optimal distortions across the two dimensions trades off the benefits from overconfident beliefs against these cognitive costs, as well as against the material consequences from holding (multi-dimensional) distorted beliefs.

The above model generates the prediction that attributions towards own performance will be positively biased ($\gamma_s^1 > 1$), due to the assumed benefits of overconfidence. However, the model allows for either positive or negative attributions regarding the performance of teammate 2 ($\gamma_s^2 \lesseqgtr 1$). The intuition for this result is that negative attributions towards one's teammate do increase self-serving beliefs (excess blame on the teammate reduces one's own responsibility by construction), a benefit, but also increase the financial costs, through more biased weighting choices. Implicit in the derivation of these optimal distortions, context-specific cognitive costs will impact individuals' abilities to distort γ_s^i away from one. This allows for the possibility that individuals may face varying levels of ease or difficulty in distorting beliefs tied to different dimensions (Engelmann et al., 2019) – see Appendix A for more details.[15]

Following Section 3.3, belief updating depends on the four possible states. Given the above potential distortions, for teammate 1 it follows that the model of updating with self-serving attribution bias (denoted by AB) takes the following functional form for positive and negative signals respectively.

[15]Our model assumes that the cognitive costs of mis-attributions across the two sources are independent. In our concluding discussion we discuss the possibility of relaxing this assumption.

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

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

Regarding updating about the teammate:

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

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

These parameters have the following interpretations. As noted earlier, when $\gamma_s^1 = \gamma_s^2 = 1$, updating is Bayesian. The larger γ_s^1 is, the greater are the positive attributions that the individual makes towards themselves, with an analogous relationship holding between γ_s^2 and the teammate. For example, a larger value of γ_s^1 increases the perceived likelihood that the states TT and TB generated a signal s, the states of the world where own performance is in the top half. Similarly, greater values of γ_s^2 increase the perceived likelihood that the states TT and BT generated a signal s. Our specification of the bias can thus be interpreted as an extension of the one-dimensional biased updating model of Gervais and Odean (2001).

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

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

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

distorted weighting decisions.[16]

The optimal $\gamma_s^1 \geq 1$ and $\gamma_s^2 \leq \gamma_s^1$ are such that $\left[b_{t+1}^{1,AB}|s_t = s\right] \geq \left[b_{t+1}^{1,BAYES}|s_t = s\right]$, i.e. posteriors about own performance are biased upwards. However, whether the biased posterior for teammate 2, $\left[b_{t+1}^{2,AB}|s_t = s\right]$, is smaller, equal, or larger than the Bayesian $\left[b_{t+1}^{2,BAYES}|s_t = s\right]$ depends on the value of γ_s^2.[17] Regardless of the direction, an implication of the framework is that future decisions involving the external fundamental, such as changing environments, will be further distorted, consequently heightening the likelihood of sub-optimal outcomes.

Finally we note that we can examine the nested case of the model, where distortions only occur over one dimension of uncertainty, relating to own performance, as typical in existing literature (Buser et al., 2018; Coutts, 2019a; Eil and Rao, 2011; Ertac, 2011; Grossman and Owens, 2012; Möbius et al., 2022). In this special case, $\gamma_s^2 = 1$. Because this is a restricted case, self-serving beliefs will be necessarily lower.

4 Hypotheses

The theoretical model compares belief updating to a benchmark in which updating follows Bayes' rule (Section 3). 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.

4.1 Prior Belief Formation

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

Our first hypothesis of interest concerns whether there is overconfidence in the Main treatment for teammate 1, relative to the Control treatment benchmark. Let $b_0^{1,M}$ be the average initial ($t = 0$) belief about one's own probability of scoring in the top half, where the superscript M stands for Main treatment and 1 indicates that it is teammate 1. Similarly, $b_0^{1,C}$ refers to the initial belief for teammate 1 in the Control treatment, regarding another person. By belief we refer to a participants' reported probability of being in the top half of performances. The null hypothesis is that the initial beliefs are the same across the Main and Control treatments ($b_0^{1,M} = b_0^{1,C}$). We test the following alternative hypothesis:

Hypothesis 1:

Initial beliefs about one's probability of scoring in the top half are higher in the Main than in the Control treatment.

[16] It is important to note that changes to the material incentives can create contexts where negative attributions are unambiguously optimal. Appendix C provides one example of such incentives.

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

$(b_0^{1,M} > b_0^{1,C})$

4.2 Belief Updating

Here we examine the implications of the model for the empirical framework, which follows Grether (1980) and Möbius et al. (2022); see Benjamin (2019) for additional references. Bayes' rule can be written in the following form, considering binary signals, s_t, for positive and negative signals respectively:

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

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

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

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

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

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

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

Similarly, the perceived likelihood ratio, $\hat{LR}_t^1(n)$, is:[18]

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

Denote the Bayesian likelihood ratios, calculated by setting $\gamma_s^i = 1$, by $LR_t^i(s)$. Inserting the perceived likelihood ratio in Equation 10, taking natural logarithms of both sides, and adding

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

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

$$logit(b^i_{t+1}) = logit(b^i_t) + I\{s_t = p\} \ln\left(\hat{LR}^i_t(p)\right) + I\{s_t = n\} \ln\left(\hat{LR}^i_t(n)\right). \tag{11}$$

The empirical model nests this Bayesian benchmark as follows,

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

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

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

While 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, 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 $\hat{LR}^1_t(n) \geq LR^1_t(n)$. Larger perceived likelihood ratios with self-serving attribution bias indicate that individuals perceive both positive and negative signals as being more indicative of their performance being in the top half than it really is.[19] As a result, in the empirical framework their response to positive signals will register as larger ($\beta_1 > 1$), while their response to negative signals will register as smaller ($\beta_0 < 1$).[20]

For teammate 2, the analogous distortions could result in the empirical framework registering over-response to positive and under-response to negative signals (*positive bias*) or vice-versa (*negative bias*). 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. Our null hypothesis is that participants update their beliefs about one's self and the teammate equally across Main and Control treatments ($\beta^{1,M}_1 = \beta^{1,C}_1$; $\beta^{1,M}_0 = \beta^{1,C}_0$ and $\beta^{2,M}_1 = \beta^{2,C}_1$; $\beta^{2,M}_0 =$

[19] This implication simultaneously explains the intuition for why the γ^i_s are distorted in a way which leads to larger perceived likelihood ratios – to arrive at self-serving beliefs. If any of these conditions were violated it would imply that signals are perceived as less indicative of being in the top half than they really are. If this were the case then Bayesian updating would in fact give the individual higher utility (see also Appendix A).

[20] β_1 is biased upwards because, since $ln(\hat{LR}^1_t(p)) \geq 0$, a Bayesian response to $\hat{LR}^1_t(p)$ 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) $LR^1_t(n)$.

$\beta_0^{2,C}$).[21] We test the following alternative hypothesis:

Hypothesis 2:

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

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

5 Results

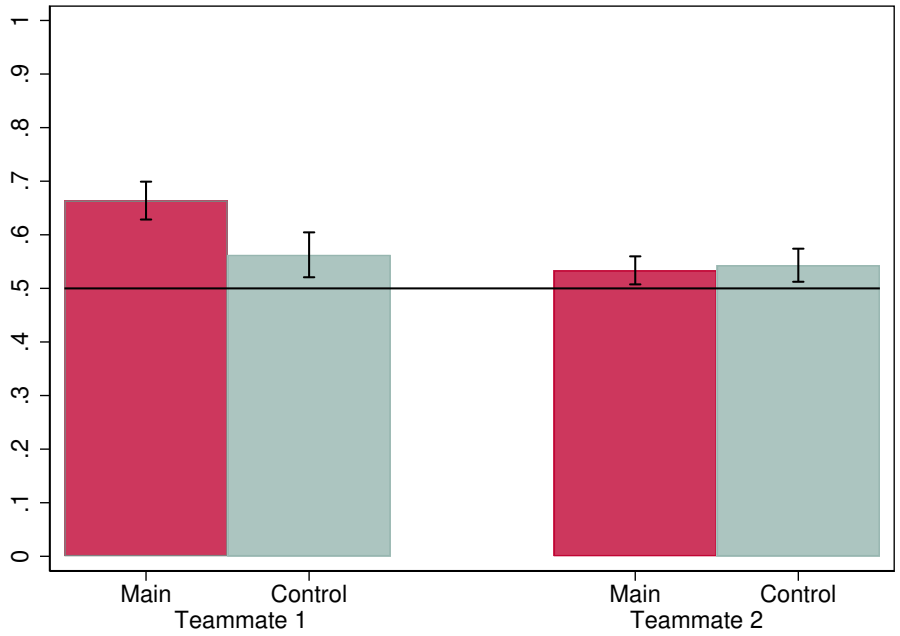
5.1 Initial Beliefs

Figure 2 presents the first round beliefs in Main and Control treatments for both teammates. In the Main treatment, where individuals estimate beliefs about their own performance, the average reported belief about being in the top half is 66.4%, significantly different from 50% in a two-sided Wilcoxon signed-rank test at the 1% level (p-value 0.0000).[22] In the Control treatment, where individuals estimate the performance of another, randomly selected individual in the position of teammate 1, the average reported belief is 56.3%. Intriguingly, this is also significantly different from 50% at the 1% level using a Wilcoxon signed-rank test (p-value 0.0046). Similarly, the beliefs that teammate 2 scores in the top half are 53.4% and 54.3% in the Main and Control treatment, respectively. These beliefs are also significantly different from 50% (Wilcoxon signed-rank tests p-values 0.0012 and 0.0017 respectively).

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

[22] For those individuals in the top half, 83% hold prior beliefs greater than 50% (compared to 76% in Control). For those in the bottom half, 54% hold prior beliefs greater than 50% (compared to 29% in Control). Note also that we use two-sided tests throughout the paper. Non-parametric tests are used as we reject normality in belief distributions, see Online Appendix Section 5.

Figure 2: Prior Beliefs by Treatment



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

These results hence appear to present evidence for "overconfidence", according to the test of Benoît and Dubra (2011). However, as these beliefs do not involve estimation of one's own performance, we regard them as a general over-estimation that is not driven by differences in Main or Control, or in teammate 1 or teammate 2 framing: a Kruskal–Wallis test does not find a significant difference across performance beliefs about teammate 1 in Control and teammate 2 in Main and Control (p-value 0.2654). Also, there are no significant differences in initial beliefs about teammate 2 between the Main and Control treatment (Wilcoxon ranksum p-value: 0.5723).

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

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

Lastly, we also note that our hard-easy manipulation affects the initial beliefs as expected (Larrick et al., 2007; Moore and Small, 2007). Individuals rate themselves in the top half with 72% probability when the test was easy, and with 62% when the test was hard (for more details, and a test of hard-easy effects on belief updating, see Online Appendix Section 2). While not our main focus, we also find evidence that men are more overconfident than women (further

details, also concerning gender differences in belief updating are provided in Online Appendix Section 3).[23]

5.2 Belief Updating

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

5.2.1 Structural Framework

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

[23] In short, we find that the main results appear to be driven by men. Men also perform better, on average, on the test, which we discuss further in Online Appendix Section 3.

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	0.59	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. δ is the coefficient on the log prior odds ratio. β_1 and β_0 are coefficients on the log likelihood of observing positive and negative signals respectively. Constant omitted because of collinearity. Bayesian updating corresponds to $\delta = \beta_1 = \beta_0 = 1$. $\beta_1, \beta_0 < 1$ indicates conservative updating. $\beta_1 - \beta_0 > 0$ indicates positive asymmetric updating.

Updating is not Bayesian in either Main or Control. All coefficients in Table 2 are significantly different from the Bayesian prediction of 1, indicated by asterisks. Column 1 reveals that positive signals are given significantly more weight than negative signals when updating is about own performance ($\beta_1^{1,M} > \beta_0^{1,M}$, significant at the 1% level). No such asymmetry is observed in column 2, in the Control treatment, for updating about another's performance.[24]

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

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

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

Table 3: Updating Beliefs about Teammate 2

Regressor	(1) Main Treatment	(2) Control Treatment
δ	0.770***	0.717***
	(0.048)	(0.050)
β_1	0.398***	0.491***
	(0.056)	(0.070)
β_0	0.248***	0.418***
	(0.043)	(0.061)
P-Value ($\delta = 1$)	0.0000	0.0000
P-Value ($\beta_1 = 1$)	0.0000	0.0000
P-Value ($\beta_0 = 1$)	0.0000	0.0000
P-Value ($\beta_1 = \beta_0$)	0.0358	0.3708
R^2	0.55	0.50
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. δ is the coefficient on the log prior odds ratio. β_1 and β_0 are coefficients on the log likelihood of observing positive and negative signals respectively. Constant omitted because of collinearity. Bayesian updating corresponds to $\delta = \beta_1 = \beta_0 = 1$. $\beta_1, \beta_0 < 1$ indicates conservative updating. $\beta_1 - \beta_0 > 0$ indicates positive asymmetric updating.

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

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

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

As noted earlier in Section 3 and detailed in Appendix A, some positively biased updat-

ing about teammate 2 can be optimal since it permits self-serving beliefs, while reducing the material costs of such beliefs, due to more moderate weighting between the two teammates. Interestingly, for positive signals, $\beta_1^{1,M}$ in Table 2 column 1 is significantly greater than $\beta_1^{2,M}$ in Table 3 column 1 (Chow test p-value 0.0062). For negative signals, the respective $\beta_0^{1,M}$ and $\beta_0^{2,M}$ coefficients do not differ significantly (Chow test p-value 0.8637). Taken together, the difference in asymmetry ($\beta_1^{1,M} - \beta_0^{1,M}$) versus ($\beta_1^{2,M} - \beta_0^{2,M}$) across the first columns in Tables 2 and 3 is significant at the 10% level (Chow test p-value 0.0963).[25] Hence, while we find positive asymmetry for both self and teammate 2, it is stronger when updating about one's self.

There are a few potential alternative explanations for the observation of positively biased updating for both teammate 1 and teammate 2 in the Main treatment. We briefly discuss three more prominent ones here and address them in more detail in Online Appendix Section 7: first, that anchoring causes individuals to update similarly about teammate 2, second that participants selectively discount or ignore negative signals, and third that positively biased updating for teammates is driven by an in-group bias.

First, if individuals update in a self-serving manner for themselves, which mechanically anchors their updating about the teammate, then we should see similar patterns in the Follow-up experiment. As will be detailed in Section 6.1.2, this is not the case – instead updating is not self-serving which suggests the identity of the dimension of uncertainty (a human teammate) is central to the results. Second, participants in our Main treatment selectively ignore negative signals at equivalent rates to those in the Control treatment. Third, should an in-group bias drive the results, we would anticipate elevated prior beliefs for teammate 2 – however initial prior beliefs for teammate 2 are not statistically different across Main and Control. With that said, we cannot exclude a type of in-group bias that is specific only to information processing. Note that a variation of such a bias could however be incorporated into our theoretical framework, e.g., by assuming cognitive costs of negative attributions towards an in-group target.

5.2.2 Matching on Priors

After having shown that beliefs are updated differently in the Main versus Control treatments in a quasi-Bayesian framework, we also examine the extent to which updating differs across treatments without any reliance on the Bayesian benchmark. Appendix D presents a matching strategy that compares the posteriors at the end of Part 2 for Main and Control participants, conditional on having the same (1) initial prior beliefs in round 1, and (2) total number of negative signals received.

This matching strategy reveals that, given the same prior and proportion of signals observed, individuals updating about their own performance (Main treatment) end up with posteriors significantly higher than those updating about the performance of a randomly chosen teammate 1. Appendix D further shows that this effect is strongest for individuals receiving all negative signals. Regarding teammate 2, differences are in the same direction but not

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

statistically significant.

Result 4: *In line with the findings from the structural framework, individuals who update about their own performance (Main treatment) end up with posteriors that are 6.5 to 7.5 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 differences for updating about teammate 2 go into the same direction, but are smaller in magnitude and not statistically significant at conventional levels.*

5.3 Willingness to Change Teammates

We now examine whether the observed distortions in updating concerning the teammate yield broader consequences within our experimental setting. To do so, we provided our participants with a surprise opportunity to change teammates. In wave 2 we measured the participants' 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 participant is willing to change teammates. While we also study the intensive margin in Appendix E, that analysis is confounded by the fact that the value of switching teammates depends also on beliefs about own performance.

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

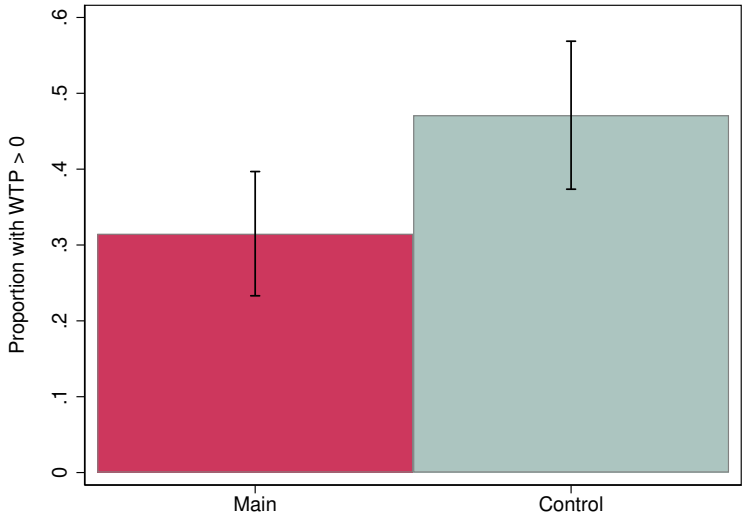
Result 5: *As a result of biased updating about teammate 2, participants in the Main treatment are 34% less likely to want to change teammates than their Control counterparts.*

Note that this result does not derive from participant's more overconfident initial beliefs in the Main compared to the Control treatment. Before feedback, the proportion of those willing to switch teammates should be the same in both treatments. The reason is that before feedback, the decision to change teammates depends only on the belief about teammate 2's performance. Result 5 thus confirms that the biased updating patterns we observed translate into actual differences in future decision making. Moreover, it suggests that participants 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.

The fact that self-serving motives can motivate distorted beliefs about others which impact behaviour has critical implications for whether and how individuals change environments. Importantly, our result that overconfident individuals are less likely to switch teammates contrasts with the recent theory literature involving multi-dimensional uncertainty. When there

is only one-dimension of distortion, Hestermann and Le Yaouanq (2021) show the opposite – overconfident individuals should be less satisfied with their environments, and therefore more likely to change them. As a result, they show that overconfidence should not persist over the long run, though underconfidence will due to analogous reasoning. Importantly, our result shows that allowing for distortion about other fundamentals can lead to scenarios where individuals are less likely to change environments, and overconfidence is thus likely to persist. This could help explain why real world evidence has suggested instances of both overconfidence and underconfidence (Dunning, 2005).[26]

Figure 3: Willingness to Switch



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

6 Follow-up and Validation Experiments

6.1 Follow-up

6.1.1 Overview

Our theoretical framework highlights the pivotal role of the source of uncertainty. Context-specific cognitive costs – illustrated by factors like the ease of distorting beliefs about a teammate – can influence the magnitude of distortions, thereby underscoring their importance in shaping self-serving biases. In our Follow-up lab experiment, we explored precisely this theme, namely whether belief distortions differ depending on the source of uncertainty: human teammate versus random fundamental.

[26] In their model of misguided learning with multi-dimensional uncertainty, Heidhues et al. (2018) also make the point that overconfident agents will exit environments more frequently. However, their focus on more extreme forms of overconfidence where no learning or severely limited learning is assumed, means that, by design, overconfident individuals never learn the truth and overconfidence does persist in the long run.

The Follow-up experiment sessions were conducted in-person using the same software and participant-pool at the University of Hamburg during the 2021-22 academic year, with 219 participants. These sessions were identical to the Main treatment, but with the critical difference that teammate 2 was replaced with a random fundamental. Thus, instead of being paired with another participant in the same session, participants were (truthfully) told that they had been matched with a random fundamental (referred to as a random factor in the instructions), that could take on one of two values: HIGH or LOW. Everything else about the experiment was identical, with a HIGH or LOW value of the random fundamental being equivalent to teammate 2 being in the top half or bottom half, respectively.[27]

To ensure prior beliefs about the random fundamental were similar to beliefs about a human teammate, participants were given a range which corresponded to the probability that the random fundamental was HIGH. This range was ± 15 percentage points from a randomly selected prior belief about teammate 2's performance (taken from the Main experiment). For instance, for a specific prior belief of 50%, the range for the random fundamental to be HIGH would have been given as 35% to 65%.[28]

6.1.2 Results

First, we confirm that initial prior beliefs in the Follow-up experiment are similar to the Main treatment of the Primary experiment. For teammate 1 average prior beliefs about own performance being in the top half is 67.7%, which is not significantly different from the Main treatment (66.4%, Wilcoxon ranksum test p-value 0.6665). Average prior beliefs for the random fundamental are 54.5% (53.4% for the human teammate 2 in Main, Wilcoxon ranksum test p-value 0.8895).

Table 4 presents the same specifications from previous Tables 2 and 3, showing belief updating for self (teammate 1) and the random fundamental (teammate 2). Immediately, one can see that there is no asymmetry in belief updating, neither for self nor the random fundamental. Comparing the results for teammate 1 to the Control treatment (Table 2, column 2) reveal nearly identical response to signals; Chow tests confirm no significant differences for positive (β_1) or negative (β_0) signals (nor overall asymmetry, $\beta_1 - \beta_0$). For the random fundamental, though response to signals is slightly smaller, there are similarly no significant differences with teammate 2 in the Control treatment (Table 3, column 2).

As the material incentives were identical in the Follow-up and Primary experiments, the lack of asymmetric belief distortion is potentially surprising. It suggests that the differences in updating behaviour must derive from the differences between the human versus non-human (random fundamental) nature of the teammate. Our theoretical framework posited that cogni-

[27] Recruitment aimed for gender balance, 51% of participants reported to be women. One difference was that due to difficulties in recruitment, we adapted the design to permit sessions with fewer than 20 participants. In all cases, participants were told that should a session involve fewer than 20 participants, past-participants would be added to generate the ranking. This did not affect the matching process which was always done within-session.

[28] The reason to include a range was to generate additional uncertainty to better match the human version. Prior beliefs that were either $< 15\%$ or $> 85\%$ were excluded (11% of priors).

tive costs of distortion account for how difficult it is to distort beliefs. Through this lens, such a result would arise when individuals find it less costly to distort beliefs when matched with a human, enabling the self-serving beliefs found in the Main treatment.[29] To study precisely this aspect of belief distortion, we conducted an online Validation experiment, described in detail in the next subsection.

Table 4: Updating Beliefs in Follow-up

Regressor	(1) Teammate 1	(2) Random Fundamental
δ	0.835***	0.646***
	(0.038)	(0.063)
β_1	0.516***	0.352***
	(0.064)	(0.058)
β_0	0.507***	0.364***
	(0.062)	(0.043)
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 ($\beta_1 = \beta_0$)	0.9197	0.8690
R^2	0.69	0.41
Observations	610	706
P-Value [Chow-test] comparing to Control in Column (2) of Tables 2 and 3 respectively:		
δ	0.1488	0.3779
β_1	0.9184	0.1266
β_0	0.9948	0.4666
$(\beta_1 - \beta_0)$	0.9345	0.4292

Analysis uses OLS regression. Difference is *significant from 1* at * 0.1; ** 0.05; *** 0.01. Robust standard errors clustered at individual level. 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.

6.2 Validation Experiment

6.2.1 Overview

To reconcile the contrasting results between the Main and Follow-up experiments, in 2023 we conducted a Validation experiment on Prolific.co with $N = 600$ participants. The experiment

[29] This result is also interesting in light of the blame-shifting literature. Bartling and Fischbacher (2012) showed evidence suggesting that delegating to another human reduces responsibility more than delegating to random processes (such as a die roll), with Oexl and Grossman (2013) finding that individuals are punished even when they have no autonomy over their choices. Based on this literature, we might have anticipated greater attribution (stronger response to signals) when teammate 2 was human, though as noted the differences observed were not statistically significant.

tested the hypothesis that individuals would show a greater tendency to distort beliefs concerning a human than a random fundamental.[30] Further experimental details are provided in Online Appendix Section 9.

The experiment involved two between-subject treatments, where a participant was either matched with a person who previously participated in our lab experiments (treatment Human; $N = 301$) or with a random fundamental (treatment RF; $N = 299$). The matched person in Human (the random fundamental in RF) could take the values of Top (High) or Bottom (Low). Participants in the Human treatment knew that their matched person participated in an IQ quiz and their scores were ranked from 1 to 20 where ranks from 1-10 were Top while ranks from 11-20 were Bottom. Analogously, participants in the RF treatment knew that their matched random fundamental was determined by the random draw of a number between 1-20, where numbers 11-20 were High and numbers from 1-10 were Low. They were asked to provide a probability estimate that their match was Top/High depending on treatment (*initial estimate*). The probability estimate was incentivised with a binarised scoring rule (BSR) (Hossain and Okui, 2013) so that the more accurate estimates would have a higher chance of earning 50 tokens (1 token = £0.01).

To test whether individuals were more likely to distort their beliefs about a human relative to a random fundamental, after providing their initial probability estimate, participants were given an opportunity to revise their estimate (*revised estimate*) under a unique incentive scheme. For this scheme, in addition to the standard BSR incentives, they received a fixed amount of 1 token for every 5 percentage points of change from their initial estimate, upwards or downwards. Theoretically, under this scheme the optimal action is to distort beliefs. For example, for a risk-neutral individual the optimal distortion is 20 percentage points, see Online Appendix Section 9.

Following our theoretical framework, we assume individuals face cognitive costs of distorting beliefs from their subjectively held accurate beliefs. Hence in determining the optimal distortion, individuals trade-off the material benefits from distortion against these cognitive costs. It follows that, when the cognitive costs are greater, the willingness to distort will be lower. Our primary hypothesis is that distortion will thus be greater in the Human than RF treatment.

6.2.2 Results

As expected, initial beliefs about being Top/High were nearly identical between the Human and RF treatments: participants believed to be matched to a Top Human and High RF with 55.1% and 55.0% chance, respectively (Wilcoxon ranksum p-value 0.9464). Coming to the main result of interest, participants were more likely to revise their estimates (irrespective of direction) in the Human treatment: 76.1% versus 67.6% ($\chi 2$ test p-value 0.0200). The average absolute revision was 16.1 versus 14.6 percentage points (Wilcoxon ranksum p-value 0.0653) for the Human vs RF treatments respectively. Hence the main results of the Validation experiment

[30] As in the Follow-up experiment, within the Validation experiment we used the identical language of "Random Factor" to refer to the random fundamental.

confirm the hypothesis that individuals are more willing to distort their beliefs when matched with a human than when matched with a (non-human) random fundamental.[31]

7 Discussion

Previous literature has often focused on the relatively narrow view of biased information processing as a one-dimensional phenomenon: self-serving attribution at the expense of "other factors". Yet our theoretical framework underscores how multiple dimensions of uncertainty can unlock additional levers that enable self-serving biases. Our series of experimental results highlights the importance of material incentives and context, showcasing how changes in the environment can enable or constrain belief distortions.

These insights offer significant contributions to the existing literature on self-serving biases. While prior work in psychology focused on negative attributions towards other factors as a means of ego protection and enhancement (Campbell and Sedikides, 1999), our theory and results show that attributions can be strategic responses to economic incentives and other features of the environment, not limited to the negative direction. This could help explain some of the mixed evidence on the strength and direction of attributions (Miller and Ross, 1975; Zuckerman, 1979).

Within economics, there has been recent interest in the reverse perspective: studying how attributions are affected by initial confidence biases (Heidhues et al., 2018; Hestermann and Le Yaouanq, 2021). In particular, Heidhues et al. (2018) establish conditions for misguided negative attribution, showcasing how an overconfident but otherwise Bayesian agent can develop increasing pessimism towards an external fundamental, a finding empirically supported by Goette and Kozakiewicz (2020) and Marray et al. (2021).[32] Critically, when we allow for (multi-dimensional) belief distortion, our theoretical and empirical results show that it is possible to generate the opposite finding: positive attributions, as was the case in our Primary experiment.

Beyond this, our results highlight the relevance of environmental features in shaping self-serving beliefs. In our Primary experiment, we observe self-serving information processing, driven by a positive bias towards a teammate which mitigated negative financial consequences in our experiment. Yet in our Follow-up experiment with the same material incentives, we find no distortions when updating about a random fundamental, which appears to have constrained self-serving beliefs. Such patterns are consistent with our theoretical framework when individuals

[31] It is more difficult to shed light on the mechanisms underlying the result of greater distortion when matched with a human. Online Appendix Section 9 provides details about potential mechanisms. We examined a measure of warmth towards the matched Human/RF, to study whether higher reported feelings of warmth were associated with more or less belief distortion. We also examined a measure of cognitive uncertainty (Enke and Graeber, 2023), to examine whether greater subjective uncertainty was related to belief distortion. We find no evidence that warmth towards the matched Human/RF matters for explaining revision. We do find that participants reporting more cognitive uncertainty are more likely to revise their beliefs, however the interaction of cognitive uncertainty and treatment was not significant.

[32] Both papers show evidence for self-defeating learning, finding that beliefs about the external fundamental become less accurate for overconfident participants. Beyond the difference with our focus on distorted belief updating, our experimental design intentionally shutdown the link between actions and feedback, which drives the self-defeating learning they study.

find it easier to manipulate their beliefs about another human – a result we find support for in our Validation experiment.

This set of results has important implications. First, other dimensions of uncertainty can impact the extent of self-serving beliefs in different ways. Given the variation in environments used to study self-serving belief updating in economics, our findings may also help explain the decidedly mixed literature on self-serving belief updating in economics (Benjamin, 2019; Drobner, 2022).[33] More specifically, the distinction between human and non-human dimensions we identify indicates a nuanced psychological mechanism at play. In real-world settings, such as the workplace, an employee might distort their perception of a colleague's performance to bolster their own ego, but they might be less likely to do so with non-human factors like market conditions or automated tools such as artificial intelligence. If human relationships and interactions have a distinct influence on how we update our beliefs, this suggests important considerations for how organisations and teams manage feedback, performance evaluations, and team dynamics.

A second implication relates to how belief distortion evolves over the long run, and how this impacts individual decisions. In our Primary experiment, we find evidence that distorted beliefs affect decision-making, through a reduced willingness to change teammates. As joining a different team provides a new, independent source of information, this can slow down the learning process, providing a potential explanation for why overconfidence is sometimes observed to be persistent. More broadly, as we note that distortions towards other dimensions can vary in direction based on the incentives present in the environment; this has wide ranging implications for learning. In particular, in distorting other dimensions of uncertainty to arrive at self-serving beliefs, individuals may end up more or less likely to change environments, and therefore more or less likely to learn (Hestermann and Le Yaouanq, 2021).[34]

Our results raise important questions and multiple avenues for future research. A first step would be to better understand the costs of belief distortion. Our model allows for distinct and independent cognitive costs across dimensions, and our results suggest differences in our ability to distort our perceptions of human versus non-human sources of uncertainty. The end of Section 5.2.1 mentions potential alternative explanations for the empirical findings in the Primary experiment. While we find that some observed patterns are inconsistent with an in-group bias, such a bias could manifest itself as further cognitive costs of negative distortions

[33] This empirical literature is typically focused on asymmetry in updating with one dimension of uncertainty. Different authors have found: Positive asymmetry (Drobner and Goerg, 2022; Eil and Rao, 2011; Möbius et al., 2022), no asymmetry (Buser et al., 2018; Grossman and Owens, 2012), and negative asymmetry (Coutts, 2019b; Ertac, 2011) have all been observed. Buser et al. (2018) do find positive asymmetry in some sub-samples. Reactions to feedback have also been studied in less comparable or non ego-relevant settings, see Barron (2021), Burks et al. (2013), Charness and Dave (2017), Eberlein et al. (2011), Erkal et al. (2022), Ertac and Szentes (2011), Gotthard-Real (2017), Pulford and Colman (1997), and Wozniak et al. (2014).

[34] Hestermann and Le Yaouanq (2021) show that with Bayesian updating, underconfidence, not overconfidence should persist in the long run, as overconfident individuals will change environments more frequently, and thus learn from their encounters with varying external fundamentals. Our result of positive bias provides one example where the opposite is true. Note that in our experiment, the opportunity to change teammates came as a surprise to participants. To the extent that such opportunities can sometimes be predictable in the real world, we might expect this would limit the welfare consequences. We thank an anonymous referee for bringing this point to our attention.

towards an in-group target. Beyond this, one could consider a world where the costs of distortion across different sources of uncertainty are not independent. For example, does distortion in one dimension make distortion in another dimension more costly?

More broadly, our results present a way forward for thinking about how individuals select into or leave certain environments, to nurture their preferred worldview. Do people choose to work with others in anticipation of how they will rationalise good or bad outcomes? Do they choose environments in which the material costs of overconfidence are lower, or in which outcomes may be more easily attributed among various sources? These questions are critical for future research. In the end, if self-serving belief formation motivates strategic behaviour in how we choose our environments, and how we process information within those environments, we should not be surprised to find that for many individuals overconfidence could persist over the long run.

8 Supplementary data

The data and codes for this paper are available on the Journal repository. They were checked for their ability to reproduce the results presented in the paper. The replication package for this paper is available at the following address: https://doi.org/10.5281/zenodo.10535890.

References

Azrieli, Yaron, Christopher P Chambers, and Paul J Healy, "Incentives in experiments: A theoretical analysis," *Journal of Political Economy*, 2018, *126* (4), 1472–1503.

Baldiga, Katherine, "Gender differences in willingness to guess," *Management Science*, 2014, *60* (2), 434–448.

Baron, Jonathan and John C. Hershey, "Outcome bias in decision evaluation," *Journal of Personality and Social Psychology*, 1988, *54* (4), 569–579.

Barron, Kai, "Belief updating: does the 'good-news, bad-news' asymmetry extend to purely financial domains?," *Experimental Economics*, 2021, *24* (1), 31–58.

Bartling, Björn and Urs Fischbacher, "Shifting the blame: On delegation and responsibility," *The Review of Economic Studies*, 2012, *79* (1), 67–87.

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

_ and _, "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," *The Quarterly Journal of Economics*, 2011, *126* (2), 805–855.

Benjamin, Daniel J, "Errors in probabilistic reasoning and judgment biases," *Handbook of Behavioral Economics: Applications and Foundations 1*, 2019, *2*, 69–186.

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*, 2012, *75* (1), 67–80.

Brownback, Andy and Michael A. Kuhn, "Understanding outcome bias," *Games and Economic Behavior*, 2019, *117*, 342–360.

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

Burks, Stephen V, Jeffrey P Carpenter, Lorenz Goette, and Aldo Rustichini, "Overconfidence and social signalling," *Review of Economic Studies*, 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*, 2018, *56*, 165–192.

Campbell, W Keith and Constantine Sedikides, "Self-threat magnifies the self-serving bias: A meta-analytic integration," *Review of General Psychology*, 1999, *3* (1), 23–43.

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

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

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

Drobner, Christoph, "Motivated Beliefs and Anticipation of Uncertainty Resolution," *American Economic Review: Insights*, 2022, *4* (1), 89–105.

_ **and Sebastian J. Goerg**, "Motivated Belief Updating and Rationalization of Information," *SSRN Electronic Journal*, 2022.

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

Eberlein, Marion, Sandra Ludwig, and Julia Nafziger, "The effects of feedback on self-assessment," *Bulletin of Economic Research*, 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*, 2011, *3* (2), 114–138.

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

Enke, Benjamin and Thomas Graeber, "Cognitive uncertainty," *The Quarterly Journal of Economics*, 2023, *138* (4), 2021–2067.

Erkal, Nisvan, Lata Gangadharan, and Boon Han Koh, "By chance or by choice? Biased attribution of others' outcomes when social preferences matter," *Experimental Economics*, 2022, *25* (2), 413–443.

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

_ **and Balazs Szentes**, "The effect of information on gender differences in competitiveness: Experimental evidence," Technical Report, Working Paper, Koç University 2011.

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

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

Goette, Lorenz and Marta Kozakiewicz, "Experimental evidence on misguided learning," Technical Report, Working Paper, University of Bonn 2020.

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

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

Grether, David M, "Bayes rule as a descriptive model: The representativeness heuristic," *The Quarterly Journal of Economics*, 1980, *95* (3), 537–557.

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

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

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

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

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

Hestermann, Nina and Yves Le Yaouanq, "Experimentation with self-serving attribution biases," *American Economic Journal: Microeconomics*, 2021, *13* (3), 198–237.

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

Hossain, Tanjim and Ryo Okui, "The binarized scoring rule," *Review of Economic Studies*, 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*, 1980, *31* (1), 457–501.

Kunda, Ziva, "The case for motivated reasoning," *Psychological Bulletin*, 1990, *108* (3), 480.

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*, 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), 299–305.

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

Marray, Kieran, Nikhil Krishna, and Jarel Tang, "How do expectations affect learning about fundamentals? some experimental evidence," Technical Report, Working Paper, University of Oxford 2021.

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, Markus M, Muriel Niederle, Paul Niehaus, and Tanya S Rosenblat, "Managing self-confidence: Theory and experimental evidence," *Management Science*, 2022, *68* (11), 7793–7817.

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.

Oexl, Regine and Zachary J. Grossman, "Shifting the blame to a powerless intermediary," *Experimental Economics*, 2013, *16* (3), 306–312.

Oster, Emily, Ira Shoulson, and E Ray Dorsey, "Optimal expectations and limited medical testing: Evidence from Huntington disease," *American Economic Review*, 2013, *103* (2), 804–830.

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*, 1997, *23* (1), 125–133.

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

_ , Egon Tripodi, and Joël J. van der Weele, "Self-Persuasion: Evidence from Field Experiments at International Debating Competitions," *American Economic Review*, 2022, *112* (4), 1118–1146.

Schwarz, Norbert and Gerald L. Clore, "Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states," *Journal of Personality and Social Psychology*, 1983, *45* (3), 513–523.

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*, 1973, *5* (2), 207–232.

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

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

Zimmermann, Florian, "The Dynamics of Motivated Beliefs," *American Economic Review*, 2020, *110* (2), 337–363.

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

Appendix

A Model of Optimal Information Distortion

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

We introduce the possibility that individuals receive direct utility over the belief that they are in the top half, through a linear function $\alpha \cdot b_{t+1}^1$.[35] $\alpha \in [0, \infty)$ indicates the extent to which the individual benefits from holding overconfident beliefs. This can be thought of as a reduced form interpretation of the benefits to overconfidence, for example direct hedonic utility benefits, signalling to others, or benefits from motivation. Importantly, we assume that individuals do not derive any benefit from beliefs about others' ability, nor do they derive direct benefit from beliefs about the four states TT, TB, BT, BB. Of course, since $b_{t+1}^1 = b_{t+1}^{TT} + b_{t+1}^{TB}$, indirectly they can benefit from these beliefs.

We follow the literature and assume that a subconscious process trades off these benefits from overconfidence against the costs, which we posit to be material costs from inefficient decision making as well as mental costs of distorting the updating process. In the experiment, these material costs are the lower expected probability of earning $P = €10$. Following Bracha and Brown (2012), we assume mental cost functions $J_i(\gamma_s^i, 1)$ that are convex and strictly increasing in $|\gamma_s^i - 1|$, i.e. minimised at the Bayesian information processing parameter $\gamma_s^i = 1$.[36] We will further assume that the mental costs of distorting γ_s^1 and γ_s^2 are separable, noting that we allow them to take different potential functional forms.

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.[37] We first note that if participants 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

[35] 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.

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

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

two states, $\hat{b}^{TB}_{t+1}$ and $\hat{b}^{BT}_{t+1}$. Given the form of the bias for updating about own ability, this will imply an over-weighting of the likelihood of state TB by γ^1_s, and an over- or under-weighting of the likelihood of state BT by γ^2_s.

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

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

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

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

Taking the second derivative, and letting $\bar{x}_1=\gamma^2_p\Phi_{TT}b^{TT}_t+\Phi_{TB}b^{TB}_t$:

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

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

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:[38]

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

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

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

$$\max_{\{\gamma_s\}} \Big\{ \alpha \cdot \hat{b}^1_{t+1} + b^{TT}_{t+1} \cdot u(P) + b^{TB}_{t+1} \cdot \sqrt{\hat{\omega}^*_{t+1}} \cdot u(P) + b^{TB}_{t+1} \cdot (1 - \sqrt{\hat{\omega}^*_{t+1}}) \cdot u(0) \tag{14}$$
$$+ b^{BT}_{t+1} \cdot \sqrt{1 - \hat{\omega}^*_{t+1}} \cdot u(P) + b^{BT}_{t+1} \cdot (1 - \sqrt{1 - \hat{\omega}^*_{t+1}}) \cdot u(0) + b^{BB}_{t+1} \cdot u(0)$$
$$- J_1(\gamma^1_s, 1) - J_2(\gamma^2_s, 1) \Big\}.$$

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 maximised when $\gamma^1_s = \gamma^2_s$, resulting in an unbiased weight. The final two terms are mental costs, which are minimised when $\gamma^i_s = 1$, i.e. updating is Bayesian.

By the properties of the mental cost function $J_i(\gamma^i_s, 1)$, 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.[39] Substituting biased beliefs and weights into the maximisation, and substituting the values of Φ from the experiment, the first order condition with respect to γ^1_s is (where $u(P) - u(0) = \Delta u$):

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

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

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

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

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

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

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

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

[39] 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 $\gamma_s^2 = \gamma_s^1$.[40]

B Deriving the Condition for $\Theta < 0$

B.1 Theoretical Result

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

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

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

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

The expression of interest in the first round is thus:

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

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

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

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

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

[40] Our model assumes that the process determining the optimal γ_s^1, γ_s^2 occurs at the subconscious level. In other words, individuals might not be able to explicitly compute these optimal parameters, yet they act as if they can. It's worth noting that while this optimisation is complex, the belief elicitation interface in our experiment ensures that individuals have an intuition of how belief distortions influence the submitted weight, and consequently, their payoffs.

for the four states back into the initial expression of interest, Equation 17:

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

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

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

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:

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

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:

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

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 fewer than 2% of cases is this expression positive.

C Theoretical Alternative: Negative Attributions

Our results in Section 3.4 on self-serving attribution bias showed that either (i) negative or (ii) positive attributions towards teammate 2 are consistent with our theory. First, by blaming others one can directly increase self-serving beliefs (success is then over-attributed to self, failure is over-attributed to other). But second, positive attributions counterbalance biased weighting allocations that result from self-serving beliefs. While our experiment can resolve this ambiguous result for our context, here we want to emphasise the importance of different incentive structures on generating different predictions.

Consider the following change to the payoffs, where the weighting now only affects the states TT (both rank in the top half) and BB (both rank in the bottom half). When one teammate scores in the top half and the other one in the bottom half, the payoffs are fixed.

$$\Pi^t(\omega_t, A_1, A_2) = \begin{cases} (P, 0; \sqrt{\omega_t}) & \text{if } TT \\ P & \text{if } TB \\ 0 & \text{if } BT \\ (P, 0; \sqrt{1-\omega_t}) & \text{if } BB \end{cases} \tag{18}$$

Analogous to the earlier distorted weight shown in Equation 9, the optimal weight is distorted by the parameters γ_s^1 and γ_s^2.[41]

$$\omega_t^* = \frac{1}{1 + \left(\frac{b_t^{BB}}{9\gamma_s^1 \gamma_s^2 b_t^{TT}}\right)^2}. \tag{19}$$

This weight is distorted, with material payoff consequences, whenever $\gamma_s^1 \gamma_s^2 \neq 1$. This means that individuals now have incentives to counterbalance positive self-attributions ($\gamma_s^1 > 1$) with negative other-attributions ($\gamma_s^2 < 1$). Thus, unlike the incentives our experiment, under these conditions individuals' incentives would be aligned towards negative attributions towards the teammate: both because of the benefits of self-serving attributions, but also the benefits of

[41] The $\frac{1}{9}$ term enters because of the ratio of the likelihoods $\frac{0.1}{0.9}$ of the two states.

counterbalancing the material costs of submitted a distorted weight.[42] Though we do not study such a treatment in our experiment, it is important to showcase how changes in the incentives can alter the theoretical predictions.

D Matching on Priors

Here we present a non-parametric analysis of updated beliefs, which utilises a matching strategy that conditions the Main and Control participants on their initial prior beliefs in round 1, and then compares their posteriors at the end of Part 2 after four rounds of feedback.[43] By matching on initial prior beliefs we step away from the reliance on the Bayesian benchmark, and instead ask the following question: given the same prior, do participants 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 participants face the same number of positive and negative signals, we force exact matching on the total number of negative signals received over the four rounds of feedback. Matching on both priors and the proportion of negative signals received summarises all of the information that individuals have about the teammates' abilities.[44]

Table D.1 presents the results of this exercise reporting average treatment effects (ATE). We find a significant difference in posterior beliefs between the Main treatment, which involves updating about one's own performance, and the treatment involving updating about the performance of a randomly chosen teammate 1. Conditional on having the same priors and exposure to an equal proportion of negative (versus positive) signals, the posterior beliefs of individuals in the Main treatment are estimated to be 6.5 to 7.5 percentage points higher. This finding corroborates evidence of divergent information processing between the two treatments.

[42] As shown in Equation 9, the incentives in our experiment lead to distortion whenever $\frac{\gamma_s^2}{\gamma_s^1} \neq 1$, which generate incentives to counterbalance positive self-attributions ($\gamma_s^1 > 1$) with positive other-attributions ($\gamma_s^2 > 1$).

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

[44] Matching follows a k-nearest neighbour strategy, searching for the Control individual with the closest prior (to a maximum caliper of 0.03, with replacement). The exact matching requires that the Control individual(s) received the exact same number of negative signals as the Main individual. Main treatment observations are dropped when there is no common support (when the prior is greater than the maximum or less than the minimum prior among Control individuals) – less than 12% of the sample.

Table D.1: Main vs Control: Belief Teammate 1 Top

	(1) 1 Neighbour	(2) 2 Neighbours
ATE	0.075**	0.065**
	(0.032)	(0.028)
Observations	373	373

Analysis uses nearest neighbour matching, with replacement when > 1 neighbour. 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 D.2: Main vs Control: Belief Teammate 1 Top by Proportion of Negative Signals Received

	(1) 0 −	(2) 1 −	(3) 2 −	(4) 3 −	(5) 4 −
ATE	−0.021	0.061	0.139***	−0.025	0.179**
	(0.065)	(0.083)	(0.046)	(0.087)	(0.083)
Observations	73	68	99	60	73

Analysis uses nearest neighbour matching with 1 neighbour. 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).

Our structural analysis suggests this difference in updating is driven primarily by under-responsiveness to negative signals. To investigate this in our non-parametric framework, Table D.2 presents matching estimates for each of the possible distributions of observed signals separately. Consistent with the structural framework, receiving 4 negative signals (0 positive) turns out to reveal the greatest difference between Main versus Control: participants with the same initial priors end up an estimated 17.9 percentage points more confident when they are estimating their own performance. The only other significant effect is found for a balanced distribution of 2 positive and 2 negative signals.

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 D.3 and D.4. The estimates suggest that posterior beliefs about one's teammate are between 4.9 and 5.2 percentage points greater in Main relative to Control, however this is not statistically significant at conventional levels (respective p-values: 0.1314 and 0.1625). Examining the ATE estimates separately for different distributions of negative signals received, receiving all negative signals is associated with a large and significant effect. Individuals with the same priors about teammate 2 in Main and Control who receive only negative signals end up with posteriors about teammate 2 that are approximately

14 percentage points greater in Main relative to Control. Again, this supports our structural results.

Table D.3: Main vs Control: Belief Teammate 2 Top

	(1) 1 Neighbour	(2) 2 Neighbours
ATE	0.052 (0.037)	0.049 (0.033)
Observations	376	376

Analysis uses nearest neighbour matching, with replacement when > 1 neighbour. 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 D.4: Main vs Control: Belief Teammate 2 Top by Proportion of Negative Signals Received

	(1) 0 −	(2) 1 −	(3) 2 −	(4) 3 −	(5) 4 −
ATE	−0.014 (0.098)	0.077 (0.095)	0.033 (0.071)	−0.013 (0.088)	0.139** (0.065)
Observations	69	74	92	52	89

Analysis uses nearest neighbour matching with 1 neighbour. 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).

E Willingness to Pay to Switch Teammates

In wave 2 we provided participants with the opportunity to be randomly re-matched to a new teammate 2, using the BDM mechanism. Participants i could bid $x_i \in$ €$[0, 5]$, where €5 is the risk-neutral maximum value of switching.[45] 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 participants we are able to calculate whether it would be optimal for them to switch teammates, assuming risk neutrality. Before receiving feedback,

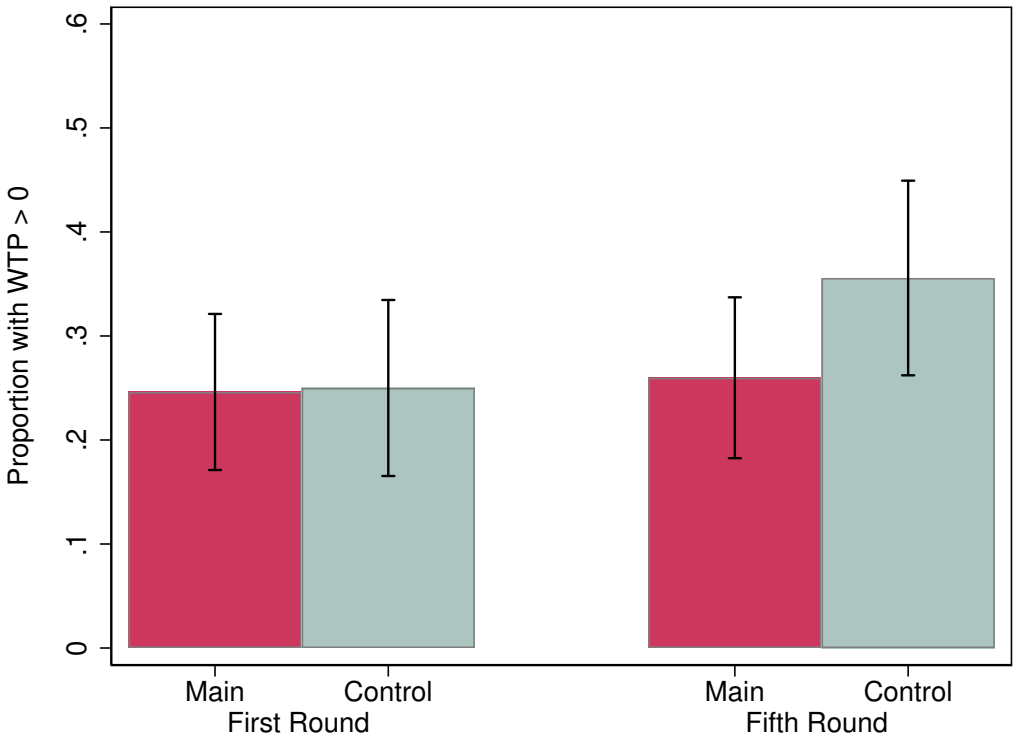
[45] Note that the worst outcome for participants is when both teammates are in the bottom half, where they will earn €0 with certainty. If one is in the top half, they can select ω accordingly to ensure a high probability of earning €10. Since there is a 50% probability a randomly selected person is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

this decision depends entirely on the belief about teammate 2. If participants believe their teammate is in the top half with probability less than 50% they should pay to switch, otherwise they should not be willing to pay any positive amount.[46]

Since initial beliefs about teammate 2 are not statistically different across Main and Control treatments, we would predict that the number of participants willing to pay a positive amount to switch 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 participants which should opt to switch teammates.

While initial prior beliefs are such that there are no differences across Main and Control treatments, beliefs after four rounds of feedback (Round 5) are such that in fact a higher proportion of individuals in Control should be willing to switch teammates. This is because in Control, participants update in a symmetric way about their teammate, and end up with more moderate beliefs.[47] In Main, because of the positive bias in updating about the teammate, there is no corresponding increase in the proportion that should switch teammates. As was shown in Figure 3, this is indeed the case for actual participant decisions.

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



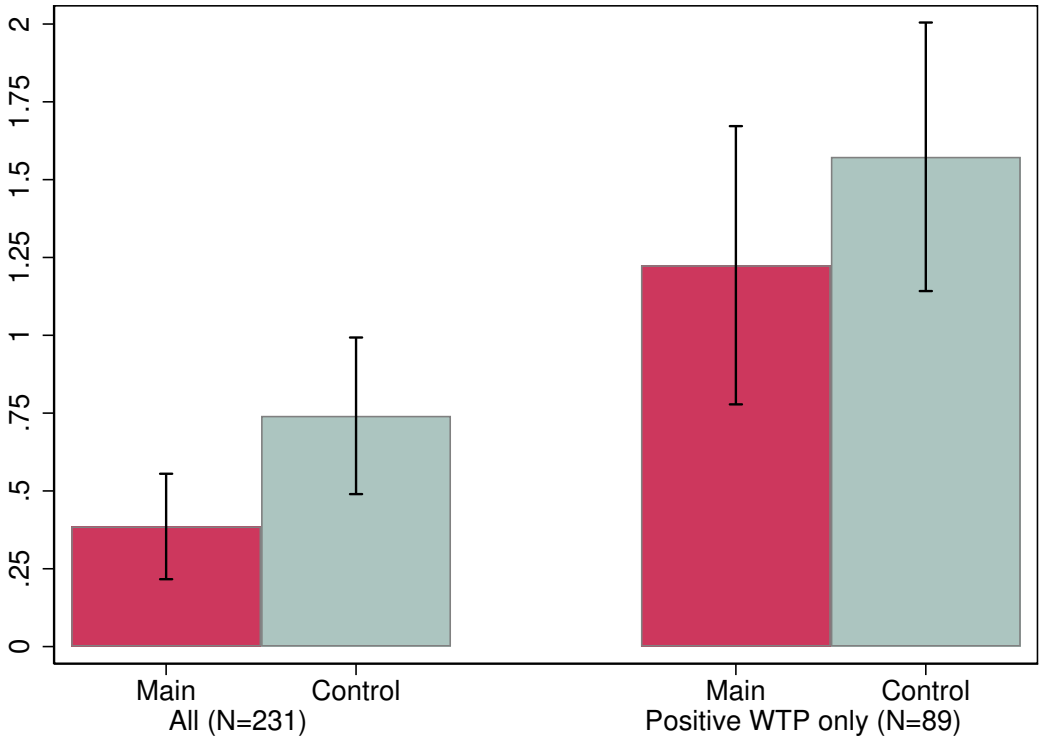
Given participant beliefs, this shows the proportion of participants that would (hypothetically) gain from switching teammates. 95% confidence intervals shown.

[46] 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.

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

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 (Wilcoxon ranksum p-value 0.0061). Restricting the sample only to positive WTP, the Wilcoxon ranksum p-value is 0.1321, $N = 89$. 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 Euros) of participants to switch teammate 2. Left side includes all data, right side includes only positive values of WTP. Wave 2 only. 95% confidence intervals shown.