THE ECONOMIC JOURNAL

The Economic Journal, **134** (*July*), 1835–1874 https://doi.org/10.1093/ej/ueae005 © The Author(s) 2024 Published by Oxford University Press on behalf of Royal Economic Society. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
Advance Access Publication Date: 5 March 2024

WHAT TO BLAME? SELF-SERVING ATTRIBUTION BIAS WITH MULTI-DIMENSIONAL UNCERTAINTY*

Alexander Coutts, Leonie Gerhards and Zahra Murad

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 experiment with a random fundamental replacing the teammate, unbiased updating occurs. A validation experiment shows that belief distortion is greater when outcomes originate from human actions. Overall, our experiments highlight how multi-dimensional environments can enable self-serving biases.

Researchers have amassed a wealth of evidence suggesting that people hold self-serving beliefs regarding personal traits such as ability, beauty or health (Eil and Rao, 2011; Oster *et al.*, 2013; Benoît *et al.*, 2015). 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]

* Corresponding author: Zahra Murad, Department of Accounting, Economics and Finance, Richmond Building, The University of Portsmouth, Portsmouth, PO1 2UP, UK. Email: zahra.murad@port.ac.uk

This paper was received on 20 October 2020 and accepted on 22 January 2024. The Editor was Steffen Huck.

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.

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] Specifically, benefits may arise from (*i*) direct utility from holding overconfident beliefs, for example arising from self-esteem or ego protection (Brunnermeier and Parker, 2005; Möbius *et al.*, 2022), (*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).

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

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 that can enable greater levels of self-serving beliefs.

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

[3] Engelmann *et al.* (2019) found 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 (Kunda, 1990; Bracha and Brown, 2012).

[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 by Kelley (1973) and Weiner (2010). Moreover, the study of self-serving biases in psychology is often framed as one of the 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

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

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 that optimally balances their teammate's performance (lower weight) and their own performance (higher weight). Accurate beliefs are incentivised because 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 replacing 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

work has focused on misattribution due to mood (Schwarz and Clore, 1983). Relatedly, outcome bias occurs when the randomness or luck inherent to outcomes disproportionately influences 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) studied outcome bias in the context of a principal-agent framework, and provided a useful review of this literature within economics.

[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.

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

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 that examined the extent of belief distortion across two between-subject 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 experiment. 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.

1. Experimental Design

1.1. *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 2017–8 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 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 10. We defer discussion of the follow-up and validation experiments to Section 5.

1.1.1. *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 ten minutes to complete a trivia and logic test consisting

[6] Wave 1 had 192 participants, while wave 2 had 234; recruitment aimed for gender balance, with 52% reported to be women. Experimental sessions in the first wave lasted approximately one 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 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. Because of a small glitch, some participants inadvertently skipped entering beliefs, which leaves us with 3,155 out of 3,170 observations.

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

Table 1. *Experimental Flow.*

Part 1	• IQ task (ten 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 four 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 four times:** • Receive feedback • Report posterior beliefs about teammate 1 and teammate 2 • Submit the weight

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]

Each correct answer would earn 2.5 points, while an incorrect answer would be penalised by one 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' identities 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, 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 ten (top half) or the bottom ten (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.

[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 4.1 below.

[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.

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

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.

1.1.2. *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 the control treatment, 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 the control treatment 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 consistently denote beliefs reported about own performance (in the main treatment) and teammate 1's performance (in the control treatment) as performance beliefs about teammate 1 and, similarly, denote beliefs reported about the teammate's performance (main treatment) and teammate 2's performance (control treatment) as performance beliefs about teammate 2.

1.2. *Weighting Decision and Belief Elicitation*

Participants were informed that their earnings from part 2 would depend on their team's performance that 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

[9] While access to information about the questions and answers is the same, we note that participants could have private information about their ability that affects 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 that such patterns would be consistent with overconfident beliefs. We thank two anonymous referees for raising these issues.

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

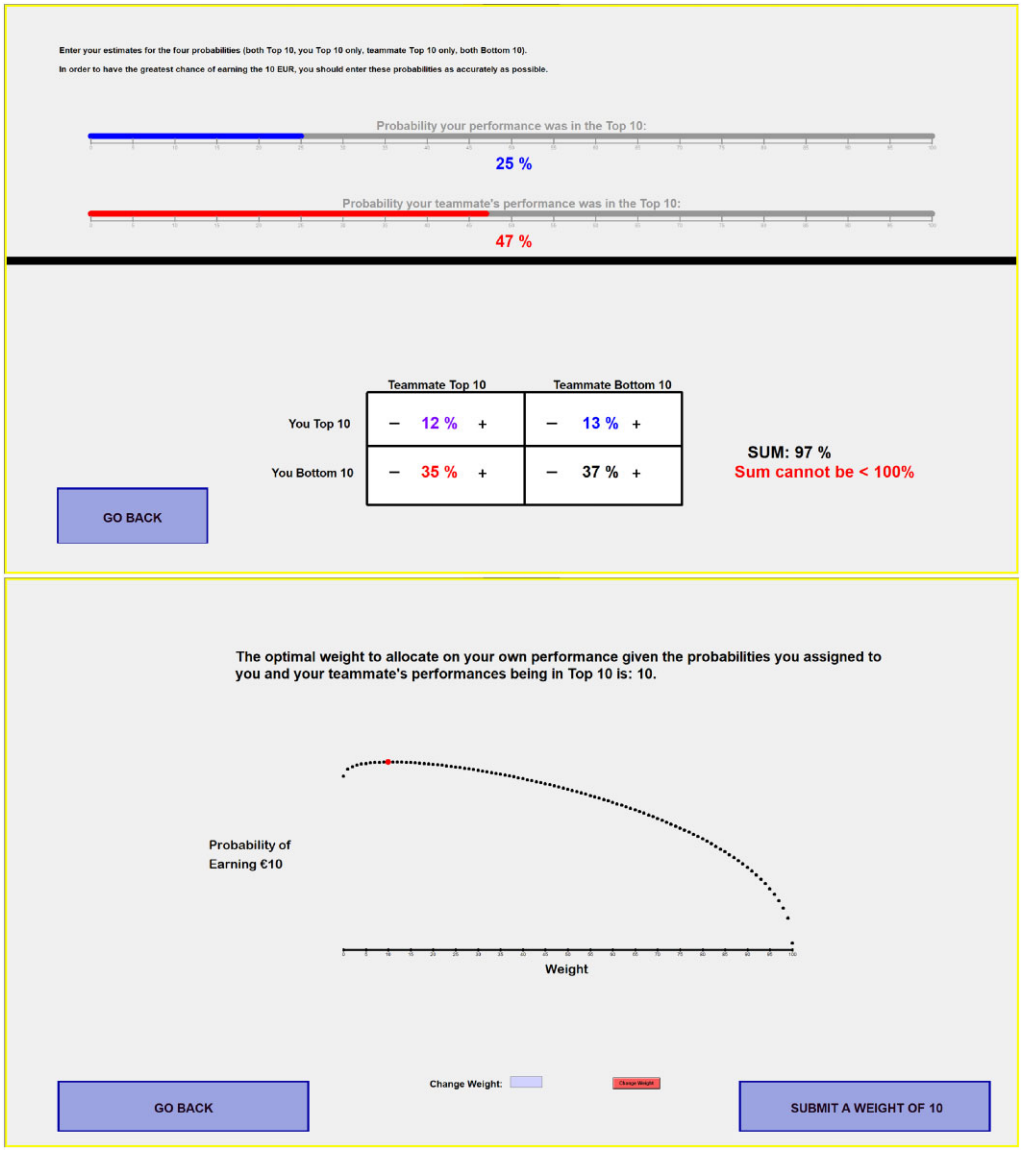

Fig. 1. *Screenshot of the Mapping from Chosen Weight to Probability of Winning €10 that Was Calculated for Every Participant, Conditional on the Beliefs They Entered.*

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 Online Appendix 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 with other elicitation procedures such as matching probabilities (Holt and Smith, 2009;

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

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 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 2.2 below. 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 1.

1.3. *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).

[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 by 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.

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

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 that 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 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.

1.4. *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 equal to or above the chosen price and keep their teammate if this maximum WTP is below that price. Our focus is on differences in WTP across the main and control treatments.

2. Theoretical Framework

2.1. *Preliminaries*

We first set-up our framework that follows from the experimental design. An individual faces an environment with two sources of uncertainty: (*i*) the ability of teammate 1 (own ability in the main treatment) 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

[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.

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

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:

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

2.2. *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 teammates 1 and 2 are in state T. Section 2.4 below describes 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

$$b_t^{TT} \cdot u(P) + b_t^{TB} \cdot \sqrt{\omega_t} \cdot u(P) + b_t^{TB} \cdot (1 - \sqrt{\omega_t}) \cdot u(0) + b_t^{BT} \cdot \sqrt{1-\omega_t} \cdot u(P) + b_t^{BT} \cdot (1 - \sqrt{1-\omega_t}) \cdot u(0) + b_t^{BB} \cdot u(0).$$

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

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

This leads to the optimal weight

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

Note that the optimal weight does not depend on the curvature of the utility function, $u(\cdot)$, and is hence 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.

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

Two observations are worth noting. First, given the functional form of the expected utility, the optimum in (1) 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 teammates 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]

2.3. *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[14]

$$[b_{t+1}^{1,BAYES} | s_t = p] = \frac{\Phi_{TT} b_t^{TT} + \Phi_{TB} b_t^{TB}}{\Phi_{TT} b_t^{TT} + \Phi_{TB} b_t^{TB} + \Phi_{BT} b_t^{BT} + \Phi_{BB} b_t^{BB}},$$

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

Analogously, for teammate 2,

$$[b_{t+1}^{2,BAYES} | s_t = p] = \frac{\Phi_{TT} b_t^{TT} + \Phi_{BT} b_t^{BT}}{\Phi_{TT} b_t^{TT} + \Phi_{TB} b_t^{TB} + \Phi_{BT} b_t^{BT} + \Phi_{BB} b_t^{BB}},$$

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

2.4. *Self-Serving Attribution Bias*

In this section we present an updating framework that maintains the structure of Bayes' rule, but allows for strategic misattribution of feedback across different sources. In our model, misattribu-

[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 by 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 (1) 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 the expected abilities of teammates 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 the expected ability of teammate 1 b_0^1.

[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 $(\Phi_{TT} b_t^{TT} + \Phi_{TB} b_t^{TB})/b_t^1$. The probability of being in the top half is b_t^1, and the perceived probability of receiving a signal $s_t = p$ is $\Phi_{TT} b_t^{TT} + \Phi_{TB} b_t^{TB} + \Phi_{BT} b_t^{BT} + \Phi_{BB} b_t^{BB}$.

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

tion will correspond directly to misperceiving 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 misattribution 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 that result from subsequent 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 that 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 positive (p) or negative (n) 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 that signal s is more likely to occur when the state is $A_1 = T$, while $\gamma_s^2 < 1$ would lead them to believe that 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

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

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 2.3, 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 AB) takes the following functional form for positive and negative signals, respectively:

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

Regarding updating about the teammate,

$$[b_{t+1}^{2,AB}|s_t = p] = \frac{\gamma_p^1\gamma_p^2\Phi_{TT}b_t^{TT} + \gamma_p^2\Phi_{BT}b_t^{BT}}{\gamma_p^1\gamma_p^2\Phi_{TT}b_t^{TT} + \gamma_p^1\Phi_{TB}b_t^{TB} + \gamma_p^2\Phi_{BT}b_t^{BT} + \Phi_{BB}b_t^{BB}},$$
$$[b_{t+1}^{2,AB}|s_t = n] = \frac{\gamma_n^1\gamma_n^2(1-\Phi_{TT})b_t^{TT} + \gamma_n^2(1-\Phi_{BT})b_t^{BT}}{\gamma_n^1\gamma_n^2(1-\Phi_{TT})b_t^{TT} + \gamma_n^1(1-\Phi_{TB})b_t^{TB} + \gamma_n^2(1-\Phi_{BT})b_t^{BT} + (1-\Phi_{BB})b_t^{BB}}. \quad (3)$$

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, the greater 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 that suggests that one might expect that teammate 2 is a likely target of negative misattribution, i.e., blaming teammate 2 that leads to more pessimistic beliefs about their performance. On the other hand, a positive misattribution 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:

$$\hat{\omega}_{t+1}^{*} = \frac{1}{1 + (\gamma_s^2 b_t^{BT}/\gamma_s^1 b_t^{TB})^2}. \quad (4)$$

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]

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

[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.

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

The optimal $\gamma_s^1 \geq 1$ and $\gamma_s^2 \leq \gamma_s^1$ are such that $[b_{t+1}^{1,AB}|s_t = s] \geq [b_{t+1}^{1,BAYES}|s_t = s]$, i.e., posteriors about own performance are biased upwards. However, whether the biased posterior for teammate 2, $[b_{t+1}^{2,AB}|s_t = s]$, is smaller, equal to or larger than the Bayesian $[b_{t+1}^{2,BAYES}|s_t = s]$ 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 (Eil and Rao, 2011; Ertac, 2011; Grossman and Owens, 2012; Buser *et al.*, 2018; Coutts, 2019a; 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.

3. Hypotheses

The theoretical model compares belief updating to a benchmark in which updating follows Bayes' rule (Section 2). 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 refer to the Bayesian benchmark.

3.1. *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 participant's 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 treatment than in the control treatment ($b_0^{1,M} > b_0^{1,C}$).

3.2. *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

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

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

signals, respectively:

$$\frac{b^i_{t+1}}{1-b^i_{t+1}} = \frac{b^i_t}{1-b^i_t} \cdot LR^i_t(s). \tag{5}$$

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

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

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

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

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

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

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

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

Denote the Bayesian likelihood ratios, calculated by setting $\gamma^i_s = 1$, by $LR^i_t(s)$. Inserting the perceived likelihood ratio into (5), taking natural logarithms of both sides and adding an indicator function $I\{s_t = s\}$ for the type of signal observed, we obtain

$$\text{logit}(b^i_{t+1}) = \text{logit}(b^i_t) + I\{s_t = p\} \ln[\hat{LR}^i_t(p)] + I\{s_t = n\} \ln[\hat{LR}^i_t(n)].$$

The empirical model nests this Bayesian benchmark as

$$\begin{aligned}\text{logit}(b^i_{j,t+1}) = \delta \text{logit}(b^i_{j,t}) &+ \beta_1 I(s_{j,t} = p) \ln[\hat{LR}^i_t(p)] \\ &+ \beta_0 I(s_{j,t} = n) \ln[\hat{LR}^i_t(n)] + \epsilon_{j,t+1}.\end{aligned} \tag{6}$$

Here 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. Parameter $\epsilon_{j,t+1}$ captures non-systematic errors, noting the use of j to identify the experimental subject.

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

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

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

While Bayesian posteriors result in a weight of $\beta_1 = 1$ or $\beta_0 = 1$ on $LR_t^1(p)$ or $LR_t^1(n)$, respectively, what are the implications of self-serving attribution bias for this framework? First note that $\hat{LR}_t^1(p) \geq LR_t^1(p)$ and $\hat{LR}_t^1(n) \geq LR_t^1(n)$. 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 ($\beta_1^{1,M} = \beta_1^{1,C}$; $\beta_0^{1,M} = \beta_0^{1,C}$ and $\beta_1^{2,M} = \beta_1^{2,C}$; $\beta_0^{2,M} = \beta_0^{2,C}$).[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 the main treatment compared to the control treatment ($\beta_1^{1,M} > \beta_1^{1,C}$; $\beta_0^{1,M} < \beta_0^{1,C}$).

Furthermore, updating beliefs about teammate 2 is biased:

Positive bias: individuals over-weight positive and under-weight negative signals about teammate 2 in the main compared to the control *treatment* ($\beta_1^{2,M} > \beta_1^{2,C}$; $\beta_0^{2,M} < \beta_0^{2,C}$).

Or negative bias: individuals under-weight positive and over-weight negative signals about teammate 2 in the main compared to the control treatment ($\beta_1^{2,M} < \beta_1^{2,C}$; $\beta_0^{2,M} > \beta_0^{2,C}$).

[19] This implication simultaneously explains the intuition for why the γ_s^i are distorted in a way that 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] Parameter β_1 is biased upwards because, since $\ln(\hat{LR}_t^1(p)) \geq 0$, a Bayesian response to $\hat{LR}_t^1(p)$ will manifest itself as an over-response to the smaller unbiased $LR_t^1(p)$. Parameter β_0 is biased downwards because $\ln(\hat{LR}_t^1(n)) \leq 0$, so a Bayesian response to $\hat{LR}_t^1(n)$ will manifest itself as an under-response to the smaller (more negative, i.e., larger in absolute value) $LR_t^1(n)$.

[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 misattributions 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 Online Appendix 8. While they generate stark predictions, neither is able to explain our results, in part due to their rigidity.

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

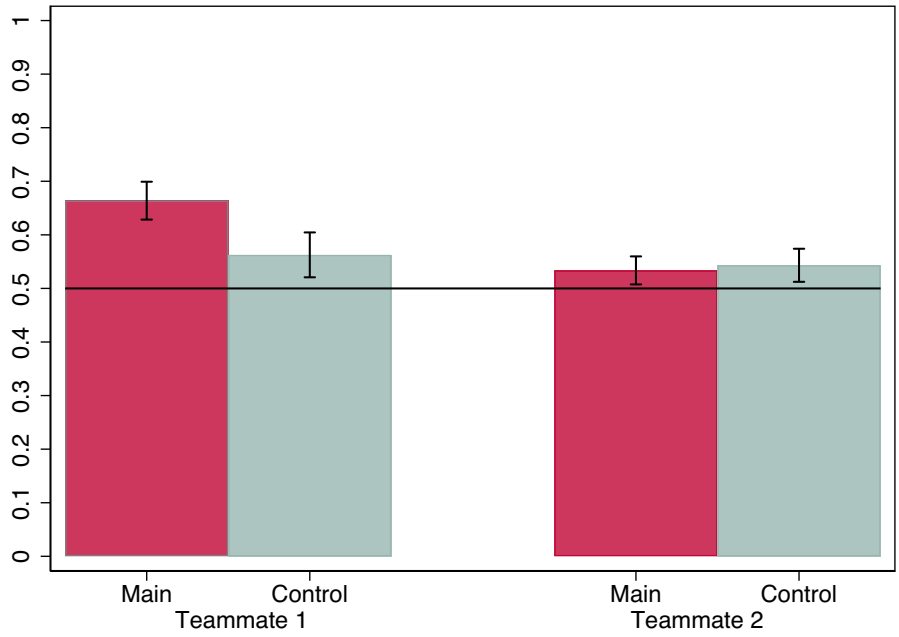

Fig. 2. *Prior Beliefs by Treatment.*
Notes: For teammate 1: In the main treatment individuals estimate beliefs about their own performance, whereas in the control treatment individuals estimate beliefs about other teammate 1's performance. For teammate 2: in both treatments individuals estimate beliefs about other teammate 2's performance. We include 95% confidence intervals.

4. Results

4.1. *Initial Beliefs*

Figure 2 presents the first-round beliefs in the 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 treatments, respectively. These beliefs are also significantly different from 50% (Wilcoxon signed-rank test p-values 0.0012 and 0.0017, respectively).

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

[22] For those individuals in the top half, 83% hold prior beliefs greater than 50% (compared to 76% in the control treatment). For those in the bottom half, 54% hold prior beliefs greater than 50% (compared to 29% in the control treatment). 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 5.

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

the main or control treatment, 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 the control treatment and teammate 2 in the main and control treatments (p-value 0.2654). Also, there are no significant differences in initial beliefs about teammate 2 between the main and control treatments (Wilcoxon rank sum p-value 0.5723).

On the other hand, when we test Hypothesis 1 and compare initial beliefs about teammate 1 across the two treatments, main (self) and control (other), we can clearly reject equality of beliefs (Wilcoxon rank sum test p-value 0.0005). The results are thus in line with Hypothesis 1. This provides robust evidence that what we are observing in the main treatment does reflect true overconfidence. It further suggests that 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 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 3).[23]

4.2. *Belief Updating*

To study self-serving attribution bias discussed in Section 2 and to test the hypotheses from Section 3, we use (6) for our primary empirical analysis. Later, in Section 4.2.2 we investigate updating biases taking a non-parametric approach, free of structural assumptions. This allows us to statistically distinguish posteriors in the main treatment versus the control treatment, accounting for differences in initial priors, utilising a matching strategy. Moreover, we discuss individuals' WTP to be matched to a new teammate 2 in Section 4.3. For the interested reader, we present an additional analysis of the resulting weights in Online Appendix 4, and examine the average evolution of beliefs in Online Appendix 5 by treatment.

4.2.1. *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 parts 2 and 3. Samples excluding part 3 are presented in Online Appendix 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 6.

Updating is not Bayesian in either the main or control treatment. All coefficients in Table 2 are significantly different from the Bayesian prediction of 1, indicated by asterisks. Column (1)

[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 3.

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

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

Notes: Analysis uses OLS regression. Difference is *significant from* 1 at *** 0.01. Robust SEs are clustered at the individual level. Here δ is the coefficient on the log prior odds ratio; β_1 and β_0 are coefficients on the loglikelihood 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.

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 the main treatment than in the control treatment. 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 the main treatment than in the control treatment. 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.*

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 the main treatment compared to the control treatment.

To identify which of these patterns are visible, Table 3 presents belief regressions for teammate 2 in the main treatment (column (1)) and the control treatment (column (2)) that are analogous to those 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 a member of

[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).

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

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

Notes: Analysis uses OLS regression. Difference is *significant from* 1 at *** 0.01. Robust SEs are clustered at the individual level. Here δ is the coefficient on the log prior odds ratio; β_1 and β_0 are coefficients on the loglikelihood 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.

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 2 and detailed in Appendix A, some positively biased updating about teammate 2 can be optimal since it permits self-serving beliefs, while reducing the material costs of such beliefs, due to more moderate weighting between the two teammates. Interestingly, for positive signals, $\beta_1^{1,M}$ in Table 2, column (1) is significantly greater than $\beta_1^{2,M}$ in Table 3, column (1) (Chow test p-value 0.0062). For negative signals, the respective $\beta_0^{1,M}$ and $\beta_0^{2,M}$ coefficients do not differ significantly (Chow test p-value 0.8637). Taken together, the difference in asymmetry ($\beta_1^{1,M} - \beta_0^{1,M}$) versus ($\beta_1^{2,M} - \beta_0^{2,M}$) across the first columns in Tables 2 and 3 is significant at the 10% level (Chow test p-value 0.0963).[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 7: first, that anchoring causes individuals to update similarly about teammate 2, second that participants

[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).

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

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 5.1.2, this is not the case—instead, updating is not self-serving, which suggests that 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 the main and control treatments. 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.

4.2.2. *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 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 in the same direction, but are smaller in magnitude and not statistically significant at conventional levels.*

4.3. *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 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 also depends 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

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

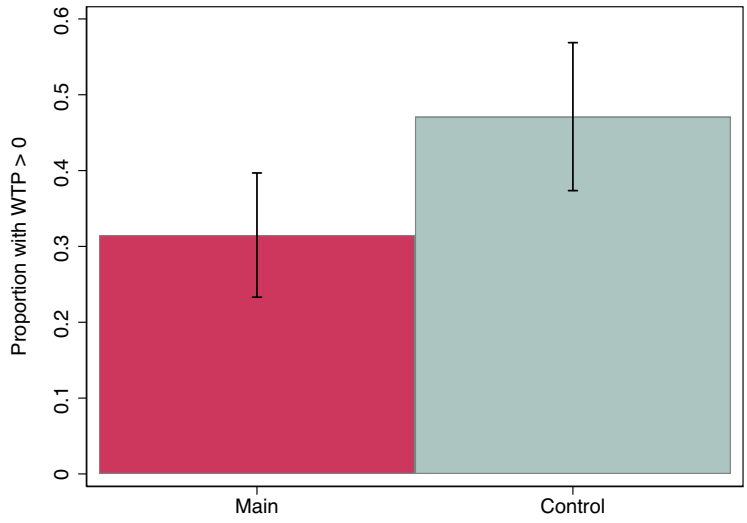

Fig. 3. *Willingness to Switch.*
Notes: Proportion of participants in wave 2 only ($N = 231$) who submitted strictly positive WTP to change teammate 2. We include 95% confidence intervals.

participants in the main treatment 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. Thirty-one percent 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 treatment 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 that falls outside of the purview of the elicitation procedure.

The fact that self-serving motives can motivate distorted beliefs about others that 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) showed the opposite—overconfident individuals should be less satisfied with their environments, and therefore more likely to change them. As a result, they showed that overconfidence should not persist in the long run, though

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

under-confidence 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 under-confidence (Dunning, 2005).[26]

5. Follow-up and Validation Experiments

5.1. *Follow-up*

5.1.1. *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.

The follow-up experiment sessions were conducted in person using the same software and participant pool at the University of Hamburg during the 2021–2 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 that prior beliefs about the random fundamental were similar to beliefs about a human teammate, participants were given a range that 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]

5.1.2. *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 rank sum test p-value 0.6665). Average prior beliefs for the random

[26] In their model of misguided learning with multi-dimensional uncertainty, Heidhues *et al.* (2018) also made 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.

[27] Recruitment aimed for gender balance, with 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 that 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).

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

Table 4. *Updating Beliefs in the Follow-Up Experiment.*

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
Compared to control treatment in Column (2) of:	Table 2	Table 3
p-value [Chow test] for δ	0.1488	0.3779
p-value [Chow test] for β_1	0.9184	0.1266
p-value [Chow test] for β_0	0.9948	0.4666
p-value [Chow test] for ($\beta_1 - \beta_0$)	0.9345	0.4292

Notes: Analysis uses OLS regression. Difference is *significant from* 1 at *** 0.01. Robust SEs are clustered at the individual level. Here δ is the coefficient on the log prior odds ratio; β_1 and β_0 are coefficients on the loglikelihood 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. Bottom four rows conduct Chow tests comparing results with the analogous control treatment: Column (1) is compared to column (2) in Table 2; Column (2) is compared to column (2) in Table 3.

fundamental are 54.5% (53.4% for the human teammate 2 in the main treatment, Wilcoxon rank sum test p-value 0.8895).

Table 4 presents the same specifications from 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)) reveals 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 the 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 cognitive 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

[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.

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

belief distortion, we conducted an online validation experiment, described in detail in the next subsection.

5.2. *Validation Experiment*

5.2.1. *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 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 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 the human treatment (the random fundamental in the RF treatment) 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 (one 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 one 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 9.

Following our theoretical framework, we assume that 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 treatment than in the RF treatment.

5.2.2. *Results*

As expected, participants' initial beliefs about their matched human/RF being top/high were nearly identical between the two treatments: they believed their match was top (human) or high (RF) with probabilities of 55.1% and 55.0%, respectively (Wilcoxon rank sum 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% (χ^2 test p-value 0.0200).

[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.

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

The average absolute revision was 16.1 versus 14.6 percentage points (Wilcoxon rank sum p-value 0.0653) for the human versus RF treatments, respectively. Hence, the main results of the validation experiment 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]

6. 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) established 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 that 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 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

[31] It is more difficult to shed light on the mechanisms underlying the result of greater distortion when matched with a human. Online Appendix 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 shuts down the link between actions and feedback, which drives the self-defeating learning they study.

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

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 4.2.1 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 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

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

[34] Hestermann and Le Yaouanq (2021) showed that, with Bayesian updating, under-confidence, 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.

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

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.

Appendix A. Model of Optimal Information Distortion

In this appendix 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 that 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 (2) and (3).

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] By $\alpha \in [0, \infty)$ we indicate 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 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 two states, $\hat{b}_{t+1}^{TB}$ and $\hat{b}_{t+1}^{BT}$. Given the form of the bias for updating about own ability, this will imply an over-weighting of the likelihood of state TB by γ_s^1, and an over- or under-weighting of the likelihood of state BT by γ_s^2.

Under this formulation we present again the resulting biased posterior beliefs for teammates 1 and 2, as shown in (2) and (3). We show the case for a positive signal, noting that the results are

[35] We choose this for simplicity, though our results would hold for both concave belief value functions, as well as 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 by Bracha and Brown (2012). Finally, experimental evidence suggests that such mental costs are necessary if one wishes to take models of belief distortion seriously (Coutts, 2019a; Engelmann *et al.*, 2019).

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

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

unchanged by replacing $\Phi_{A_1A_2}$ by the negative signal equivalent $1 - \Phi_{A_1A_2}$:

$$[\hat{b}^1_{t+1}|s_t = p] = \frac{\gamma^1_p\gamma^2_p\Phi_{TT}b^{TT}_t + \gamma^1_p\Phi_{TB}b^{TB}_t}{\gamma^1_p\gamma^2_p\Phi_{TT}b^{TT}_t + \gamma^1_p\Phi_{TB}b^{TB}_t + \gamma^2_p\Phi_{BT}b^{BT}_t + \Phi_{BB}b^{BB}_t},$$

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

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 $[\hat{b}^1_{t+1}|s_t = p] = 1/(1 + x_2/x_1)$. Taking the derivative with respect to γ^1_p,

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

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

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

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^2_p,

$$\begin{aligned}\frac{\partial[\hat{b}^1_{t+1}|s_t = p]}{\partial\gamma^2_p} &= \frac{1}{(1 + x_2/x_1)^2} \cdot \frac{x_2}{x_1^2} \cdot (\gamma^1_p\Phi_{TT}b^{TT}_t) - \frac{1}{(1 + x_2/x_1)^2} \cdot \frac{1}{x_1} \cdot (\Phi_{BT}b^{BT}_t) \\ &= \frac{1}{x_1^2(1 + x_2/x_1)^2} \cdot (x_2 \cdot \gamma^1_p\Phi_{TT}b^{TT}_t - x_1 \cdot \Phi_{BT}b^{BT}_t) \\ &= \frac{1}{x_1^2(1 + x_2/x_1)^2} \cdot ((\gamma^2_p\Phi_{BT}b^{BT}_t + \Phi_{BB}b^{BB}_t) \cdot \gamma^1_p\Phi_{TT}b^{TT}_t \\ &\qquad - (\gamma^1_p\gamma^2_p\Phi_{TT}b^{TT}_t + \gamma^1_p\Phi_{TB}b^{TB}_t) \cdot \Phi_{BT}b^{BT}_t) \\ &= \frac{\gamma^1_p}{x_1^2(1 + x_2/x_1)^2} \cdot (\Phi_{TT}b^{TT}_t \cdot \Phi_{BB}b^{BB}_t - \Phi_{TB}b^{TB}_t \cdot \Phi_{BT}b^{BT}_t) \\ &< 0.\end{aligned}$$

Given our specification of the signal structure $\Phi_{A_1A_2}$, $\Theta = \Phi_{TT}b^{TT}_t \cdot \Phi_{BB}b^{BB}_t - \Phi_{TB}b^{TB}_t \cdot \Phi_{BT}b^{BT}_t < 0$, as detailed in Appendix B. Hence, $\partial[\hat{b}^1_{t+1}|s_t = p]/\partial\gamma^2_p < 0$, and similarly for γ^2_n.

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

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

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

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's performance, $\hat{b}^2_{t+1}$. 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 (1), setting $\Phi_{BT} = \Phi_{TB} = 0.5$, we have[38]

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

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

$$\begin{aligned}\max_{\{\gamma_s\}} \Big\{ &\alpha \cdot \hat{b}^1_{t+1} + b^{TT}_{t+1} \cdot u(P) + b^{TB}_{t+1} \cdot \sqrt{\hat{\omega}^*_{t+1}} \cdot u(P) + b^{TB}_{t+1} \cdot \left(1 - \sqrt{\hat{\omega}^*_{t+1}}\right) \cdot u(0) \\ &+ b^{BT}_{t+1} \cdot \sqrt{1 - \hat{\omega}^*_{t+1}} \cdot u(P) + b^{BT}_{t+1} \cdot \left(1 - \sqrt{1 - \hat{\omega}^*_{t+1}}\right) \cdot u(0) + b^{BB}_{t+1} \cdot u(0) \\ &- J_1(\gamma_s^1, 1) - J_2(\gamma_s^2, 1)\Big\}.\end{aligned}$$

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

By the properties of the mental cost function $J_i(\gamma_s^i, 1)$, extreme values of γ_s^i are never optimal, and thus we restrict our attention to an interior solution. We also restrict our focus to solutions with $\gamma_s^1 \geq 1$, 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

[38] We note that, given the biased updating process, this is simplified from the following equation (analogously for a negative signal):

$$\frac{\hat{b}^{BT}_{t+1}}{\hat{b}^{TB}_{t+1}} = \frac{\gamma_p^2 \Phi_{BT} b_t^{BT} / (\gamma_p^1 \gamma_p^2 \Phi_{TT} b_t^{TT} + \gamma_p^1 \Phi_{TB} b_t^{TB} + \gamma_p^2 \Phi_{BT} b_t^{BT} + \Phi_{BB} b_t^{BB})}{\gamma_p^1 \Phi_{TB} b_t^{TB} / (\gamma_p^1 \gamma_p^2 \Phi_{TT} b_t^{TT} + \gamma_p^1 \Phi_{TB} b_t^{TB} + \gamma_p^2 \Phi_{BT} b_t^{BT} + \Phi_{BB} b_t^{BB})} = \frac{\gamma_p^2 \Phi_{BT} b_t^{BT}}{\gamma_p^1 \Phi_{TB} b_t^{TB}}.$$

[39] Note that self-serving beliefs can arise from setting $\gamma_s^1 > 1$ or $\gamma_s^2 < 1$. Regarding the latter case, while unlikely in our setting, it does not preclude that $\gamma_s^1 < 1$. 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 $\gamma_s^2 \geq \gamma_s^1/2$, or $\gamma_s^2 \geq \frac{1}{2}$ to rule out $\gamma_s^1 < 1$.

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

first-order condition with respect to γ_s^1 is (where $u(P) - u(0) = \Delta u$)

$$\alpha \cdot \frac{\partial[\hat{b}_{t+1}^1 | s_t]}{\partial \gamma_s^1} + \frac{\gamma_s^2 \cdot (b_{t+1}^{TB} \cdot b_{t+1}^{BT})^2 \cdot \Delta u}{((\gamma_s^2 b_{t+1}^{BT})^2 + (\gamma_s^1 b_{t+1}^{TB})^2)^{3/2}} \cdot (\gamma_s^2 - \gamma_s^1) - J_i'(\gamma_s^1, 1).$$

The first order-condition with respect to γ_s^2 is

$$\alpha \cdot \frac{\partial[\hat{b}_{t+1}^1 | s_t]}{\partial \gamma_s^2} + \frac{\gamma_s^1 \cdot (b_{t+1}^{TB} \cdot b_{t+1}^{BT})^2 \cdot \Delta u}{((\gamma_s^2 b_{t+1}^{BT})^2 + (\gamma_s^1 b_{t+1}^{TB})^2)^{3/2}} \cdot (\gamma_s^1 - \gamma_s^2) - J_i'(\gamma_s^2, 1).$$

RESULT A.1. When $\alpha = 0$, there will be no belief distortion.

This result derives directly from setting the two first-order conditions (FOCs) equal to zero. When $\alpha = 0$, the unique optimal solution is to set $\gamma_s^1 = \gamma_s^2 = 1$.

RESULT A.2. It holds that $\gamma_s^1 \geq \gamma_s^2$.

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

If $\alpha = 0$, the optimal $\gamma_s^1 = \gamma_s^2 = 1$. When $\alpha > 0$, $\gamma_s^1 > 1$, while the optimal γ_s^2 may be less than, equal to or greater than 1, though $\gamma_2^s < \gamma_1^s$. The reason why γ_s^2 is not unambiguously smaller than one is that there is a benefit to updating in a biased way about teammate 2, which counterbalances the biased updating about teammate 1, leading to a closer-to-optimal 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 $\partial[\hat{b}_{t+1}^2 | s_t = s]/\partial \gamma_s^2 > 0$ and $\partial[\hat{b}_{t+1}^2 | s_t = s]/\partial \gamma_s^1 < 0$. 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]

Appendix B. Deriving the Condition for $\Theta < 0$

B.1. *Theoretical Result*

In this appendix 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 $0.9 \cdot 0.1 - 0.5 \cdot 0.5 = -0.16 < 0$.

[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 is 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.

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

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

$$\begin{aligned}&\Phi_{TT}(b_0^1 \cdot b_0^2) \cdot \Phi_{BB}((1 - b_0^1) \cdot (1 - b_0^2)) - \Phi_{TB}(b_0^1 \cdot (1 - b_0^2)) \cdot \Phi_{BT}((1 - b_0^1) \cdot b_0^2) \\ &\quad = (b_0^1 \cdot b_0^2)((1 - b_0^1) \cdot (1 - b_0^2)) \cdot [\Phi_{TT} \cdot \Phi_{BB} - \Phi_{TB} \cdot \Phi_{BT}]. \qquad \text{(B1)}\end{aligned}$$

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. We have

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

We note that the denominator of beliefs for all four states will be identical. Denote it by $\mathcal{D}_1 = \gamma_p^1 \gamma_p^2 \Phi_{TT} \cdot b_0^{TT} + \gamma_p^1 \Phi_{TB} \cdot b_0^{TB} + \gamma_p^2 \Phi_{BT} \cdot b_0^{BT} + \Phi_{BB} \cdot b_0^{BB}$. We now substitute these expressions for the four states back into the initial expression of interest, (B1):

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

We now note that this is simply an iteration of (B1). As such, it reduces to

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

We continue this inductive process once more:

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

We define $\mathcal{D}_2 = \gamma_p^1 \gamma_p^2 \Phi_{TT} \cdot b_1^{TT} + \gamma_p^1 \Phi_{TB} \cdot b_1^{TB} + \gamma_p^2 \Phi_{BT} \cdot b_1^{BT} + \Phi_{BB} \cdot b_1^{BB}$, and so

$$[b_2^{TT} | s_t = p] = \frac{\gamma_p^1 \gamma_p^2 \Phi_{TT} \cdot \gamma_p^1 \gamma_p^2 \Phi_{TT} b_0^{TT} / \mathcal{D}_1}{\mathcal{D}_2} = \frac{(\gamma_p^1 \gamma_p^2 \Phi_{TT})^2 \cdot b_0^{TT}}{\mathcal{D}_1 \cdot \mathcal{D}_2}.$$

Thus, we arrive at the third term

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

Following this process, assume that the kth posterior is given by

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

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

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

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

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 this expression is positive.

Appendix C. Theoretical Alternative: Negative Attributions

Our results in Section 2.4 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:

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

Analogous to the earlier distorted weight shown in (4), the optimal weight is distorted by the parameters γ_s^1 and γ_s^2:[41]

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

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 in 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 counterbalancing the material costs of submitting 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.

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

[42] As shown in (4), the incentives in our experiment lead to distortion whenever $\gamma_s^2/\gamma_s^1 \neq 1$, which generate incentives to counterbalance positive self-attributions ($\gamma_s^1 > 1$) with positive other attributions ($\gamma_s^2 > 1$).

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

Table D.1. *Main versus Control Treatment: Belief Teammate 1 is in the Top Half.*

	(1) One neighbour	(2) Two neighbours
ATE	0.075**	0.065**
	(0.032)	(0.028)
Observations	373	373

Notes: Analysis uses nearest-neighbour matching, with replacement when > 1 neighbour. Significantly different from zero at ** 0.05. Abadie-Imbens robust SEs are reported in parentheses. All matches received the exact same distribution of signals.

Appendix 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 (ATEs). 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.

Our structural analysis suggests that 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 four negative signals (zero positive) turns out to reveal the greatest difference between the main versus control treatment: 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 two positive and two 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 treatment), analogous

[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.

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

Table D.2. *Main versus Control Treatment: Belief Teammate 1 is in the Top Half by the 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

Notes: Analysis uses nearest-neighbour matching with one neighbour. Significantly different from zero at ** 0.05; *** 0.01. Abadie-Imbens robust SEs are reported in parentheses. Each column restricts the sample to a specific distribution of negative signals received (out of four total signals).

Table D.3. *Main versus Control Treatment: Belief Teammate 2 is in the Top Half.*

	(1) One neighbour	(2) Two neighbours
ATE	0.052	0.049
	(0.037)	(0.033)
Observations	376	376

Notes: Analysis uses nearest-neighbour matching, with replacement when > 1 neighbour. Abadie-Imbens robust SEs are reported in parentheses. All matches received the exact same distribution of signals.

Table D.4. *Main versus Control Treatment: Belief Teammate 2 is in the Top Half by the Proportion of Negative Signals Received.*

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

Notes: Analysis uses nearest-neighbour matching with one neighbour. Significantly different from zero at ** 0.05. Abadie-Imbens robust SEs are reported in parentheses. Each column restricts the sample to a specific distribution of negative signals received (out of four total signals).

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 the main relative to the control treatment; however, this is not statistically significant at conventional levels (respective p-values of 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 the main and control treatments who receive only negative signals end up with posteriors about teammate 2 that are approximately 14 percentage points greater in the main relative to the control treatment. Again, this supports our structural results.

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

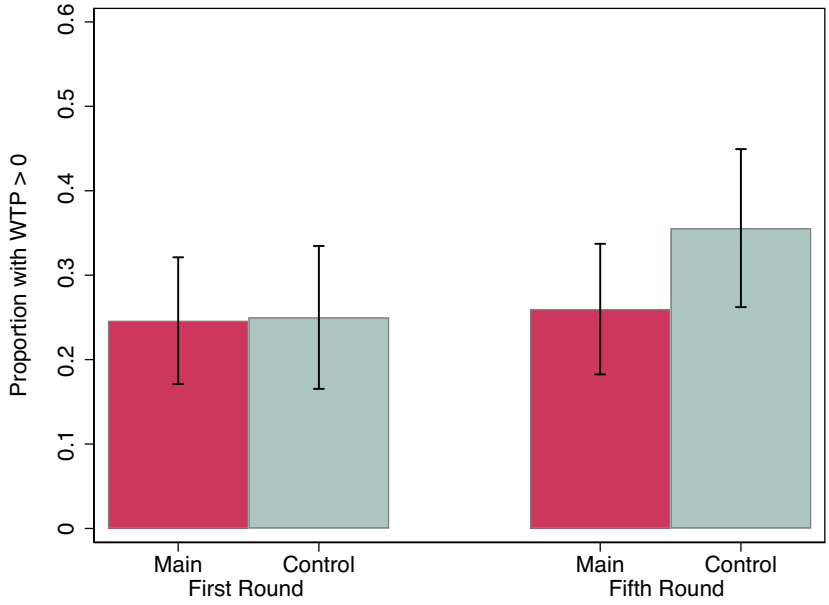

Fig. E.1. *(Calculated) Optimal Proportion Willing to Switch.*
Notes: Given participant beliefs, this shows the proportion of participants that would (hypothetically) gain from switching teammates. We include 95% confidence intervals.

Appendix 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, 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 con-

[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.

[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.

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

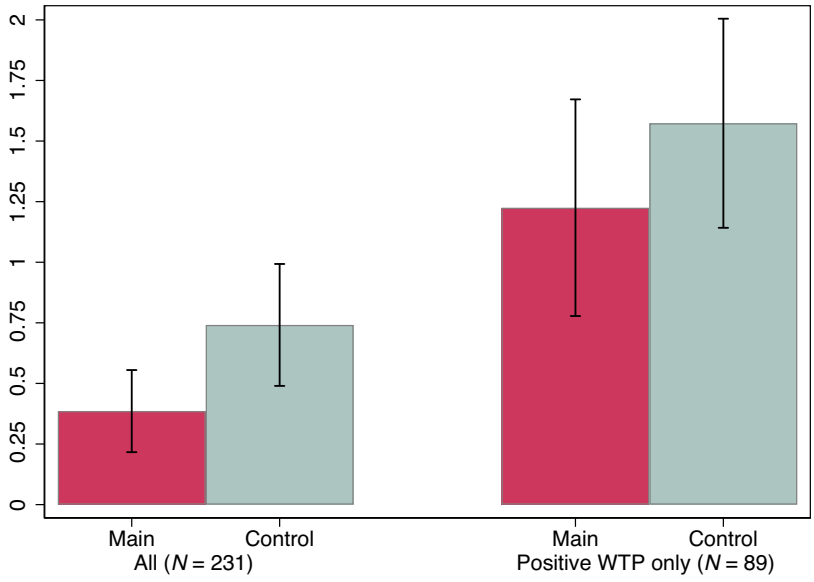

Fig. E.2. *Willingness to Pay.*
Notes: WTP (in euros) of participants (in wave 2 only) to be matched to a new teammate 2. Left side includes all data; right side includes only positive values of WTP. We include 95% confidence intervals.

firms this is the case given prior beliefs in the main and control treatments (round 1). This figure plots the theoretically optimal proportion of participants that 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 the control treatment should be willing to switch teammates. This is because, in the control treatment, participants update in a symmetric way about their teammate, and end up with more moderate beliefs.[47] In the main treatment, 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.2 presents the actual values of WTP submitted. The average WTP in the main treatment is €0.39, while in the control treatment it is €0.74, significantly different at the 1% level (Wilcoxon rank sum p-value 0.0061). Restricting the sample to only positive WTP, the Wilcoxon rank sum p-value is 0.1321, $N = 89$. Thus, while there is lower WTP among this restricted sample in the main treatment relative to the control treatment, this can be accounted for by the more overconfident beliefs in the main treatment, for which there is less material benefit to having a new teammate.

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

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

York University, Canada
King's College London, UK
The University of Portsmouth, UK & Azerbaijan State University of Economics (UNEC), Azerbaijan

Additional Supporting Information may be found in the online version of this article:

Online Appendix
Replication Package

References

Azrieli, Y., Chambers, C.P. and Healy, P.J. (2018). 'Incentives in experiments: A theoretical analysis', *Journal of Political Economy*, vol. 126(4), pp. 1472–503.
Baldiga, K. (2014). 'Gender differences in willingness to guess', *Management Science*, vol. 60(2), pp. 434–48.
Baron, J. and Hershey, J.C. (1988). 'Outcome bias in decision evaluation', *Journal of Personality and Social Psychology*, vol. 54(4), pp. 569–79.
Barron, K. (2021). 'Belief updating: Does the "good-news, bad-news" asymmetry extend to purely financial domains?', *Experimental Economics*, vol. 24(1), pp. 31–58.
Bartling, B. and Fischbacher, U. (2012). 'Shifting the blame: On delegation and responsibility', *The Review of Economic Studies*, vol. 79(1), pp. 67–87.
Becker, G.M., Degroot, M.H. and Marschak, J. (1964). 'Measuring utility by a single-response sequential method', *Behavioral Science*, vol. 9(3), pp. 226–32.
Bénabou, R. and Tirole, J. (2002). 'Self-confidence and personal motivation', *The Quarterly Journal of Economics*, vol. 117(3), pp. 871–915.
Bénabou, R. and Tirole, J. (2009). 'Over my dead body: Bargaining and the price of dignity', *American Economic Review*, vol. 99(2), pp. 459–65.
Bénabou, R. and Tirole, J. (2011). 'Identity, morals, and taboos: Beliefs as assets', *The Quarterly Journal of Economics*, vol. 126(2), pp. 805–55.
Benjamin, D.J. (2019). 'Errors in probabilistic reasoning and judgment biases', in (B.D. Bernheim, S. DellaVigna and D. Laibson, eds.), *Handbook of Behavioral Economics: Applications and Foundations 1*, vol. 2, pp. 69–186, Amsterdam: North-Holland.
Benoît, J.P. and Dubra, J. (2011). 'Apparent overconfidence', *Econometrica*, vol. 79(5), pp. 1591–625.
Benoît, J.P., Dubra, J. and Moore, D.A. (2015). 'Does the better-than-average effect show that people are overconfident?: Two experiments', *Journal of the European Economic Association*, vol. 13(2), pp. 293–329.
Bracha, A. and Brown, D.J. (2012). 'Affective decision making: A theory of optimism bias', *Games and Economic Behavior*, vol. 75(1), pp. 67–80.
Brownback, A. and Kuhn, M.A. (2019). 'Understanding outcome bias', *Games and Economic Behavior*, vol. 117, pp. 342–60.
Brunnermeier, M.K. and Parker, J.A. (2005). 'Optimal expectations', *American Economic Review*, vol. 95(4), pp. 1092–118.
Burks, S.V., Carpenter, J.P., Goette, L. and Rustichini, A. (2013). 'Overconfidence and social signalling', *Review of Economic Studies*, vol. 80(3), pp. 949–83.
Buser, T., Gerhards, L. and Van Der Weele, J. (2018). 'Responsiveness to feedback as a personal trait', *Journal of Risk and Uncertainty*, vol. 56, pp. 165–92.
Campbell, W.K. and Sedikides, C. (1999). 'Self-threat magnifies the self-serving bias: A meta-analytic integration', *Review of General Psychology*, vol. 3(1), pp. 23–43.
Charness, G. and Dave, C. (2017). 'Confirmation bias with motivated beliefs', *Games and Economic Behavior*, vol. 104, pp. 1–23.
Coutts, A. (2019a). 'Good news and bad news are still news: Experimental evidence on belief updating', *Experimental Economics*, vol. 22(2), pp. 369–95.
Coutts, A. (2019b). 'Testing models of belief bias: An experiment', *Games and Economic Behavior*, vol. 113, pp. 549–65.
Drobner, C. (2022). 'Motivated beliefs and anticipation of uncertainty resolution', *American Economic Review: Insights*, vol. 4(1), pp. 89–105.
Drobner, C. and Goerg, S.J. (2022). 'Motivated belief updating and rationalization of information', Preprint, http://dx.doi.org/10.2139/ssrn.4268790.
Dunning, D. (2005). *Self-Insight: Roadblocks and Detours on the Path to Knowing Thyself*, New York: Psychology Press.
Eberlein, M., Ludwig, S. and Nafziger, J. (2011). 'The effects of feedback on self-assessment', *Bulletin of Economic Research*, vol. 63(2), pp. 177–99.

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

Eil, D. and Rao, J.M. (2011). 'The good news-bad news effect: Asymmetric processing of objective information about yourself', *American Economic Journal: Microeconomics*, vol. 3(2), pp. 114–38.
Engelmann, J., Lebreton, M., Schwardmann, P., van der Weele, J.J. and Chang, L.A. (2019). 'Anticipatory anxiety and wishful thinking', Preprint, http://dx.doi.org/10.2139/ssrn.3408017.
Enke, B. and Graeber, T. (2023). 'Cognitive uncertainty', *The Quarterly Journal of Economics*, vol. 138(4), pp. 2021–67.
Erkal, N., Gangadharan, L. and Koh, B.H. (2022). 'By chance or by choice? Biased attribution of others' outcomes when social preferences matter', *Experimental Economics*, vol. 25(2), pp. 413–43.
Ertac, S. (2011). 'Does self-relevance affect information processing? Experimental evidence on the response to performance and non-performance feedback', *Journal of Economic Behavior & Organization*, vol. 80(3), pp. 532–45.
Ertac, S. and Szentes, B. (2011). 'The effect of information on gender differences in competitiveness: Experimental evidence', Working paper, Koç University.
Fischbacher, U. (2007). 'z-tree: Zurich toolbox for ready-made economic experiments', *Experimental Economics*, vol. 10, pp. 171–8.
Gervais, S. and Odean, T. (2001). 'Learning to be overconfident', *The Review of Financial Studies*, vol. 14(1), pp. 1–27.
Goette, L. and Kozakiewicz, M. (2020). 'Experimental evidence on misguided learning', Working paper, University of Bonn.
Golman, R., Hagmann, D. and Loewenstein, G. (2017). 'Information avoidance', *Journal of Economic Literature*, vol. 55(1), pp. 96–135.
Gotthard-Real, A. (2017). 'Desirability and information processing: An experimental study', *Economics Letters*, vol. 152, pp. 96–9.
Grether, D.M. (1980). 'Bayes rule as a descriptive model: The representativeness heuristic', *The Quarterly Journal of Economics*, vol. 95(3), pp. 537–57.
Grossman, Z. and Owens, D. (2012). 'An unlucky feeling: Overconfidence and noisy feedback', *Journal of Economic Behavior & Organization*, vol. 84(2), pp. 510–24.
Hastorf, A.H., Schneider, D.J. and Polefka, J. (1970). *Person Perception*, Reading, MA: Addison-Wesley Publishing.
Heider, F. (1944). 'Social perception and phenomenal causality', *Psychological Review*, vol. 51(6), pp. 358–74.
Heider, F. (1958). *The Psychology of Interpersonal Relations*, Hillsdale, NJ: Lawrence Erlbaum Associates.
Heidhues, P., Köszegi, B. and Strack, P. (2018). 'Unrealistic expectations and misguided learning', *Econometrica*, vol. 86(4), pp. 1159–214.
Hestermann, N. and Le Yaouanq, Y. (2021). 'Experimentation with self-serving attribution biases', *American Economic Journal: Microeconomics*, vol. 13(3), pp. 198–237.
Holt, C.A. and Smith, A.M. (2009). 'An update on Bayesian updating', *Journal of Economic Behavior & Organization*, vol. 69(2), pp. 125–34.
Hossain, T. and Okui, R. (2013). 'The binarized scoring rule', *Review of Economic Studies*, vol. 80(3), pp. 984–1001.
Karni, E. (2009). 'A mechanism for eliciting probabilities', *Econometrica*, vol. 77(2), pp. 603–6.
Kelley, H.H. (1973). 'The processes of causal attribution', *American Psychologist*, vol. 28(2), pp. 107–28.
Kelley, H.H. and Michela, J.L. (1980). 'Attribution theory and research', *Annual Review of Psychology*, vol. 31(1), pp. 457–501.
Kunda, Z. (1990). 'The case for motivated reasoning', *Psychological Bulletin*, vol. 108(3), pp. 480–98.
Larrick, R.P., Burson, K.A. and Soll, J.B. (2007). 'Social comparison and confidence: When thinking you're better than average predicts overconfidence (and when it does not)', *Organizational Behavior and Human Decision Processes*, vol. 102(1), pp. 76–94.
Lassiter, G.D., Geers, A.L., Munhall, P.J., Ploutz-Snyder, R.J. and Breitenbecher, D.L. (2002). 'Illusory causation: Why it occurs', *Psychological Science*, vol. 13(4), pp. 299–305.
Machina, M.J. (1982). ' "Expected utility" analysis without the independence axiom', *Econometrica*, vol. 50(2), pp. 277–323.
Marray, K., Krishna, N. and Tang, J. (2021). 'How do expectations affect learning about fundamentals? Some experimental evidence', Working paper, University of Oxford.
Mezulis, A.H., Abramson, L.Y., Hyde, J.S. and Hankin, B.L. (2004). '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*, vol. 130(5), pp. 711–47.
Miller, D.T. and Ross, M. (1975). 'Self-serving biases in the attribution of causality: Fact or fiction?', *Psychological Bulletin*, vol. 82(2), pp. 213–25.
Möbius, M.M., Niederle, M., Niehaus, P. and Rosenblat, T.S. (2022). 'Managing self-confidence: Theory and experimental evidence', *Management Science*, vol. 68(11), pp. 7793–817.
Moore, D.A. and Small, D.A. (2007). 'Error and bias in comparative judgment: On being both better and worse than we think we are', *Journal of Personality and Social Psychology*, vol. 92(6), pp. 972–89.
Oexl, R. and Grossman, Z.J. (2013). 'Shifting the blame to a powerless intermediary', *Experimental Economics*, vol. 16(3), pp. 306–12.
Oster, E., Shoulson, I. and Dorsey, E.R. (2013). 'Optimal expectations and limited medical testing: Evidence from Huntington disease', *American Economic Review*, vol. 103(2), pp. 804–30.
Pryor, J.B. and Kriss, M. (1977). 'The cognitive dynamics of salience in the attribution process', *Journal of Personality and Social Psychology*, vol. 35(1), pp. 49–55.

© The Author(s) 2024.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

Pulford, B.D. and Colman, A.M. (1997). 'Overconfidence: Feedback and item difficulty effects', *Personality and Individual Differences*, vol. 23(1), pp. 125–33.
Schwardmann, P. and van der Weele, J. (2019). 'Deception and self-deception', *Nature Human Behaviour*, vol. 3(10), pp. 1055–61.
Schwardmann, P., Tripodi, E. and van der Weele, J.J. (2022). 'Self-persuasion: Evidence from field experiments at international debating competitions', *American Economic Review*, vol. 112(4), pp. 1118–46.
Schwarz, N. and Clore, G.L. (1983). 'Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states', *Journal of Personality and Social Psychology*, vol. 45(3), pp. 513–23.
Tetlock, P.E. and Levi, A. (1982). 'Attribution bias: On the inconclusiveness of the cognition-motivation debate', *Journal of Experimental Social Psychology*, vol. 18(1), pp. 68–88.
Tversky, A. and Kahneman, D. (1973). 'Availability: A heuristic for judging frequency and probability', *Cognitive Psychology*, vol. 5(2), pp. 207–32.
Weiner, B. (2010). 'Attribution theory', in (T. O'Connor and C. Sandis, eds.), *A Companion to the Philosophy of Action*, pp. 366–73, Oxford: John Wiley and Sons.
Wozniak, D., Harbaugh, W.T. and Mayr, U. (2014). 'The menstrual cycle and performance feedback alter gender differences in competitive choices', *Journal of Labor Economics*, vol. 32(1), pp. 161–98.
Zimmermann, F. (2020). 'The dynamics of motivated beliefs', *American Economic Review*, vol. 110(2), pp. 337–63.
Zuckerman, M. (1979). 'Attribution of success and failure revisited, or: The motivational bias is alive and well in attribution theory', *Journal of Personality*, vol. 47(2), pp. 245–87.

Downloaded from academic.oup.com/ej/article/134/661/1835/7619130 by guest on 11 November 2026

© The Author(s) 2024.

The Economic Journal, **134** (*July*), 1835–1874 https://doi.org/10.1093/ej/ueae005 © The Author(s) 2024 Published by Oxford University Press on behalf of Royal Economic Society. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
Advance Access Publication Date: 5 March 2024