# Biases in Belief Reports[§]

Dominik Folli
*Graduate School of Decision Sciences, University of Konstanz*
dominik.3.bauer@uni.kn

Irenaeus Wolff
*Thurgau Institute of Economics/ University of Konstanz*
wolff@twi-kreuzlingen.ch

Forthcoming in: *Journal of Economic Psychology*

**Abstract:** Belief elicitation is important in many different fields of economic research. We show that how a researcher elicits such beliefs—in particular, whether the belief is about the participant's opponent, an unrelated other, or the population of others—strongly affects the belief reports. We study the underlying processes and find a clear consensus effect. Yet, when matching the opponent's action would lead to a low payoff and the researcher asks for the belief about this opponent, *ex-post* rationalization kicks in and beliefs are re-adjusted again. Hence, we recommend to ask about unrelated others or about the population in such cases, as 'opponent beliefs' are even more detached from the beliefs participants had when deciding about their actions in the corresponding game. We find no evidence of wishful thinking in any of the treatments.

*JEL classification:* C72, C91, D84

*Keywords:* Belief Elicitation, Belief Formation, Belief-Action Consistency, Framing Effects, Projection, Consensus Effect, Wishful Thinking, *Ex-Post* Rationalization

# 1 Introduction

Subjective beliefs play a central role in economic theory. When facing a decision, people often do not know the true probabilities of the relevant states of the world. Standard economic theory assumes that in such situations, people form subjective beliefs and act on those subjective beliefs as if they were the true probabilities (Savage, 1954). Because of this assumption, eliciting subjective beliefs

[§]We would like to thank Ariel Rubinstein, Yuval Salant, Robin Cubitt, Marie Claire Villeval, Bård Harstad, Dirk Sliwka, and Roberto Weber, the research group at the Thurgau Institute of Economics and the members of the Graduate School of Decision Sciences of the University of Konstanz, as well as participants of several conferences and seminars for their helpful comments. The four data-sets discussed in this paper are available under https://dx.doi.org/10.23663/x2679.

often is extremely helpful to test economic models, as well as for understanding behaviour more generally. The list of examples for this approach is long (for a list of examples from numerous domains, see, *e.g.*, Trautmann & van de Kuilen, 2015).

Both for model-testing purposes and for understanding behaviour, we need to know the true beliefs that underlie behaviour, which may or may not correspond to what we elicit as belief reports. Thus, it is crucial to know whether our elicitation methods trigger any additional processes—or that we at least know the biases that our methods come with. And, indeed, there is a sizeable literature on belief elicitation (for recent reviews, see Schotter & Trevino, 2014, or Schlag *et al.*, 2015). However, the literature has focused mainly on two questions: how to incentivize belief reports,[1] and whether to ask for beliefs before or after actions are chosen.[2]

We will focus on a different aspect: the belief's 'target', namely whether participants are asked about their situation-specific opponent or about unrelated others (we say that the belief "targets" a particular player—or group of players—if the belief represents the expectation of what that player/these players will do). Virtually all studies in the literature use a population treatment (asking about all other participants in the session) or an opponent treatment (asking about the participants' direct interaction partner), but the specific choice is rarely motivated.

Importantly, this choice correlates with the results of a study. All major studies in economics documenting a consensus effect (forming beliefs about others using oneself as a model) use a population treatment.[3] In contrast, studies on belief-action consistency typically use opponent treatments and do not find a

[1] *E.g.*, Armantier & Treich (2013), Erkal *et al.* (2020), Harrison *et al.* (2014), Holt & Smith (2016), Hossain & Okui (2013), Karni (2009), Palfrey & Wang (2009), Trautmann & van de Kuilen (2015).

[2] *E.g.*, Costa-Gomes & Weizsäcker (2008); out of the 20 studies mentioned in ftns. 3 and 4, 15 ask for beliefs only after choices at least in some treatments, 9 do so exlusively, and 6 use different treatments to control for the timing of the belief question (one study does not give information about the elicitation order). Additional topics are hedging (Blanco *et al.*, 2010), the usefulness of second-order beliefs (Manski & Neri, 2013), the precision with which probabilities can be expressed (Delavande *et al.*, 2011a), or a central-tendency bias (Crosetto *et al.*, 2020).

[3] Selten & Ockenfels (1998), Charness & Grosskopf (2001), Van Der Heijden, Nelissen & Potters (2007), Engelmann & Strobel (2012), Iriberri & Rey-Biel (2013), Blanco *et al.* (2014), Danz, Madarász & Wang (2014), Molnár & Heintz (2016), Rubinstein & Salant (2016), Proto & Sgroi (2017).

consensus effect.[4]

Therefore, our first contribution is to answer the question of why a consensus effect seems to be linked to not asking about the opponent. We show that asking about the opponent's behaviour does not eliminate the consensus effect *per se*.[5] Rather, an opponent treatment will make *ex-post* rationalization (fitting one's belief to a prior action in order to appear consistent) override the consensus effect when actions are strategic substitutes. However, this is exactly the main type of situation that allows to distinguish a consensus effect from other effects. To see that, suppose that actions were not strategic substitutes, such as in pure coordination games. In such cases, *ex-post* rationalization or wishful thinking would make the agent report a higher probability of the opponent choosing the same action as the agent, too.

Our second contribution is to provide additional evidence on whether '*ex-ante* rationalization' (choosing the optimal alternative given a belief, as posited by game theory), the consensus effect, and *ex-post* rationalization are the only processes that matter for belief reports. In light of the huge number of biases that people have been found to exhibit, it is not obvious that no other process would play a role for reported beliefs. And yet, the literature that uses (as opposed to: "studies") belief elicitation discusses exactly the above-mentioned three processes when making sense of empirical observations.

We went through a long list of potential biases to see which of the biases would conceptually fit the setup we had in mind for answering our first research question. For this study, we focus on processes that affect how beliefs change after choices have been made (which will become clearer conceptually in Section 2).[6] Our focus implicitly means that we define in particular one cognitive

[4] Costa-Gomes & Weizsäcker (2008), Danz *et al.* (2012), Hyndman *et al.* (2012), Hyndman *et al.* (2013), Manski & Neri (2013), Nyarko & Schotter (2002), Rey-Biel (2009), Sutter *et al.* (2013), Trautmann & van de Kuilen (2015), Wolff (2018).

[5] We are not able to observe any of the processes directly, and therefore, any of the corresponding statements should be read as "the results are consistent with the interpretation we give." In the above example, the sentence should be understood as: We show that asking about the opponent's behaviour does not eliminate the (observed) effect that would be consistent with a consensus effect. We stick to the stronger statements in the text for better readability.

[6] The biases that fit our setup if conceptualized appropriately were bias blind spot, cognitive dissonance, confirmation bias, conservatism in updating, correlation neglect, illusion of control, salience bias, social-desirability bias, and wishful thinking. Biases that did not make sense within our setup were: base-rate fallacy, belief bias, conjunction fallacy, contrast effect, fundamental attribution error, gambler's fallacy, hindsight bias, hot-hand fallacy, and status-quo bias.

process—the consensus effect—in a way that differs from prior literature. The consensus effect usually is meant to affect belief formation (also) before an action is chosen. Here, we disregard any effect on the belief that happens prior to the action choice because the overarching question of our study is on how to elicit the 'true belief' (the belief at the time of the action choice). Hence, when we talk about a "consensus effect", we refer to how a consensus effect may shift the belief after an action has been chosen.

From the list of biases, we found 11 biases that seem applicable to our setting, 9 of which happen after choices are made (For an overview, see Table 2 at the beginning of Section 2). However, four of the 11 applicable biases are possible root-causes of *ex-post* rationalization, and two others cannot be isolated from the consensus effect, which is why we group them into two 'bias groups'. In the end, we will be able to distinguish one process in addition to what has been discussed in the literature on belief reports: wishful thinking. Reassuringly for the interpretation of existing studies, we do not find evidence for wishful thinking to affect belief reports.

Our paper has two main parts, comprising three experiments. In Experiment 1-DISC, we use a pure discoordination game and elicit beliefs in the two standard treatments, the opponent treatment and the population treatment. As pointed out, we replicate the systematic differences from the literature: a consensus effect in the population treatment, and higher observed best-response rates in the opponent treatment.

The population and opponent treatments differ in four ways (which is why we refrain from calling them "frames"; when we do talk of "frames", we refer to the mental representation of the question in participants' heads). The four differences are (i) the participants' interaction with the belief's 'target' (in the population treatment, the belief question is mostly about the behaviour of people the participant is *not* interacting with; in the opponent treatment, the question is only about the person the participant *is* interacting with); (ii) how many people are the belief's target; (iii) the exact incentivization; and (iv) whether we ask about a percentage or a probability. To find out which features of the main treatments are responsible for the differences between them, we add a third treatment which we call 'random-other treatment'. A random-other treatment asks for participants' beliefs about the behaviour of some other individual who is not the matching partner, and allows for *ceteris-paribus* comparisons with the corre-

**Opponent treatment**

Object: Single person, the matching partner

*"With what probability did your matching partner choose each of the respective boxes of the current set-up?"*

Incentivization: $Pr(\text{win}) = 1 - \frac{1}{2}\Big(\big[1 - r(a_{\text{true}})\big]^2 + \sum_{a_j \neq a_{\text{true}}} r(a_j)^2\Big)$, where $r(a_j)$ is the reported probability of the 'Object' playing action $a_j$ and $a_{\text{true}}$ is the 'Object's' true choice.

**Random-other treatment**

Object: Single person, not the matching partner

*"With what probability did a person who is not your matching partner choose each of the respective boxes of the current set-up?"*

Incentivization: $Pr(\text{win}) = 1 - \frac{1}{2}\Big(\big[1 - r(a_{\text{true}})\big]^2 + \sum_{a_j \neq a_{\text{true}}} r(a_j)^2\Big)$.

**Population treatment**

Object: Many people, almost all of them not the matching partner

*"What is the percentage of other participants of today's experiment choosing each of the respective boxes of the current set-up?"*

Incentivization: $Pr(\text{win}) = 1 - \frac{1}{2}\sum_j \big[r(a_j) - f(a_j)\big]^2$, where $f(a_j)$ denotes action $a_j$'s relative frequency in the population.

**Table 1:** The three types of treatments we use (differences underlined).

sponding opponent treatment.

Table 1 contrasts the three types of treatments, show-casing all four differences betweeen opponent and population treatments. Our data shows that the relevant difference is between the random-other and opponent treatments, and not between the random-other and the population treatments. Thus, it is the *interaction* with the belief's target that makes the difference.

The second part of our paper varies the environment in which we elicit beliefs (in particular, the game people play). We use two experiments to disentangle the processes that lead to biased belief reports, using a 'to-your-left game' (a rock-paper-scissors-type of game) in Experiment 2-TYL and a battle-of-the-sexes game with alternating (but unobservable) moves in Experiment 2-BOS.

Experiment 2-TYL rules out potentially active biases that might have affected belief reports in Experiment 1-DISC. To test for a consensus effect in the opponent treatment, Experiment 2-BOS eliminates the 'cognitive need' for *ex-post* rationalization. We find as much of a consensus effect in the opponent treatment as in the population treatment. We thus conclude that initially, opponent treatments lead to as much consensus effect as the other treatments. However, opponent treatments trigger subsequent *ex-post* rationalization whenever the beliefs that result from the consensus effect would lead to cognitive dissonance.

# 2 The applicable processes and our treatments

Table 2 gives a short description of processes known to affect probability judgments and indicates whether a process is applicable within our setting(s). The three bold-faced processes are the processes that have been discussed prominently as determinants of belief reports.

We next describe the applicable processes and identify in which treatment(s) they could matter. We summarize our predictions in Table 3 at the end of this Section. Section 3 then describes the experiments in detail, and Sections 4 and 5 relate them to our general predictions from the current Section.

Before we discuss the processes in detail, however, Figure 1 shows our conceptualization of the process leading to a belief report. Salience bias (being attracted by salient items) and bias blind spot (assuming that only others are affected by a bias, in this case, salience bias) will happen when players form their belief. '*Ex-ante* rationalization' (forming a belief and best-responding to it) then leads to the chosen action.

After the players have chosen their action, we (as the researchers) ask them for their belief. At this point, biases like *ex-post* rationalization, the consensus effect, or wishful thinking (and the corresponding underlying processes) play out and re-shape the latent belief into a final belief report.[7] As pointed out in the introduction, we will have to re-adjust Figure 1 at the end of our study, eliminating wishful thinking, and placing *ex-post* rationalization after the consensus effect.

[7] It is conceivable that a consensus effect or potentially even wishful thinking affect beliefs even prior to the action being taken. However, we are not interested in that part of the process as it would enter the 'true' belief underlying the action. If any of these processes enter the true belief, then a researcher typically wants belief-elicitation to capture that, too. We are focusing on the differences between 'true' belief and belief report, and thus focus only on the ('part of' the) processes that happen after the choice.

| Process | Short description | Applica-bility | Focus of ex-periment... |
|---|---|---|---|
| **‘ex-ante rationalization’** | forming a belief, then best-responding to the belief | ✓ | 2-TYL |
| bias blind spot | everybody else is falling for a fallacy, but not me | (✓) | |
| **consensus effect** | belief that others are like me ⇒ they will act as I do/would | ✓ | |
| conservatism in updating | (partially) ignoring new information | ✓ | 2-TYL/2-BOS |
| correlation neglect | ignoring that two events are correlated | ✓ | |
| ***ex-post* rationalization** | fitting a belief to an action after that action has been taken | ✓ | |
| cognitive dissonance | when my action is inconsistent with my belief, I adjust the belief as to correct the inconsistency | ✓ | |
| illusion of control | belief that I can influence pure-chance moves | ✓ | 2-TYL/2-BOS |
| social-desirability bias | reporting behaviour/opinions that conforms untruthfully closely to social norms | ✓ | |
| confirmation bias | when I have a theory, I only search for confirming evidence | ✓ | |
| salience bias | being attracted by salient choices/labels | (✓) | |
| wishful thinking | when people assign a higher probability to favourable outcomes just because they are favourable | ✓ | 2-TYL |
| base-rate fallacy | ignoring prior probabilities | ✗ | |
| belief bias | if the conclusion is right, the argument must be right, too | ✗ | |
| conjunction fallacy | ignoring that the conjunction of two events can never be more likely than either event separately | ✗ | |
| contrast effect | draws more attention to items/characteristics that change strongly | ✗ | |
| fundamental attribution error | attributing too much to the characteristics of a person and too little to the characteristics of the situation | ✗ | |
| gambler’s fallacy | belief that prior realisations of an i.i.d. process change future probabilities, to move the observed mean closer to its expected value | ✗ | |
| hindsight bias | not being able to abstract from knowledge acquired after a choice was made, when assessing that choice | ✗* | |
| hot-hand fallacy | belief that s.b. who has been lucky several times in a row is more likely to be lucky the next time, too | ✗ | |
| status-quo bias | a preference for the current state relative to any changes, irrespective of what the current state is | ✗ | |

*A hindsight bias could in principle apply to our setting in the following way: A player knows her own action at the time of stating her belief about her opponent. If she cannot abstract from the knowledge about her action when forming her belief about the other player’s choice, she might adjust her belief such that a best-response to her own action by her opponent becomes very likely. We explored that possibility in the working-paper version Bauer & Wolff (2018), finding no evidence for it.

Table 2: Overview of all processes considered. Processes that have been prominent in the literature as affecting belief reports are marked in bold face. Processes that are considered jointly with or as underlying causes of other processes are indented and directly follow the corresponding ‘process category’. Further note that some of the “non-applicable” ones are non-applicable because they would have required feedback about others’ choices.

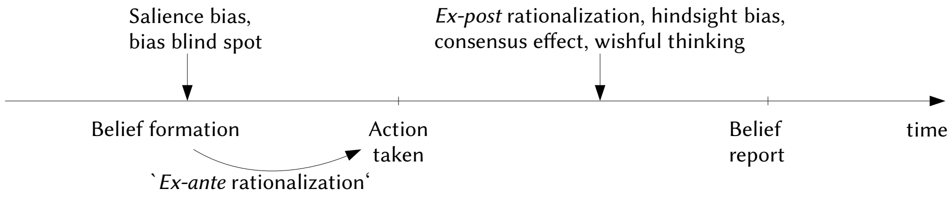



**Figure 1**: Timing of when and which processes are expected to be active in our setting.

Having looked at the broad picture, let us introduce some notation that we will need on the following pages to fix ideas. In all of our Experiments, participants interact in pairs facing an action set $\mathcal{A} = \{a_1, a_2, ..., a_N\}$ that is the same for both players. Participants have to make a choice $c^{(i)}$ from $\mathcal{A}$. The choice $c^{(i)}$ translates into a probability distribution $a(c^{(i)})$ over $\mathcal{A}$. In all but one cases, this translation is trivial: $a^{(k)} = 1$ if $c^{(i)} = a_k$, and $a^{(k)} = 0$, otherwise, where $a^{(k)}$ is the probability with which $a_k$ is the payoff-relevant action. Only in Experiment 2-TYL, $a^{(k)} = 5/8$ if $c^{(i)} = a_k$, and $a^{(k)} = 1/8$, otherwise, because we introduce uniformly-random implementation errors with probability $1/2$. Participants' (expected) payoffs from the game $EU_\Gamma$ are determined by the joint distribution $a(c^{(i)}) \times a(c^{(j)})$, which in all Experiments but 2-TYL can be represented by the vector of choices $(c^{(i)}, c^{(j)})$. Finally, $BR(p)$ is the set of $EU_\Gamma$-defined best-responses to a probability distribution $p$.

After participants have made their choices, they have to make a belief report, $r = (r_1, r_2, ..., r_N)$. Participants' expected payoff from the belief elicitation, $EU_E$ then depends on $r$ and the underlying belief. We consider three different types of beliefs that are relevant for $EU_E$ in the different treatments. We denote the three types of beliefs by $b_t$, $B_t$, and $b_t^-$, where $t \in \{0, 1\}$ is the point in time when the belief is formed, $0$ is the time of the action choice and $1$ is the time of the report.

The belief $b_t = (b_{t,1}, b_{t,2}, ..., b_{t,N})$, is participants' best estimate of what their opponent in the game will do (or will have done), where $b_{t,k}$ is the probability that the participant assigns at time $t$ to the event that the opponent chooses $c^{(j)} = a_k$. $B_t = (B_{1,t}, B_{2,t}, ..., B_{N,t})$ is the participant's belief about the population of possible opponents, where $B_{k,t}$ is the fraction of the population that the agent expects to choose $c^{(j)} = a_k$. Finally, $b_t^- = (b_{t,1}^-, b_{t,2}^-, ..., b_{t,N}^-)$ is the belief that is relevant in the random-other treatments: the participant who is not the opponent but whose action determines the belief-elicitation payment in that treatment.

We will assume that agents fulfil Hossain & Okui's (2013) monotonicity constraint and that their report thus will be a truthful representation of their current belief, so that $r = b_t$ in the opponent treatments, $r = B_t$ in the population treatments, and $r = b_t^-$ in the random-other treatments. Note that social-image concerns with a belief that the experimenter prefers consistent behaviour—which might lead to untruthful reports—are behaviourally equivalent to cognitive dissonance which we discuss under the heading of *ex-post* rationalization below. Given that in our view, such social-image concerns would be the primary source of monotonicity-violations, we are positive that the above assumption is not a restrictive constraint within the context of our study.

### '*Ex-Ante* Rationalization'

'*Ex-ante* rationalization' corresponds to the standard game-theoretic model. Thus, agents choose $c^{(i)}$ such as to maximize their expected utility given $b_0$:

$$\max_{c^{(i)}} EU_\Gamma(a(c^{(i)})|b_0) \Rightarrow c^{(i)} \in BR(b_0).$$

In all experiments but 2-TYL, the above holds by definition. In Experiment 2-TYL, each player 'makes implementation errors' with probability $1/2$, and thus, $EU_\Gamma(a(c^{(i)})|b_0) = \frac{1}{4}EU_\Gamma(c^{(i)}|b_0) + \frac{1}{4}EU_\Gamma(c^{(i)}|(\frac{1}{N}, \frac{1}{N}, ..., \frac{1}{N})) + \frac{1}{4}\sum_{k=1}^{N}\frac{1}{N}EU_\Gamma(a_k|b_0) + \frac{1}{4}\sum_{k=1}^{N}\frac{1}{N}EU_\Gamma(a_k|(\frac{1}{N}, \frac{1}{N}, ..., \frac{1}{N}))$. Agents can only influence the first two terms, and $\frac{1}{4}EU_\Gamma(c^{(i)}|(\frac{1}{N}, \frac{1}{N}, ..., \frac{1}{N}))$ is a constant in the games we study. Hence, $c^{(i)} \in BR(b_0)$ solves the agents' problem in Experiment 2-TYL, too.

After choosing $c^{(i)}$, participants have to choose their report $r$ optimally, given the belief that corresponds to the treatment. As an example, in the population treatment, they have to solve

$$\max_{r} EU_E(r|B_t).$$

However, given that we are maintaining the idea of a perfectly rational Bayesian agent, that the agent cannot distinguish between different opponents, and that there is no reason for beliefs to change over time, $B_t = b_0$ (and also $b_t^- = b_0$ in the random-other treatments). Thus, $r = b_0$ in all three treatments under *ex-ante* rationalization.

### *Ex-Post* Rationalization

Humans are extremely good at rationalizing whatever they do (so much that certain psychologists even think that beliefs virtually always go second; Chater, 2018). The specific reasons for such *ex-post* rationalization may vary. In the context of our setup, they include cognitive dissonance (Festinger, 1957); social-desirability bias (Edwards, 1953; if participants believe that experimenters expect or like to see consistent behaviour); illusion of control (Langer, 1975; if participants have the perception that they can magically influence the matching); and confirmation bias (Wason, 1960; conceptually more of a stretch). For the purpose of this paper, we subsume all of the above processes under the header of *ex-post* rationalization.

To derive our predictions, we assume cognitive dissonance to be the driving force behind *ex-post* rationalization, noting that the other processes will lead to similar predictions. Next, note that for cognitive dissonance to be relevant, there must be a non-negligible probability that $c^{(i)} \notin BR(b_0)$, which means that agents arrive at their choice $c^{(i)}$ by some process other than (pure) *ex-ante* rationalization. Focusing on the beliefs, we consider $c^{(i)}$ as given and consider the 'belief-choice' that will subsequently lead to the report $r$ (recall that we focus on the case that $r$ is a truthful representation of $b_1, B_1,$ or $b_1^-$, respectively).

Let us consider the opponent treatment first. In this type of treatment, the *ex-post*-rationalizing agent maximises

$$EU(r, b_1|c) = EU_E(r|b_1) - \delta 1_{c^{(i)} \notin BR(b_1)} - \gamma \sum_{k=1}^{N} (b_{0,k} - b_{1,k})^2,$$

where $\delta$ is a penalty for maintaining a belief $b_1$ about the opponent's behaviour that is at odds with the prior choice $c^{(i)}$, $1_{c^{(i)} \notin BR(b_1)}$ is the indicator function that is 1 whenever $c^{(i)} \notin BR(b_1)$ and 0, otherwise, and $\gamma$ is a measure of how hard the agent finds it to convince himself of having had a different belief than $b_0$. Here—as well as in the other processes that lead to a distortion of beliefs—we assume a quadratic 'cost' function to depict the idea that distorting one's belief a little will be much easier than distorting one's belief a lot. We maintain the idea that the agent first chooses the time-1 belief and then reports $r$.

Given our assumptions, $r = b_1$. If $c^{(i)} \in BR(b_0)$, $r = b_1 = b_0$ because changing the belief is cognitively costly. In contrast, if $c^{(i)} \notin BR(b_0)$, $r = b_1 = b_0$

if and only if

$$EU_E(b_0|b_0) - \delta > \max_{b_1} EU_E(b_1|b_1) - \delta 1_{c^{(i)} \notin BR(b_1)} - \gamma \sum_{k=1}^{N} (b_{0,k} - b_{1,k})^2.$$

If the inequality is not fulfilled, most often $b_1$ will be chosen such that $c^{(i)} \in BR(b_1)$ so that the $\delta$-term disappears. In any case, the inequality makes it clear that the possibility of changing the belief such that $b_1 \neq b_0$ could have a side effect if agents were sophisticated: not only does a change in belief allow to bring $b_1$ in line with $c^{(i)}$, it would also allow to increase the subjective probability of earning the belief-elicitation prize of the Binarized Scoring Rule.[8] However, we assume that agents do not exhibit that level of sophistication (*i.a.*, because it would mean that agents are aware of the fact that they are manipulating their beliefs, which is psychologically implausible).

Consider next the population treatment. Here, the *ex-post*-rationalizing agent maximises

$$EU(r, B_1|c) = EU_E(r|B_1) - \delta 1_{c^{(i)} \notin BR(b_1)} - \gamma \sum_{k=1}^{N} (B_{0,k} - B_{1,k})^2,$$
$$\text{s.t. } B_{1,k} \geq \frac{b_{1,k}}{P}, \forall k,$$

where $P$ is the size of the population that forms the 'target' of the population-treatment belief report. The side constraint is a logical constraint that excludes the possibility that an agent has beliefs $b_1$ and $B_1$ that cannot both be correct even if the agent knew which opponent (s)he was matched to. For example, if the population of others consists of 2 people and the agent maintains a $b_1$ s.t. $b_{1,1} = 1/2$, then $B_{1,1}$ has to be at least $1/4$: if the opponent plays $a_1$ with a probability of 50%, then the collective of (both) others cannot possibly play $a_1$ with an aggregate probability below 25%.

It is in the sense of the above objective function that there is no need for the agent to find arguments to change the belief, as long as there is any way to 'reconcile $c^{(i)}$ with $B_0$'. On top, for large-enough $P$, the side constraint becomes non-binding, so that the agent should virtually always report $r = B_0$ in our

[8] Assuming that $EU(r, b_1|c) = f[EU_E(r|b_0)]$ would be implausible psychologically: the agent would be assumed "still to know" the initial belief at the same time as changing it. In contrast, the $\gamma$-term in $EU(r, b_1|c)$ simply reflects the cognitive effort of finding arguments of why the belief was different in the first place.

settings ($19 \leq P \leq 27$).

Finally, in the random-other treatment, the objective function becomes

$$EU(r, b_1^-|c) = EU_E(r|b_1^-) - \delta 1_{c^{(i)} \notin BR(b_1)} - \gamma \sum_{k=1}^{N} (b_{0,k}^- - b_{1,k}^-)^2.$$

Given that the opponent and the random other are two different people, there is no logical requirement for them to act in a similar way. Hence, $c^{(i)}$ does not impose any restriction on $b_1^-$, and—given that changing the belief is cognitively costly—the agent will always report $r = b_0^-$.

### Wishful Thinking

A large body of literature studies *unrealistic optimism*, which is described as a tendency to hold overoptimistic beliefs about future events (e.g. Camerer & Lovallo 1999, Larwood & Whittaker 1977, Svenson 1981, Weinstein 1980, 1989, or Heger & Papageorge, 2018). *Wishful thinking* has been brought forward as a possible cause of unrealistic optimism and has been described as a desirability bias (Babad & Katz 1991, Bar-Hillel & Budescu, 1995). Wishful thinking hence means a subjective overestimation of the probability of favorable events (*cf.* also the closely related idea of *affect* influencing beliefs, Charness & Levin, 2005). Despite the large body of evidence on human optimism (Helweg-Larsen & Shepperd, 2001), there is some doubt about whether a genuine wishful-thinking effect truly exists (Krizan & Windschitl, 2007, Bar-Hillel *et al.* 2008, Harris & Hahn, 2011, Shah *et al.,* 2016).

In the context of this study, an agent whose belief is influenced by wishful thinking places an unduly high subjective probability on the event that others act such that the agent receives a (high) payoff. In particular, the agent's belief-choice problem in the opponent treatment is to maximize the following function over $r$ and $b_1$:

$$EU(r, b_1|\tilde{a}) = EU_E(r|b_1) + \omega EU_\Gamma(\tilde{a}|b_1) - \gamma \sum_{k=1}^{N} (b_{0,k} - b_{1,k})^2,$$

where $\omega$ measures how important it is to the agent to have a belief that yields a high probability of winning given the implemented action $\tilde{a}$. In principle, the second part of the second term on the right-hand side should be a function $EU(b_1|\tilde{a})$.

For ease of exposition, we nonetheless stick to the notation $EU_\Gamma(\tilde{a}|b_1)$ as it is immediately clear how to calculate the latter.

As before, we assume the Binarized Scoring Rule to be proper for our agents, so that $r = b_1$. Given the specification above, the agent will adjust the belief whenever that is possible (*i.e.*—with a slight abuse of notation—whenever $\sum_{k|c^{(i)}\in BR(a_k)} b_{0,k} < 1$), because the marginal costs of adjustment are 0 at $b_0$, while the marginal benefit is strictly positive.

Consider now the population treatment. Here, the objective function is:

$$EU(r, B_1|\tilde{a}) = EU_E(r|B_1) + \omega EU_\Gamma(\tilde{a}|b_1) - \gamma \sum_{k=1}^{N}(B_{0,k} - B_{1,k})^2,$$
$$\text{s.t. } B_{1,k} \geq \frac{b_{1,k}}{P}, \forall k.$$

As before, $r = B_1$, and $B_1 \approx B_0$, because the agent does not face any incentives to change the population belief as long as the population belief is 'compatible' with the overoptimistic belief about the actual opponent, $b_1$. While adjustments may be somewhat more frequent under wishful thinking compared to *ex-post* rationalization ($b_1$ will be a point belief most of the time, while the condition $c^{(i)} \in BR(b_1)$ normally does not require point beliefs), the necessary adjustments will be small (at the very most, $1/P$). In summary, we do not predict wishful thinking to be detectable in a population treatment.

Finally, let us consider the objective function in the random-other treatment:

$$EU(r, b_1^-|\tilde{a}) = EU_E(r|b_1^-) + \omega EU_\Gamma(\tilde{a}|b_1) - \gamma \sum_{k=1}^{N}(b_{0,k}^- - b_{1,k}^-)^2.$$

As discussed for *ex-post* rationalization, $b_1$ and $b_1^-$ are logically independent, and thus, in random-other treatments $r = b_0^-$ also under wishful thinking.

### Consensus Effect

The *consensus effect* is a phenomenon studied by psychologists and economists. Tversky & Kahneman (1973, 1974) link it to the *availability heuristic* and the *anchoring-and-adjustment heuristic.* Joachim Krueger describes the consensus effect in a general but simple way: *"People by and large expect that others are similar to them"* (Krueger, 2007, p. 1). The basic idea has been studied in many different contexts under many different names: [false-]consensus effect (Ross, Greene &

House, 1977; Mullen *et al.*, 1985; Marks & Miller, 1987; Dawes & Mulford, 1996), perspective taking (Epley *et al.,* 2004), social projection (Krueger, 2007; 2013), type projection (Breitmoser, 2019), evidential reasoning (al-Nowaihi & Dhami, 2015) or self-similarity bias (Rubinstein & Salant, 2016).

Engelmann & Strobel (2012) convincingly demonstrate that the consensus effect exists, but only as long as no representative information about others is available. Similarly, Engelmann & Strobel (2000) had demonstrated that participants do use the information provided by their own choice in their belief reports even when information about others' behaviour is available (but that, in their setup, participants underweight their own choice relative to the choices of others). Given that we do not provide any information about others, we expect the consensus effect to be strong in our study.

For this study, we define the consensus effect as a psychological mechanism that changes reported beliefs in the direction of a participant's own action after that action has been taken.[9] In particular, we posit that the agent updates the 'prior' belief $\beta_0 \in \{b_0, B_0, b_0^-\}$ that is applicable in the respective treatment using the observation $c^{(i)}$:

$$\beta_1 = (1-\kappa)\beta_0 + \kappa\hat{b}, \text{ where } \hat{b} = (\hat{b}_1, \hat{b}_2, ..., \hat{b}_N),$$
$$\hat{b}_k = \begin{cases} 0, & c^{(i)} \neq a_k, \\ 1, & c^{(i)} = a_k, \end{cases}$$

and where $\kappa > 0$ is the weight of the 'new observation' (which is the agent's own choice), whether or not $\kappa$ has the value that would be prescribed by Bayesian updating.

Three remarks seem in order. First, we do not impose any cognitive costs on 'distorting' the belief because the agent is 'expecting' to change the belief right from the start (rather than convincing herself that the belief has been different from the initial belief all along). Second, as long as the agent has not made a choice, the agent cannot update any 'prior' (or at the very least, the updating based on a 'considered choice' should be rather limited). And third, correlation

[9]As a consequence, it also is futile to think about what a consensus effect may mean for behaviour in our pure discoordination game in Experiment 1-disC. Whatever process leads to $b_0$ and $c^{(i)}$ is not at our focus: as in the typical experiment, we "simply" want to elicit $b_0$ as well as possible. As a side note, note that the empirical distributions over choices in Experiment 1-disC are far from uniform, and the same applies for participants' reported beliefs in any treatment.

neglect (of the correlation between others' choices and one's own; "illusion of validity" in Kahneman & Tversky, 1973) would correspond to $\kappa = 0$, while conservatism in updating (about others' choices after observing one's own, Edwards, 1968) would imply a (suboptimally) low $\kappa$ (compared to the 'rational' $\kappa$). Given that the effects we observe in our experiments suggest a substantial $\kappa$, we stick to interpreting the results as providing support for a consensus effect.

Apart from the above updating process, nothing changes with respect to the standard model. Hence, agents simply report the updated prior $\beta_1$, that is $r = \beta_1$. While it would be conceivable that the consensus effect is active in all treatments, we rest our (*ex post*: mistaken) hypothesis on the working paper of Rubinstein & Salant (2015): "The population frame highlights similarities among players" while "[t]he opponent frame highlights the strategic aspect of the game". Even though we are talking about symmetric games, the opponent treatment seems to be asking about 'the other side of the interaction' (reacting to 'me'), while the other treatments ask about 'someone/many others in the same position'. Within our model, this would mean that $\kappa$ is treatment dependent and $\kappa = 0$ in the opponent treatment. Hence, we (wrongly) expect to find a consensus effect only in the population and random-other treatments.

### Salience bias and bias blind spot

People who follow a salience bias (Taylor & Fiske, 1975) will choose salient items more often. A bias blind spot means they assume that 'everybody else falls for a bias (in our study, most plausibly a salience bias) but not me' (Pronin, Lin, & Ross, 2002). Both biases may be active in our setting. However, they will act primarily *before* a participant decides on an action (and equally so across all treatments). This might seem less clear for the bias blind spot; however, if a participant thinks everybody else's choices are going to be shaped by salience, then, the participant will have held this belief already at the time of chosing an action (which in that case will be a best-reply to the belief that everybody else chooses the salient item). We are focusing on changes in a belief that happen after an action is chosen, and therefore, we leave bias blind spot and salience bias out of the equation (we offer a short formalization of both in Online Appendix B).

| | Population | Random Other | Opponent |
|---|---|---|---|
| *Ex-ante* rationalization | ✓ | ✓ | ✓ |
| *Ex-Post* Rationalization | - | - | ✓ |
| Wishful Thinking | - | - | ✓ |
| Consensus Effect | ✓ | ✓ | - |

**Table 3:** Predictions of which processes are active under which treatment.

# 3 Experimental Design

### Rationale behind the experiments

We start this Section by describing the specific purposes of the three experiments of this paper. Experiment 1-DISC serves three purposes. First, it replicates Rubinstein & Salant's (2015) finding that beliefs are closer to participants' own actions under a population treatment than under an opponent treatment.

Second, Experiment 1-DISC highlights the consequences the elicitation treatment has for conclusions about participants' belief-action consistency. Third, and most importantly, it shows that the difference in behaviour between the population treatment and the opponent treatment stems from the 'interaction partner *vs.* another person' difference and not from any of the other differences.

Experiments 2-TYL and 2-BOS disentangle different mental processes that may underlie Experiment 1-DISC's findings. They provide evidence on which of the known biases and processes are important, and when. Experiment 2-TYL separates the consensus effect and wishful thinking from *ex-ante* and *ex-post* rationalization. In addition, we need Experiment 2-BOS to differentiate between two possible explanations of the data: under an opponent treatment, (i) the consensus effect is overridden by *ex-post* rationalization, and (ii) there is no consensus effect to begin with. Table 4 summarizes the experiments and their purposes.

### Experimental setup

In Experiment 1-DISC, participants face a series of 24 one-shot, two-player, four-action pure discoordination games. Players get a prize of 7€ if they choose different actions and nothing, otherwise. Participants play the discoordination games with randomly changing partners, and without any feedback in between.

| Exp. | Game/Treatments | Purpose |
|---|---|---|
| 1-disC | Discoordination (Pop, Opp) | - Replicating that beliefs are closer to participants's actions under a population treatment than under an opponent treatment<br>- Highlighting the consequences for measured belief-action consistency |
| | (RO) | - Identifying the critical treatment difference by the random-other treatment: interaction with the 'belief target', whether the 'target' is a single person or many, asking about a percentage *vs* a probability, or the exact incentivization |
| 2-tyl | To-your-left (with implementation errors) (RO, Opp) | - Separating the consensus effect and wishful thinking from *ex-ante* rationalization and *ex-post* rationalization |
| 2-bos | Battle-of-the-Sexes with alternating (but unobservable) moves (Pop, Opp) | - Disentangling whether in opponent treatments, ...<br>(i) a consensus effect is overridden by *ex-post* rationalization, or<br>(ii) whether there is no consensus effect in opponent treatments |

**Table 4:** Overview of the experiments and their purpose. Pop stands for the population treatment, Opp for the opponent treatment, and RO for the random-other treatment.

Participants play the discoordination games on different sets of boxes carrying labels such as "1", "2", "3", and "4", or "1", "x", "3", and "4", or "a", "a", "a", and "B". We provide the full list of label sets in Table A1 in Online Appendix A. All participants went through the same order of sets. We chose the varying sets to keep up participants' attention.

In Experiment 2-tyl, we use the same sets of labelled boxes. However, participants play one-shot "to-your-left games" (Wolff, 2021), in which a player gets a prize of 12€ if he chooses the box immediately to the left of his opponent's choice. The game works in a circular fashion, so that choosing "4" against a choice of "1" by your opponent would make you win the 12€ in a "1-2-3-4" setting. The difference in payoffs is meant to reduce expected-earnings differences accross experiments: In a discoordination game, (both) participants are likely to win fairly often, while in the "to-your-left game", participants will win at a much lower rate.

To separate wishful thinking from *ex-ante* rationalization and *ex-post* rationalization, we add random implementation errors to Experiment 2-tyl. There is a 50% probability that the computer changes a participant's decision. If the computer alters the decision, the computer chooses each box with equal probability (including the participant's chosen box). We then inform participants about whether their decision has been altered, and if so, which box the computer has chosen.

If the computer changes the decision, the computer's choice is used to determine the game payoff of the participant and of her interaction partner. However, the belief elicitation still targets the other participants' original choices, not the implemented ones. Hence, *ex-ante* and *ex-post* rationalization still mean a higher probability mass on the option to the right of the participant's originally chosen option even when the computer changes the decision. In contrast, wishful thinking implies a higher probability mass on the option to the right of the implemented decision.

We elicit probabilistic beliefs directly after each choice in the game, incentivizing the belief reports via a Binarized-Scoring Rule (McKelvey & Page, 1990, Hossain & Okui, 2013). Beliefs had to sum up to 100% before participants could go on. However, the interface allowed to enter any non-(or super-)additive belief (either by clicking into a diagramme or by entering numbers) and then click on a "scale" button that would scale the belief up or down to 100%. In other words, there was no need for participants to change their relative belief reports after inserting a non-or-super-additive belief.

In the belief-elicitation task, subjects could earn another 7€. The Binarized-Scoring Rule uses a quadratic scoring rule to assign participants lottery tickets for a given prize. The lottery procedure accounts for deviations from risk neutrality and, under a weak monotonicity condition, even for deviations from expected utility maximization (Hossain & Okui, 2013). Hence, we control for participants' risk preferences (also) in the belief task.

The exact framing of the belief-elicitation question is subject to treatment variation as described in Section 1. At the end of the experiments, we randomly select two periods for payment. In one period, we pay the outcome of the game and in the other period, the belief task. In Experiment 2-TYL, we use an opponent and a random-other treatments since they allow for a *ceteris-paribus* comparison by changing only the identity of the target.

In Experiment 2-BOS, participants face two one-shot battle-of-the-sexes games, depicted in Figure 2. In each of the two games, players move sequentially but the second-mover does not receive any information on the first-mover's choice. Following the design of Blanco *et al.* (2014), there is role-reversal between the games and belief-elicitation before choices (using a binarized scoring rule with a winning prize of 6€ and a losing prize of 3€). In contrast to Blanco *et al.*'s experiment, we randomly re-matched participants between the two games. Again, if a game was payoff-relevant, the belief payment came from the other game.

This design has the feature that a first-mover in the first game will be asked about his belief about first-mover behaviour (in the second game) directly after making his first-mover choice (in the first game). And because we are eliciting a belief about other first-movers (in a new game), cognitive dissonance does not create a need for the elicited belief to be "consistent" with the participant's previous first-mover choice (all of the above applies in exactly the same way to participants who acted as second-movers in the first game).[10]

We use a different game than in Experiment 1-DISC and Experiment 2-TYL because we need different player roles (*i.e.*, an asymmetric game) to get rid of cognitive dissonance. While, technically, implementing an alternating-move version of the discoordination or to-your-left games would suffice, in neither of the two games the alternating-move-structure would 'make sense' for participants: we conjectured that the asymmetry would not be strong enough. In contrast, in alternating-move battle-of-the-sexes games like the one we use, the alternating-move-structure has been shown to affect behaviour strongly (Cooper *et al.*, 1993). Finally, we use an opponent and a population treatments in order to induce the largest-possible treatment difference in terms of a consensus effect (judging by the results of Experiment 1-DISC).

**Procedures**

We programmed the experiments using z-Tree (Fischbacher, 2007) and conducted them in the LakeLab at the University of Konstanz. We use the data of 145 participants from Experiment 1-DISC, 70 participants from Experiment 2-TYL, and 222 participants from Experiment 2-BOS.[11]

Experiment 2-BOS was run as one out of three parts of an experimental ses-

[10]To see why the setup is appropriate, consider the following alternatives. If we asked about one's opponent in the same game, cognitive dissonance would apply. Asking about one's peers in the next game while maintaining roles would not allow for an opponent treatment. If there was only a single role we might re-introduce cognitive dissonance (the case for social-desirability concerns would be less clear). To reiterate, in order to make sure cognitive dissonance should not be playing a role, we need an asymmetric game played twice, with role-reversal.

[11]For the analysis, we exclude one participant from Experiment 1-DISC who always reported a 100% belief of not having discoordinated. This participant probably tried to hedge, but did not understand that hedging was impossible. We used all data from Experiments 2-TYL and 2-BOS.

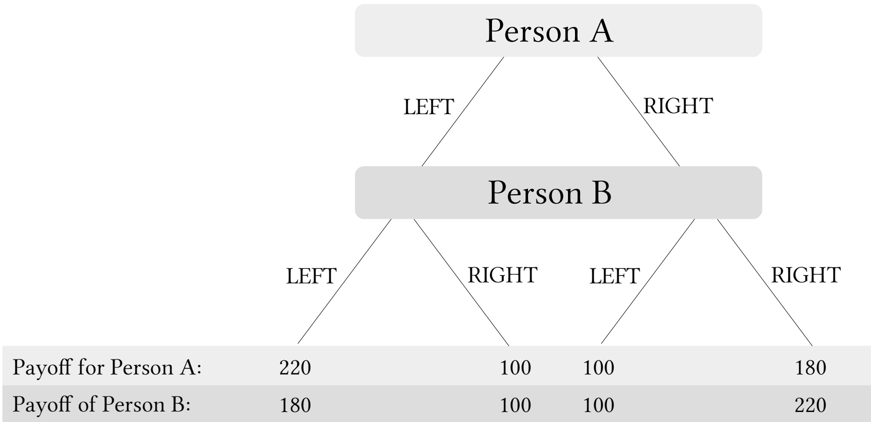



**Figure 2**: Battle-of-the-sexes game used in Experiment 2-BOS. The rounded boxes represent information sets: Person B does not learn Person A's choice before the end of the game.

sion; for 118 participants, this was the first part of the session, for another 104 participants, it was the second part of the session. In the first part, these 104 participants repeatedly had to bet on the colour of a ball after being shown differing samples of green and blue balls. There was no feedback given before the end of the experiment. In both types of sessions, one of the three parts would be paid out, with an exchange rate of 20 experimental currency units per Euro. We used ORSEE (Greiner, 2015) for recruitment. All sessions lasted between 60 and 90 minutes.

# 4 Framing effects on belief reports, behaviour, and the implications for belief-action consistency

### Predictions for Experiment 1-DISC

Recall that Experiment 1-DISC had participants play a pure discoordination game with four options. We illustrate which of the psychological processes would load on which options in Figure 3. As summarized in Table 3, we expected to observe *ex-ante* rationalization and a consensus effect in the population and random-other treatments, and *ex-ante* rationalization, *ex-post* rationalization, and wishful thinking in the opponent treatment. Consequently, we expected a lower probability mass on participants' own choices in the opponent treatment, leading

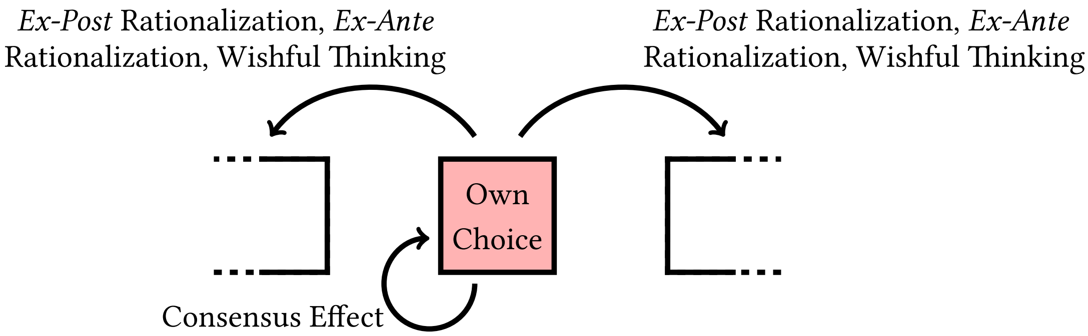



**Figure 3**: Predictions of the candidate processes in the discoordination game. We indicate the predictions by arrows: The consensus effect will increase the probability mass placed on the other player(s) making the same choice as the observed player, while the other three processes will increase the probability mass placed on the non-chosen options.

to higher observed best-response and lower observed 'worst-response' rates. A 'worst-response' means that the participant chooses the action his opponent is most likely to choose, as judged by the participant's reported belief.

### Results of Experiment 1-disC

Figure 4 summarizes beliefs and belief-action consistency for the three treatments.[12] For the analysis, we aggregate the data on the individual level across all periods, as we have one independent observation per participant (re-call that we did not give feedback). For each participant, we look at the probability that the reported belief places on the participant's own action in the corresponding game, averaged across all 24 periods. This is the participant's average subjective probability that (s)he matched the other player's/players' choice, and hence did *not* discoordinate. Similarly, we compute the best- and 'worst-response' rate to beliefs for each participant individually. Thus, the best-response rate measures how often a participant chose (one of) the action(s) that according to her reported belief was her opponent's least likely choice. And the worst-response rate measures the frequency with which a participant chose (one of) the action(s) that was her opponent's most likely choice.

[12] Figure A2 reproduces the same figures for the data from the first three periods only. It looks similar, but shows larger confidence intervals, and lower best- and higher worst-response rates.

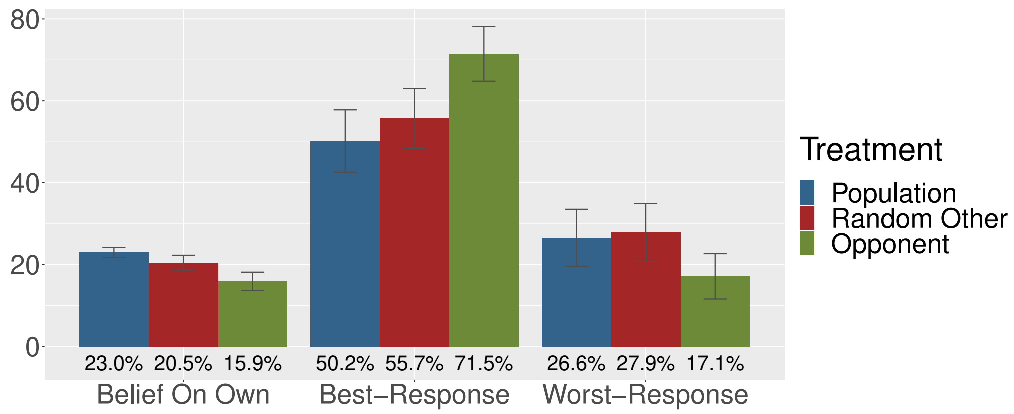



**Figure 4**: Beliefs and belief-action consistency (measured by observed best-response play and measured 'worst-response' play) in Experiment 1-DISC. Error bars indicate 95% confidence intervals. For all tests, the data is aggregated on the individual level across all periods, yielding one independent observation per participant.

The mean average belief on the participant's own action (Figure 4, left panel) is significantly higher in the population treatment and the random-other treatment compared to the opponent treatment (rank-sum tests, population/opponent: $p < 0.001$ and random-other/opponent: $p < 0.001$). The effect is strong enough to impede consistency: compared to the opponent treatment, the average observed best-response rate is lower (mid panel, $p < 0.001$ and $p = 0.004$) and the average worst-response rate is higher (right panel, $p = 0.026$ and $p = 0.019$) in the population treatment and the random-other treatment. The reduction in the observed best-response rate of 16-21 percentage-points and a 9.5 percentage-point increase in the worst-response rate in the population treatment are considerable effect sizes (in terms of the observed worst-response rates, the difference is more than 50% of the rate in the opponent treatment). The comparisons between population and random-other treatment yield $p = 0.146$ for the beliefs, $p = 0.237$ for the best-response rates, and $p = 0.822$ for the worst-response rates.

**Summary of Part 1**

Up to this point, we have documented a considerable framing effect. Most notably, beliefs differ in the *ceteris-paribus* comparison between the opponent and the random-other treatments, where we vary only whether a participant interacts directly with the 'target participant' of the belief. Additionally, the differences in reported beliefs influence observed best- and worst-response rates and hence affect the interpretation of actions and beliefs by the experimenter. What Experiment 1-DISC does not show is whether the differences between the treatments occur because there is (more) consensus under the population and random-other treatments, or because there is (more) wishful thinking, *ex-ante* or *ex-post* rationalization under the opponent treatment.[13] To disentangle these processes, we need Experiments 2-TYL and 2-BOS.

# 5 Disentangling the Processes

## 5.1 Experiment 2-TYL: Isolating Consensus Bias and Wishful Thinking

Experiment 2-TYL disentangles the influences of a consensus effect, and wishful thinking from *ex-ante*/*ex-post* rationalization. For this purpose, we use the "to-your-left game", in which a player wins a prize of 12€ if she chooses the option to the immediate left of the other player's choice (with the right-most option winning against the left-most option).

**Predictions for Experiment 2-TYL**

Figure 5 visualizes the predictions of our candidate processes in Experiment 2-TYL. Because the game is circular, only the relative position of the respective box matters and not the actual position.

[13]The fact that the average probability mass placed on a participants' own choice was below 25% for all treatments could be interpreted as suggesting that there is no consensus effect at all. However, recall that we are talking about a discoordination game in which it makes sense to choose the option that others are least likely to choose. Hence, probability masses of less than 25% are exactly what we should expect *a priori*. The consensus effect simply does not seem to be strong enough to distort beliefs so that the (average) probability mass surpasses 25%.

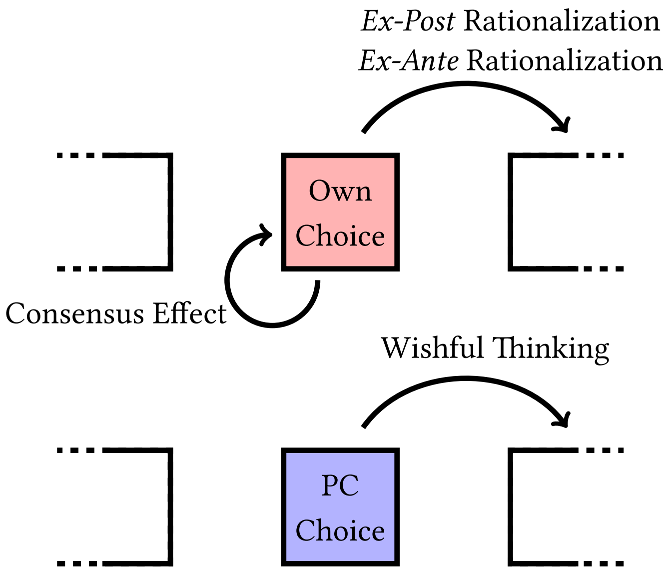



**Figure 5**: Predictions of the candidate processes in the to-your-left game with implementation errors in case of an implementation error. We color example choices and indicate the predictions by arrows: A consensus effect increases the probability mass placed on the other player(s) making the same choice as the observed player; *ex-post* and *ex-ante* rationalization increase the probability mass placed on the option to the right. Wishful thinking increases the probability of the option to the right of the option implemented by the computer.

In the to-your-left game, a consensus effect still would increase the belief-probability mass participants place on their own actions. *Ex-ante* and *ex-post* rationalization, and wishful thinking, on the other hand, would increase the probability mass on the option immediately to the right of participants' chosen actions.

To distinguish the effect of wishful thinking, we focus on periods in which the computer changed the selected box. In these periods, wishful thinking should increase the probability mass placed on the option to the right of the computer's choice. In contrast, *ex-ante* and *ex-post* rationalization yield a higher probability mass on the option to the right of the participant's choice. Depending on which box the computer selected, two different processes may increase the belief-probability mass on the same option. We control for this in the analysis.

### Results of Experiment 2-TYL

We analyze the data from Experiment 2-TYL with linear dummy regressions reported in Table 5. The dependent variable is the reported belief on a single box. Every participant reports 24 Periods $\times$ 4 Boxes = 96 belief probabilities on single boxes. We regress the beliefs on a set of dummies, indicating whether the particular reported probability would be influenced by an existing consensus effect, wishful thinking, or *ex-ante*/*ex-post* rationalization (EAR/EPR) according to the predictions indicated in Figure 5 above. Further, we use a treatment dummy which is equal to 1 in the random-other treatment and 0 in the opponent treatment. The constant of this regression is a neutral belief where all dummies are zero. Hence such a belief is unaffected by any of the processes we study.

Model 1 uses all observations where the participant made the ultimate decision.[14] Wishful thinking and EAR/EPR cannot be distinguished for the undistorted choices, as both load on the probability to the immediate right of the participant's choice. We hence have to use two separate regressions for the situations with and without implementation error because by design, the interaction EAR/EPR $\times$ Wishful Thinking is perfectly collinear with the implementation error.

Model 1 shows evidence for a consensus effect ("Belief on own") only in the random-other treatment. Further, probabilities to the right of the chosen option (influenced by EAR/EPR and/or wishful thinking) are twice the size of a neutral belief. The interaction of the "Belief to the right" with the random-other treatment ($p = 0.095$) suggests that the huge effect in the opponent treatment is reduced in the random-other treatment (the effect is substantial but the standard errors are non-negligible, too; at the same time, the effect shows up similarly in Table A2 where we use only 'implementation errors' that happened to coincide with the initial choice, $p = 0.074$). The effects are reflected in the left-hand panel of Figure 6.[15]

In our view, the reduction of "Beliefs to the the right" in the random-other treatment in Figure 6 stems from *ex-post* rationalization. *Ex-post* rationalization

[14] The observations where the computer truly altered the decision are analysed in Model 2. All results in Model 1 are robust to adding trials to the sample in which the computer decided but happened to choose the same action as the participant, as shown in Table A1 in Online Appendix A. A regression with only the trials in which the computer randomly implemented the same option as the participant shows qualitatively similar results (Table A2).

[15] Figure A3 reproduces the same figures for the data from the first three periods only. It looks similar, but shows larger confidence intervals, and larger differences for the case of implementation errors (right-hand panel).

| **Single Belief** | Model 1 | Model 2 |
|---|---|---|
| Belief on own | 0.4 (1.8) | 1.7 (1.7) |
| Belief on own × Random-Other Treatment | 7.7 (2.8)** | 0.1 (2.1) |
| Belief to the right (EAR/EPR & Wishful Thinking) | 19.8 (3.4)*** | |
| Belief to the right (EAR/EPR & Wishful Thinking) × Random-Other Treatment | -6.7 (3.9) | |
| EAR/EPR (to the right of the agent's choice) | | 9.8 (2.6)*** |
| EAR/EPR × Random-Other Treatment | | -2.3 (2.6) |
| Wishful thinking (to the right of the computer's choice) | | -0.1 (1.1) |
| Wishful thinking × Random-Other Treatment | | 1.8 (2.2) |
| Constant (Belief not directly affected by any of the processes) | 19.8 (0.7)*** | 22.2 (0.7)*** |
| Implementation error | No | Yes |
| Number of Observations | 3332 | 2532 |
| Number of Clusters | 70 | 70 |
| $R^2$ | 0.1247 | 0.0362 |

**Table 5:** Linear dummy regressions of the belief probability assigned to a given option on the processes that may affect that option's belief probability. Standard errors in parentheses clustered on subject level. EAR stands for *ex-ante* and EPR for *ex-post* rationalization. Asterisks: *** $p < 0.001$, ** $p < 0.01$, *$p < 0.05$.

should occur exclusively (or at least to a much larger degree) in the opponent treatment: believing that *some other* player chose an option that would be bad for us need not cause cognitive dissonance, because our opponent still might have chosen something else. In contrast, if we state a belief that our *opponent* chose something that would be bad for us given our action, this should indeed cause cognitive dissonance in us. Therefore, the coefficient of "Belief to the right" (with Frame = 0) should capture the added effects of *ex-ante* and *ex-post* rationalization. In contrast, the "Belief to the right" in the random-other treatment (Frame = 1) should capture *ex-ante* rationalization only. Hence, the interaction effect "Belief to the right × Frame" provides an estimate for the differential effect of *ex-post* rationalization. Like in Experiment 1-DISC, the average best-response rate is higher in the opponent treatment than in the random-other treatment when the computer does not change the decision (opponent: 62.1%, random other: 45.2%, rank-sum test $p = 0.006$; the difference in worst-response rates yields $p = 0.780$; opponent: 20.9%, random other: 22.8%).

An additional experiment reported in Online Appendix C provides more direct evidence for the treatment difference in *ex-post* rationalization. The setup

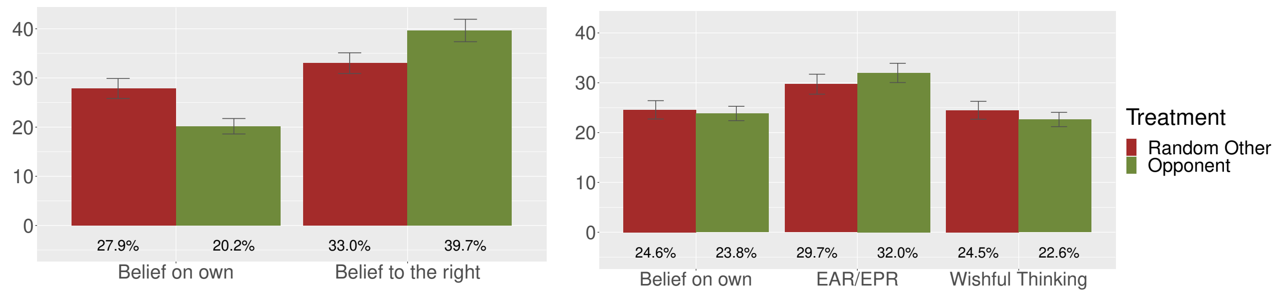



**Figure 6**: Beliefs in Experiment 2-TYL. Error bars indicate 95% confidence intervals. The left-hand panel displays choices without implementation error, the right-hand panel those after an implementation error that yielded a different action that the participant's choice (EAR/EPR refers to *ex-ante*/*ex-post* rationalization which both correspond to the option to the right of the particiant's choice, whereas wishful thinking corresponds to the option to the right of the computer's choice). The right-hand panel's categories are not mutually exclusive, however, and thus, informative only to a limited degree.

mirrors that of Experiment 1-DISC, except that we ask for beliefs before actions. The reversed order should eliminate *ex-post* rationalization as *ex-post* rationalizing a belief by an action is unintuitive: once we form a belief (as in the first stages of the additional experiment), there is no good reason to form yet a different belief that we then contradict out of a taste for consistency. We indeed no longer find a difference between the treatments, which is due to players placing a higher probability mass on their own action in the opponent treatment, in line with our prediction.

Model 2 in Table 5 includes all decisions where the computer really changed the participant's decision. Hence, Model 2 includes all observations in which the computer decided and did not choose the same action as the participant. There is no more consensus effect in either treatment. Also, there is no evidence for wishful thinking. However, EAR/EPR loads on beliefs to the right of the participant's decision also in the randomly altered trials.[16] This is not reflected well in the right-hand panel of Figure 6. Note, however, that in contrast to the regression analysis, the different effects are not well-separated in the right-hand panel of

[16] As an anonymous reviewer pointed out, one may argue that the substantial reduction of the EAR/EPR-coefficient between Models 1 and 2 is evidence of a different type of wishful thinking: Once my choice has been altered, "I wish I was wrong." However, it remains unclear where the corresponding probability mass goes. If "I wish I was wrong" had such an impact, why is there absolutely no effect in terms of "I wish the computer made the right choice for me"?

Figure 6. Finally, (neutral) beliefs are closer to uniformity in the random-action trials.

### Discussion of Experiment 2-TYL

We interpret the results in the following way: there is a consensus effect in the random-other treatment. There is *ex-ante* or *ex-post* rationalization in both treatments, but the effect tends to be stronger in the opponent treatment. We argue that the apparent difference is due to *ex-post* rationalization being less important or absent in the random-other treatment and confirm this conjecture in the additional experiment reported in Online Appendix C. As in Experiment 1-DISC, the framing differences in Model 1 affect measured belief-action consistency, with higher observed best-response rates under the opponent treatment compared to the random-other treatment.

When the computer overrides participants' decisions, a certain degree of *ex-ante* rationalization survives in the reported beliefs: also in such cases, participants *on average* seem to report beliefs that make sense given their actions, despite the fact that beliefs are closer to uniformity.[17] However, there are no more significant framing differences in beliefs or best-response rates with implementation errors. It seems as if the random implementation error detaches participants to a certain degree from the action choice altogether. We also do not see any evidence for wishful thinking, even though wishful thinking does not relate to the chosen action.

We ran Experiment 2-TYL to disentangle consensus effect and—albeit with a caveat—wishful thinking from *ex-ante*/*ex-post* rationalization. Experiment 2-BOS shows that there is as much of a consensus effect in an opponent treatment as in a population treatment, once we eliminate the cognitive need for *ex-post* rationalization in the opponent treatment.

[17] The reduced average difference to uniformity is only very partially due to a difference in the prevalence of uniform beliefs: under implementation errors, 5% of the reported beliefs are uniform, and without errors, 4%.

## 5.2 Experiment 2-BOS: Consensus Effect in Opponent Treatments?

In Experiment 2-BOS, participants play two rounds of the battle-of-the-sexes game with alternating but unobservable moves depicted in Figure 2, with role-reversal between the games, belief-elicitation before choices, and random rematching between rounds. To study whether a consensus effect exists also in an opponent treatment, we contrast beliefs in such a treatment with beliefs from a population treatment (where we know a strong consensus effect exists).

To give a concrete example of the timeline of Experiment 2-BOS, a participant starting in role of Player B in the population treatment would first be asked about her belief what the (round-one) population of Players A will do. Then, she would make her choice as Player B. Proceding to round two, she would report her belief about the (second-round) population of Players B before finally making her choice as Player A.

### Predictions for Experiment 2-BOS

As we outlined above, cognitive dissonance should not affect behaviour in Experiment 2-BOS, neither in the population nor in the opponent treatment. Hence, *ex-post* rationalization should be eliminated in the opponent treatment. If under an opponent frame, a consensus effect does not exist, we should nevertheless see a treatment difference: in that case, the probability mass placed on a participant's prior action should be higher in the population frame (where we know the consensus effect is at work) than in the opponent frame. If, on the other hand, there is a consensus effect in the opponent frame that is just 'over-written' by *ex-post* rationalization in more standard designs (such as Experiment 1-DISC or Experiment 2-TYL), we should no longer see a difference between the treatments.

### Results of Experiment 2-BOS

The data generally look as expected given the literature. Participants in both player roles chose "left" far more often than "right": 74% of As and 70% of Bs in the first game, and 75% of As and 76% of Bs in the second game. These fractions roughly correspond to participants' beliefs: in both games, As expected Bs to play "left" with an average probability of 50-51% (40% would make linear-utility

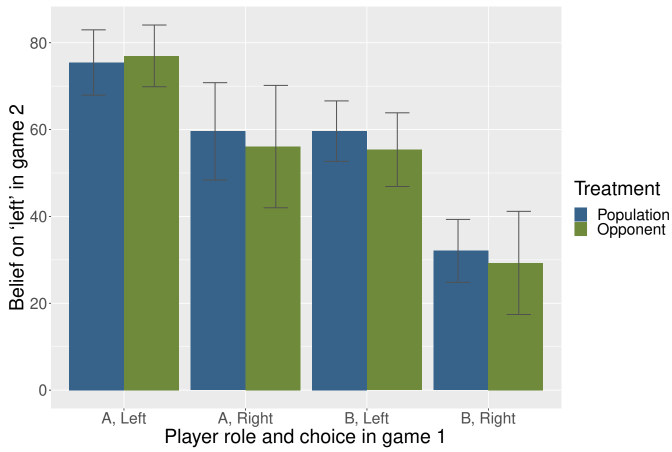



**Figure 7**: Probability mass placed on "left" in participants' belief reports for game 2, by their role and decision in game 1. Error bars indicate 95% confidence intervals.

As indifferent), and Bs expected As to play "left" with an average probability of 71%.

Given that there are no 'surprises' in the choice data, we now focus on our research question and look at participants' beliefs for game 2 depending on their choices in game 1. Figure 7 visualizes the results for both player roles and both treatments. First of all, note that we observe a clear consensus effect for either role in both treatments: players who chose "left" in game 1 place more probability mass on others (who 'now'—in game 2—have the role they used to have in game 1) also choosing "left", compared to players who chose "right". This holds for both players A and B. Moreover, there clearly are no more treatment differences between the opponent treatment and the population treatment (which also holds for best-response rates: 75% in the opponent *vs* 71% in the population treatment; Boschloo test $p = 0.383$). Note that for the computation of best-responses, we compare game-1 beliefs with game-1 actions and game-2 beliefs with game-2 actions.

To support the conclusion statistically, we run the linear-probability regression reported in Table 6. As can be seen from the Table, the participant's previous choice (when the participant was playing in the role that the belief's target is playing now) clearly has an influence on the belief, while the treatment variable (or any of its interactions) does not.

Our results mean that when participants do not have any need to *ex-post* ra-

| | Belief on Left (in %) |
|---|---|
| (Intercept) | 34.8 (4.3)*** |
| Person A in Game 1 | 18.8 (4.4)*** |
| Chose "Left" in Game 1 | 23.5 (4.8)*** |
| Opponent Frame | −3.1 (6.6) |
| Person A in Game 1 × Opponent Frame | 4.1 (6.3) |
| Chose "Left" in Game 1 × Opponent Frame | −0.6 (6.9) |
| Number of Observations | 222 |
| $R^2$ | 0.306 |

**Table 6:** Linear dummy regressions of the probability mass placed on "left" for game 2, on the participant's role and decision in game 1. Standard errors in parentheses. Asterisks: *** $p < 0.001$, ** $p < 0.01$, *$p < 0.05$, $^{\cdot}p < 0.1$.

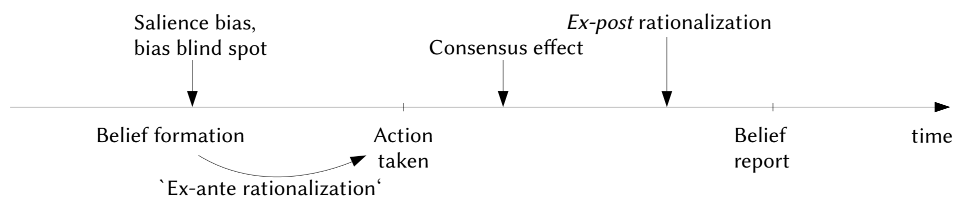



**Figure 8**: Timing of when and which processes are assumed to be active, taking into account this paper's findings. The Figure is reduced to the three processes we find evidence for, and it implies that the consensus effect and *ex-post* rationalization are serial processes rather than alternative processes that happen at the same time.

tionalize their actions, they exhibit the same degree of consensus effect under an opponent frame as under a population frame. As a consequence, we have to revise our conceptual picture from Section 2. Figure 8 shows the updated 'model' of participants' belief-report formation. It is reduced to the three processes we find evidence for, and it implies that the consensus effect and *ex-post* rationalization are not two alternative processes that might take effect at a similar point in time. Instead, consensus effect and *ex-post* rationalization seem to be *serial* processes that may be invoked one after the other.

# 6 Conclusion

When studying beliefs, researchers have several choices to make, among them, whether to ask participants about the actions of their opponent(s) or about the actions of unrelated others.[18] None of these choices is trivial, and a review of the literature reveals that different researchers make different choices. However, the choices are rarely motivated in the final publication. We claim that the reason is that the exact consequences of each alternative are unknown so far.

In this paper, we show that in particular the choice between an opponent treatment (asking about the opponent's action), a random-other treatment (asking about somebody else's action), and a population treatment (asking about everybody else's action) is by no means innocuous. Asking about others' choices induces belief reports to be affected by a consensus effect in any treatment. However, if the study uses an opponent treatment and actions are strategic substitutes, the latent belief changes (again). In such cases, the reported belief will reflect *ex-post* rationalization.

Our findings thus provide an explanation for the puzzle that, so far, all economics papers documenting a consensus effect have relied on a population treatment: It is only when actions are strategic substitutes that we can discern a consensus effect from *ex-ante* (or *ex-post*) rationalization. However, when actions are substitutes, reporting a belief that is influenced by a consensus effect seems particularly 'bad'. It would mean the participant expects others to make the same choice with a comparatively high probability, in which case the participant should have made a different choice to begin with. This is precisely the type of situation in which an opponent-oriented question leads to cognitive dissonance, and thus, *ex-post* rationalization (random-other and population treatments always offer an excuse for belief-action inconsistencies in that "*my* opponent is different"). In other words, in settings that allow to single out a consensus effect, we will observe the effect only under a belief-elicitation task that does *not* target the participants' opponent.

Our second research question was whether the literature was overlooking

[18]Researchers may avoid asking about a participant's opponent even in a one-shot design because they are afraid of hedging attempts by their participants, which is not an issue in our study. In the discoordination games we study, increased hedging when asking about the opponent would lead to the exact opposite of what we find. Further, we preclude rational hedging by never paying both an action and the corresponding belief.

other processes that are relevant for belief reports on top of *ex-ante* rationalization, the consensus effect, and *ex-post* rationalization. This would not have been surprising given the huge number of known biases in the literature. In adding potential biases to the list, we restricted ourselves to biases that we could easily apply given our main interest in understanding the interplay of belief-elicitation treatments with the three 'standard' processes.

Reassuringly for our interpretation of the literature, we find clear effects consistent only with *ex-ante* rationalization, a consensus effect, and *ex-post* rationalization. And while we cannot identify the exact process behind participants' *ex-post* rationalization, such rationalization shows exactly in those cases when cognitive dissonance or a social-desirability bias (assuming consistent behaviour to be socially desirable) would suggest it should show.

**Recommendations.** Our results show that we need to take the substantial framing differences into account when analysing existing data or designing new surveys and experiments. In particular, in designing new experiments, we propose to use random-other or population treatments, even though the reports still will be influenced by social projection. Choosing the alternative—an opponent treatment—means that reported beliefs may lose any connection to the 'true beliefs' (the belief at the time of choosing the action) altogether.

The danger of reports being disconnected from the 'true beliefs' is present particularly when actions are strategic substitutes. Having said this, note that the reports may be closer to the 'true beliefs' in opponent treatments. In particular, the reports even may match the 'true beliefs' if the two effects exactly cancel each other out. However, there is no way of assessing the relative strength of the two effects. Because of this, we prefer the treatments that trigger only one of the effects, given that they at least afford a clear interpretation (*e.g.*, as providing a lower bound for best-response play in the discoordination game).

We also recommend considering to elicit beliefs prior to actions, given that this will prevent consensus effects and *ex-post* rationalization (*cf.*, *e.g.*, Martinangeli, 2021, for an elaborate application of this approach). In our experience, it does not lead to excessively high measured best-response rates (a common concern against such a procedure; see, *e.g.*, the additional experiment reported in Online Appendix C). However, we already know that under certain circumstances, it will change behaviour (Rutström and Wilcox, 2009).[19]

Our findings suggest that it may be impossible to elicit the true beliefs that participants hold at the time of choosing their action. In our study, participants faced a strong monetary incentive to report their true beliefs. Moreover, we incentivized belief reports by a state-of-the-art mechanism that is proper even for people who do not comply with expected-utility maximization (as long as they comply with a weak monotonicity condition; Hossain & Okui, 2013). And still, we have not found a way of asking for a belief that leads to an unbiased belief report without running the risk of changing behaviour.

[19]An alternative might be eliciting beliefs on the same screen as actions, as done by, *e.g.*, Peeters and Vorsatz (2021).

# References

al-Nowaihi, A., & Dhami, S. (2015). Evidential Equilibria: Heuristics and Biases in Static Games of Complete Information. *Games*, 6(4), 637-676.

Armantier, O., & Treich, N. (2013). Eliciting beliefs: Proper scoring rules, incentives, stakes and hedging. *European Economic Review*, 62, 17-40.

Babad, E., & Katz, Y. (1991). Wishful thinking—against all odds. *Journal of Applied Social Psychology*, 21(23), 1921-1938.

Bar-Hillel, M., & Budescu, D. V. (1995). The elusive wishful thinking effect. *Thinking & Reasoning*, 1(1), 71-103.

Bar-Hillel, M., Budescu, D. V., & Amar, M. (2008). Predicting World Cup results: Do goals seem more likely when they pay off? *Psychonomic Bulletin & Review*, 15(2), 278-283.

Bauer, D., & Wolff, I. (2018). Biases in Beliefs: Experimental Evidence. *TWI Research Paper* 109.

Bellemare, C., Kröger, S., & Van Soest, A. (2008). Measuring inequity aversion in a heterogeneous population using experimental decisions and subjective probabilities. *Econometrica*, 76(4), 815-839.

Blanco, M., Engelmann, D., Koch, A. K., & Normann, H. T. (2010). Belief elicitation in experiments: is there a hedging problem?. *Experimental Economics*, 13(4), 412-438.

Blanco, M., Engelmann, D., Koch, A. K., & Normann, H. T. (2014). Preferences and beliefs in a sequential social dilemma: a within-subjects analysis. *Games and Economic Behavior*, 87, 122-135.

Breitmoser, Y. (2015). Knowing me, imagining you: Projection and overbidding in auctions. *Games and Economic Behavior*, 113, 423-447.

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

Charness, G., & Grosskopf, B. (2001). Relative payoffs and happiness: an experimental study. *Journal of Economic Behavior & Organization*, 45(3), 301-328.

Charness, G., & Levin, D. (2005). When optimal choices feel wrong: A laboratory study of Bayesian updating, complexity, and affect. *The American Economic Review*, 95(4), 1300-1309.

Chater, N. (2018). *The mind is flat: the remarkable shallowness of the improvising brain.* Yale University Press, New Haven, USA.

Cooper, R., DeJong, D., Forsythe, R. and Ross, T. (1993). Forward Induction in the Battle-of-the-Sexes Games. *American Economic Review*, 83(5), 1303-1316.

Costa-Gomes, M. A., & Weizsäcker, G. (2008). Stated beliefs and play in normal-form games. *The Review of Economic Studies*, 75(3), 729-762.

Crosetto, P., Filippin, A., Katuscak, P. & Smith, J. (2020). Central tendency bias in belief elicitation. *Journal of Economic Psychology*, 78, 102273.

Danz, D. N., Fehr, D., & Kübler, D. (2012). Information and beliefs in a repeated normal-form game. *Experimental Economics*, 15(4), 622-640.

Danz, D. N., Madarász, K., & Wang, S. W. (2014). The Biases of Others: Anticipating Informational Projection in an Agency Setting. *Working Paper*, accessed 2017/06/06, http://works.bepress.com/kristof_madarasz/42/ ,

Dawes, R. M., & Mulford, M. (1996). The false consensus effect and overconfidence: Flaws in judgment or flaws in how we study judgment?. *Organizational Behavior and Human Decision Processes*, 65(3), 201-211.

Delavande, A., Giné, X., & McKenzie, D. (2011a). Eliciting probabilistic expectations with visual aids in developing countries: how sensitive are answers to variations in elicitation design? *Journal of Applied Econometrics*, 26(3), 479-497.

Delavande, A., Giné, X., & McKenzie, D. (2011b). Measuring subjective expectations in developing countries: A critical review and new evidence. *Journal of Development Economics*, 94(2), 151-163.

Dhami, S. (2016). *The Foundations of Behavioral Economic Analysis*. Oxford University Press, Oxford, UK.

Edwards, A. (1953). The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *Journal of Applied Psychology*, 37(2), 90-93.

Edwards, W. (1968). Conservatism in human information processing. In: Kleinmutz, B. (Ed.), Formal Representation of Human Judgement. New York: Wiley, 17-52.

Engelberg, J., Manski, C. F., & Williams, J. (2011). Assessing the temporal variation of macroeconomic forecasts by a panel of changing composition. *Journal of Applied Econometrics*, 26(7), 1059-1078.

Engelmann, D., & Strobel, M. (2000). The False Consensus Effect Disappears if Representative Information and Monetary Incentives Are Given. *Experimental Economics*, 3(3), 241-260.

Engelmann, D., & Strobel, M. (2012). Deconstruction and reconstruction of an anomaly. *Games and Economic Behavior*, 76(2), 678-689.

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology*, 87(3), 327.

Epley, N., & Gilovich, T. (2016). The mechanics of motivated reasoning. *The Journal of Economic Perspectives*, 30(3), 133-140.

Erkal, N., Gangadharan, L., & Koh, B. H. (2020). Replication: Belief elicitation with quadratic and binarized scoring rules. *Journal of Economic Psychology*, 81, 102315.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3), 817-868.

Festinger, L. (1957). A theory of cognitive dissonance. Stanford: Stanford University Press.

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

Gigerenzer, G., & Selten, R. (Eds.). (2001). *Bounded rationality: The adaptive toolbox*. Cambridge: MIT press.

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. New York, Cambridge university press.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association*, 1(1), 114-125.

Guiso, L., & Parigi, G. (1999). Investment and demand uncertainty. *The Quarterly Journal of Economics*, 114(1), 185-227.

Harrison, G. W., Martínez-Correa, J., & Swarthout, J. T. (2014). Eliciting subjective probabilities with binary lotteries. *Journal of Economic Behavior & Organization*, 101, 128-140.

Harris, A. J., & Hahn, U. (2011). Unrealistic optimism about future life events: a cautionary note. *Psychological Review*, 118(1), 135.

Heger, S. A., & Papageorge, N. W. (2018). Whe should *totally* open a restaurant: How optimism and overconfidence affect beliefs. *Journal of Economic Psychology*, 67, 177-190.

Helweg-Larsen, M., & Shepperd, J. A. (2001). Do moderators of the optimistic bias affect personal or target risk estimates? A review of the literature. *Personality and Social Psychology Review*, 5(1), 74-95.

Hossain, T., & Okui, R. (2013). The binarized scoring rule. *The Review of Economic Studies*, 80(3), 984-1001.

Holt, C. A., & Smith, A. M. (2016). Belief Elicitation with a Synchronized Lottery Choice Menu That Is Invariant to Risk Attitudes. *American Economic Journal: Microeconomics*, 8(1), 110-139

Hyndman, K. B., Terracol, A., & Vaksmann, J. (2013). Beliefs and (in)stability in normal-form games. *Working paper*, accessed 2017/06/14, http://lemma.u-paris2.fr/sites/default/files/concoursMCF/Vaksman.pdf.

Hyndman, K., Ozbay, E. Y., Schotter, A., & Ehrblatt, W. Z. (2012). Convergence: an experimental study of teaching and learning in repeated games. *Journal of the European Economic Association*, 10(3), 573-604.

Iriberri, N., & Rey-Biel, P. (2013). Elicited beliefs and social information in modified dictator games: What do dictators believe other dictators do? *Quantitative Economics*, 4(3), 515-547.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80(4), 237-251.

Karni, E. (2009). A mechanism for eliciting probabilities. *Econometrica*, 77(2), 603-606.

Khwaja, A., Sloan, F., & Salm, M. (2006). Evidence on preferences and subjective beliefs of risk takers: The case of smokers. *International Journal of Industrial Organization*, 24(4), 667-682.

Krizan, Z., & Windschitl, P. D. (2007). The influence of outcome desirability on optimism. *Psychological Bulletin*, 133(1), 95.

Krueger, J. I. (2007). From social projection to social behaviour. *European Review of Social Psychology*, 18, 1-35.

Krueger, J. I. (2013). Social projection as a source of cooperation. *Current Directions in Psychological Science*, 22(4), 289-294.

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology*, 32, 311-328.

Larwood, L., & Whittaker, W. (1977). Managerial myopia: Self-serving biases in organizational planning. *Journal of Applied Psychology*, 62(2), 194

Manski, C. F. (2002). Identification of decision rules in experiments on simple games of proposal and response. *European Economic Review*, 46(4), 880-891.

Manski, C. F., & Neri, C. (2013) First- and second-order subjective expectations in strategic decision-making: Experimental evidence. *Games and Economic Behavior*, 81, 232-254.

Marks, G., & Miller, N. (1987). Ten years of research on the false consensus effect: An empirical and theoretical review. *Psychological Bulletin*, 102(1), 72.

Martinangeli, A. F. M. (2021). Do what (you think) the rich will do: Inequality and belief heterogeneity in public good provision. *Journal of Economic Psychology*, 83, 102364.

McKelvey, R. D., & Page, T. (1990). Public and private information: An experimental study of information pooling. *Econometrica*, 58, 1321-1339.

Mullen, B., Atkins, J. L., Champion, D. S., Edwards, C., Hardy, D., Story, J. E., & Vanderklok, M. (1985). The false consensus effect: A meta-analysis of 115 hypothesis tests. *Journal of Experimental Social Psychology*, 21(3), 262-283.

Molnár, A., & Heintz, C. (2016). Beliefs About People's Prosociality: Eliciting predictions in dictator games. *Working Paper*, accessed 2017/09/06, http://publications.ceu.edu/sites/default/files/publications/molnar-heintz-beliefs-about-prosociality.pdf

Nyarko, Y., & Schotter, A. (2002). An experimental study of belief learning using elicited beliefs. *Econometrica*, 70(3), 971-1005.

Palfrey, T. R., & Wang, S. W. (2009). On eliciting beliefs in strategic games. *Journal of Economic Behavior & Organization*, 71(2), 98-109.

Peeters, R., & Vorsatz, M. (2021). Simple guilt and cooperation. *Journal of Economic Psychology*, 82, 102347.

Pronin, E., Lin, D. Y., & Ross, L. (2002). The Bias Blind Spot: Perceptions of Bias in Self Versus Others. *Personality and Social Psychology Bulletin*, 28(3), 369-381.

Proto, E., & Sgroi, D. (2017). Biased beliefs and imperfect information. *Journal of Economic Behavior & Organization*, 136, 186-202.

Rey-Biel, P. (2009) Equilibrium play and best response to (stated) beliefs in normal form games, *Games and Economic Behavior*, 65(2), 572-585.

Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, 13(3), 279-301.

Rubinstein, A., & Salant, Y. (2015). "Isn't everyone like me?": On the presence of self-similarity in strategic interactions. Working paper version of Rubinstein & Salant (2016), accessed 2021/07/05, https://en-econ.tau.ac.il/sites/economy_en.tau.ac.il/files/media_server/Economics/foerder/papers/2-2015.pdf

Rubinstein, A., & Salant, Y. (2016). "Isn't everyone like me?": On the presence of self-similarity in strategic interactions. *Judgment and Decision Making*, 11(2), 168.

Rutström, E. E., & Wilcox, N. T. (2009). Stated beliefs versus inferred beliefs: A methodological inquiry and experimental test. *Games and Economic Behavior*, 67(2), 616-632.

Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica*, 5(17), 61-71.

Savage, L. J. (1954) The Foundations of Statistics. New York: John Wiley and Sons. (Second ed., Dover, 1972).

Selten, R., & Ockenfels, A. (1998). An experimental solidarity game. *Journal of Economic Behavior & Organization*, 34(4), 517-539.

Schlag, K. H., Tremewan, J., & Van der Weele, J. J. (2015). A penny for your thoughts: a survey of methods for eliciting beliefs. *Experimental Economics*, 18(3), 457-490.

Schotter, A., & Trevino, I. (2014). Belief elicitation in the laboratory. *Annual Review of Economics*, 6(1), 103-128.

Shah, P., Harris, A. J., Bird, G., Catmur, C., & Hahn, U. (2016). A pessimistic view of optimistic belief updating. *Cognitive Psychology*, 90, 71-127.

Sutter, M., Czermak, S., & Feri, F. (2013). Strategic sophistication of individuals and teams. Experimental evidence. *European Economic Review*, 64, 395-410.

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

Taylor, S. E., & Fiske, S. T. (1975). Point of view and perceptions of causality. *Journal of Personality and Social Psychology*, 32(3), 439–445.

Trautmann, S. T., & van de Kuilen, G. (2015). Belief elicitation: A horse race among truth serums. *The Economic Journal*, 125(589), 2116-2135.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2), 207-232.

Tversky, A., & Kahneman, D. (1974). Heuristics and biases: Judgment under uncertainty. *Science*, 185, 1124-1130.

Van Der Heijden, E., Nelissen, J., & Potters, J. (2007). Opinions on the tax deductibility of mortgages and the consensus effect. *De Economist*, 155(2), 141-159.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12(3), 129-140.

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, 39(5), 806.

Weinstein, N. D. (1989). Effects of personal experience on self-protective behaviour. *Psychological Bulletin*, 105(1), 31.

Wolff, I. (2018). If I Don't Trust Your Preferences, I Won't Follow Mine: Preference Stability, Beliefs, and Strategic Choice. *TWI Research Paper* 113.

Wolff, I. (2021). The lottery player's fallacy: Why labels predict strategic choices. *Journal of Economic Behavior & Organization*, 184: 16–29.

# Supplementary Materials for “Biases in Belief Reports” by Dominik Folli & Irenaeus Wolff

## Appendix A  Figures & Tables

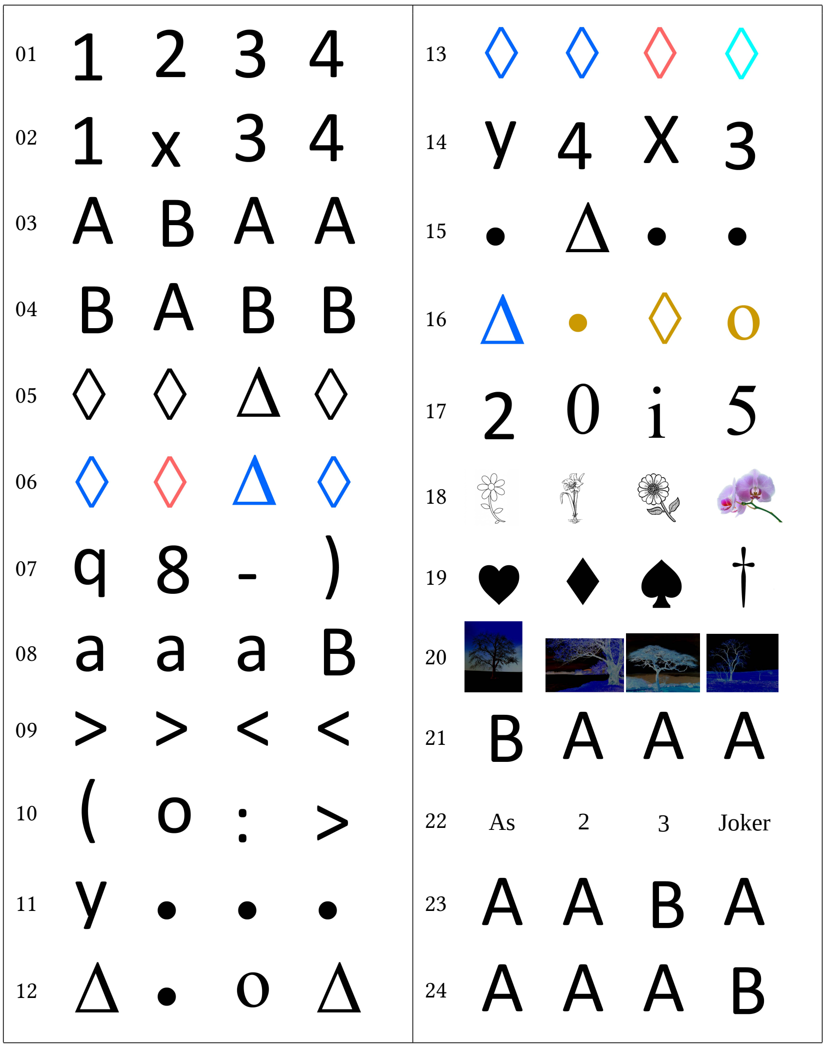



**Figure A1**: The 24 label sets, used to label the four options of the game. One set for each period.

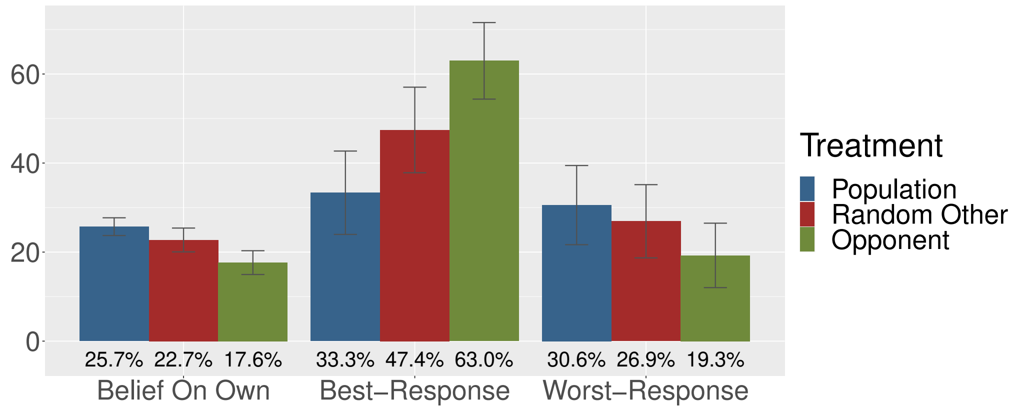



**Figure A2**: Beliefs and belief-action consistency (measured by observed best-response play and measured 'worst-response' play) in the first three periods of Experiment 1-DISC. Error bars indicate 95% confidence intervals.

| **Single Belief** | **Model 1′** | |
|---|---|---|
| Belief on own | 0.809 | (1.738) |
| Belief on own × Random-Other Treatment | 7.009* | (2.654) |
| Belief to the right (EAR/EPR & Wishful Thinking) | 19.00*** | (3.103) |
| Belief to the right (EAR/EPR & Wishful Thinking) × Random-Other Treatment | -6.375 | (3.510) |
| Constant | 19.97*** | (0.691) |
| Observations | 4188 | |
| $R^2$ | 0.117 | |

**Table A1:** OLS dummy regressions of single belief elements with interactions including trials in which the computer happened to select the same action as the participant. Standard errors in parenthesis clustered on subject level (70 clusters). Asterisks: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

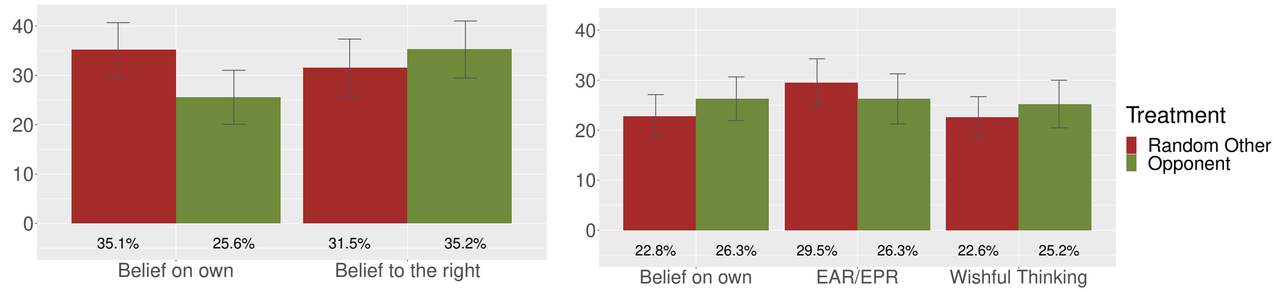



**Figure A3**: Beliefs in the first three periods of Experiment 2-TYL. Error bars indicate 95% confidence intervals. The left-hand panel displays choices without implementation error, the right-hand panel those after an implementation error that yielded a different action that the participant's choice. The right-hand panel's categories are not mutually exclusive, however, and thus, informative only to a limited degree.

| **Single Belief** | **Model 1″** | |
|---|---|---|
| Belief on own | 2.415 | (2.288) |
| Belief on own × Random-Other Treatment | 4.550 | (3.091) |
| Belief to the right (EAR/EPR & Wishful Thinking) | 15.94*** | (3.367) |
| Belief to the right (EAR/EPR & Wishful Thinking) × Random-Other Treatment | -5.613 | (3.639) |
| Constant | 20.54*** | (0.970) |
| Observations | 856 | |
| $R^2$ | 0.0894 | |

**Table A2:** OLS dummy regressions of single belief elements with interactions of only trials in which the computer happened to select the same action as the participant. Standard errors in parenthesis clustered on subject level (68 clusters). Asterisks: *** $p < 0.001$.

# Appendix B Salience bias, bias blind spot, and 'non-applicable' biases

In this section, we briefly present the processes that we left 'out of the equation'.

**Salience bias.** Suppose there is a salience pattern $s, s = (s_1, s_2, ..., s_N)$. Then, a decision-maker who is affected by salience bias will respond suboptimally to $b_0$:

$$prob(c^{(i)}) = \begin{cases} (1-\sigma)1 + \sigma s_i, & c^{(i)} \in BR(b_0), \\ \sigma s_i, & c^{(i)} \notin BR(b_0), \end{cases}$$

where $\sigma$ determines how much the decision is influenced by salience. The report will be optimal given the prior in all treatments: $r = \beta_0$.

**Bias blind spot (with salience bias).** Under the bias blind spot with respect to a salience bias, the agent believes everybody else to be affected by a salience bias, thus updating her prior $b_0$ to a $b_0^{sal}$:

$$b_0^{sal} = (1-\tau)b_0 + \tau s,$$

where $\tau$ corresponds to how strong the agent thinks the opponent will be affected by salience. Then, overlooking her own salience bias, the agent will be choosing according to:

$$prob(c^{(i)}) = \begin{cases} (1-\sigma)1 + \sigma s_i, & c^{(i)} \in BR(b_0^{sal}), \\ \sigma s_i, & c^{(i)} \notin BR(b_0^{sal}). \end{cases}$$

**Base-rate fallacy.** The base-rate fallacy is the opposite of conservatism in updating as it means ignoring prior probabilities. The only prior that could possibly be ignored is $b_0$. This simply would correspond to an extreme consensus bias.

**Belief bias.** The belief bias means the bias that people tend to think an argument must be correct if its conclusion is correct. In our case, we might interpret $c$ as the conclusion and $b_0$ as the argument. Then, the bias would simply correspond to a confirmation of *ex-ante* rationalization.

**Conjunction fallacy.** This is the fallacy that corresponds to the famous "Linda problem": ignoring that the conjunction of two events cannot be more likely than either event alone. In our case, there are no two events the likelihood of whose conjunction the agent would have to assess.

**Contrast effect.** More attention is paid to characteristics that change strongly. In our study, this might (only) be relevant for the boxes. *E.g.*, the first set of labelled boxes is (1,2,3,4), whereas the second set is (1,2,x,4). This may increase the salience of the third box labelled by "x". Which simply would mean a particular type of $\rightarrow$ salience bias.

**Fundamental attribution error.** When the agent attributes too much of somebody else's behaviour to that person's characteristics rather than to the characteristics of the situation. For the fundamental attribution error to play any role, the agent thus would have to observe a person's actions, which is not the case in any of the experiments.

**Gambler's fallacy.** The gambler's fallacy means that an agent believes that prior realisations of an i.i.d. process change future probabilities in such a way that the observed mean moves towards its expected value. In general, this presupposes the observation of an outcome, which is not given in our study.

**Hindsight bias.** An agent who shows hindsight bias is unable to abstract from knowledge acquired after a choice was made, when assessing that choice. As we write in the notes to Table 2, a hindsight bias could in principle apply to our setting in a specific way. An agent knows $c^{(i)}$ at the time of stating $r = \beta_1$. If the agent cannot abstract from the knowledge about $c^{(i)}$ when forming $\beta_1$, this can be interpreted in the way that she thinks that her opponent must have known $c^{(i)}$ when making his own choice in the game (or that the agent at least forms $\beta_1$ as if the opponent had known $c^{(i)}$). But if the opponent knew $c^{(i)}$ at the time of making his choice, the opponent will have chosen some $c^{(j)} \in BR(\mathbf{c})$, where $\mathbf{c} = (c_1, c_2, ..., c_N)$ with $c_k = 1$ for $c^{(i)} = a_k$, and 0, otherwise. In the opponent treatment, this means that $b_1$ will be such that $c^{(j)} \in BR(\mathbf{c})$. In the population treatment, $B_1$ will be adjusted such that a best-response to $c^{(i)}$ is slightly more likely (by at most $1/P$), whereas in the random-other treatment, $b_t^-$ will not be adjusted at all. We explored the possibility of a Hindsight bias affecting beliefs in the working-paper version Bauer & Wolff (2018), finding no evidence for it. We therefore chose to leave this specific interpretation of the hindsight bias out of the paper in order to reduce complexity at least to some degree (this interpretation is the only one we could think of how hindsight bias may affect our data).

**Hot-hand fallacy.** An agent who shows the hot-hand fallacy believes that somebody who has been lucky several times in a row is more likely to be lucky the next time as well. In general, this fallacy also presupposes the observation of realizations of lotteries, which is not given in our study.

***Status-quo* bias.** A preference for the current state relative to any changes, irrespective of what the current state is. Generally, there are no observable changes in the experiment (again, there is no feedback of any sort). Only if we interpret the agent's own choice $c$ as an event that may upset the '*status quo*' $b_0$, we could posit that there might be an effect—which, however, would simply reinforce the best-response relationship that *ex-ante*-rationalization gives rise to.

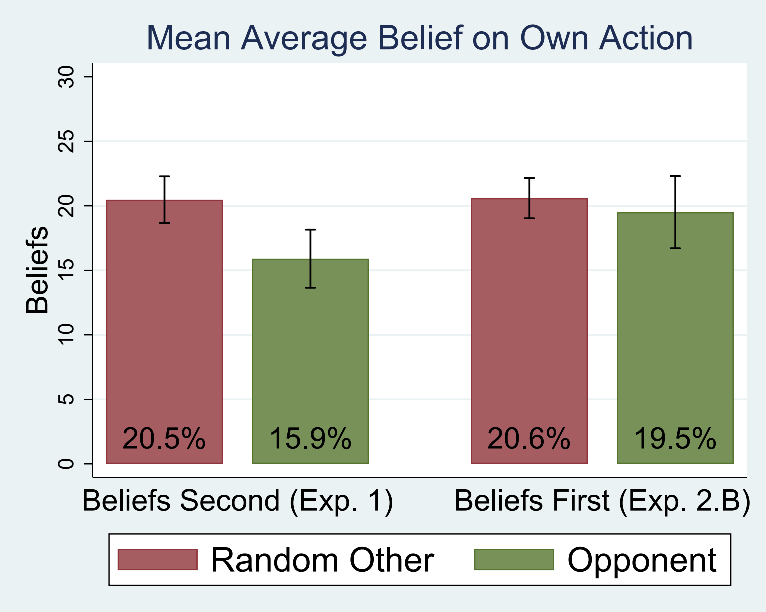



**Figure C1**: Beliefs in the Beliefs-First and the Beliefs-Second treatments. Error bars indicate 95% confidence intervals. For all tests, the data is aggregated on the individual level across all periods yielding one independent observation per participant.

# Appendix C An Additional Experiment on *ex-post* rationalization

In an additional experiment, we eliminated the potential for *ex-post* rationalization in the opponent frame by asking participants about their beliefs (directly) *before* they make their choice in the discoordination games from Experiment 1-DISC (both players obtain 7€ iff they choose different options).[20] Comparing the own-action probabilities from this treatment to the corresponding probabilities from Experiment 1-DISC yields an estimate for the importance of *ex-post* rationalization. We can interpret the probability difference in this way because we already know from Experiment 2-TYL that both the consensus effect and wishful thinking do not seem to play a role under the opponent frame. As an additional benchmark, we also ran two sessions under the random-other frame. Under this frame, we expect there to be no difference between Experiment 1-DISC and the additional experiment (as stated above, we see little scope for *ex-post* rationalization in the random-other frame). 86 subjects participated in the additional experiment.

[20] *Ex-post* rationalization of a belief by an action would be unintuitive: we may well choose an action without forming a belief in the standard setup, but once we form a belief (as in the first stages of the additional experiment), there does not seem to be a good reason to form yet a different belief that we then contradict out of a taste for consistency.

**Results** The results in Figure C1 show that removing the potential for *ex-post* rationalization indeed changes the own-action probabilities in participants' reported beliefs: under the opponent frame—the frame under which we would expect *ex-post* rationalization—average own-action probabilities are roughly four percentage points (or 25%) higher when beliefs are elicited before actions compared to when they are elicited after the action (rank-sum test, $p = 0.028$). In contrast, under the random-other frame (where we argued *ex-post* rationalization should play no role) there is no difference ($p = 0.742$), which is in line with the results of Rubinstein & Salant (2016). We interpret the results as additional evidence for *ex-post* rationalization in the opponent frame.

# Appendix D Experimental Instructions

The instructions are translated from German and show the opponent frame as an example. Boxes indicate consecutive screens shown to participants. The instructions of the additional experiment in Online Appendix C had the same content, but were slightly more complicated due to the belief elicitation before the action.

**Today's Experiment**
Today's experiment consists of 24 situations in which you will make two decisions each.

**Decision 1 and Decision 2**

In the first situation, you will see the instructions for both decisions directly before the decision. In later situations, you can display the instructions again if you need to.

**The payment of the experiment**
In every decision you can earn points. At the end of the experiment, 2 situations are randomly drawn and payed. In one of the situations, we pay the point you earned from decision 1 and in the other situation, you earn the points from decision 2. The total amount of points you earned will be converted to EURO with the following exchange rate:

**1 Point = 1 Euro**

After the experiment is completed, there will be a short questionnaire. For completion of the questionnaire, you additionally receive 7 Euro. You will receive your payment at the end of the experiment in cash and privacy. No other participant will know how much money you earned.

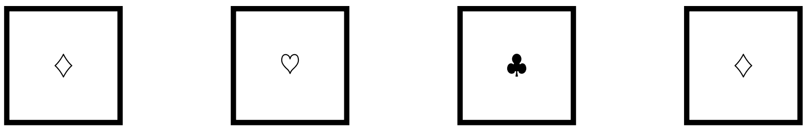
**Instructions for decision 1**
In today's experiment, you will interact with other participants. **You will be randomly rematched with a new participant of today's experiment in every situation.**
Decision 1 works in the following way: You and your matching partner see the exact same screen. On the screen, you can see an arrangement of four boxes which are marked with symbols. You and the other participant choose one of the boxes, without knowing the decision of the respective other. [One of] You can earn an price of X Euro.

*Experiment 1-DISC & 3*
[You only receive the X euro only if you choose **another** box than your matching partner. If both of you choose the same box, both do not receive points in this decision]

*Experiment 2*
[The relative position of your chosen boxes determines who wins the price. The participant wins, whose box lies to the immediate left of the other participant's box. If one participant chooses the most left box, then the other participant wins, if he chooses the most right box. If you don't win, you receive a price of 0 euro. It is of course possible, that neither you, nor the other participant wins.]

You will only learn at the end of the experiment, which box was chosen by the other participant and which payoff you receive in a certain situation.
The arrangement of symbols on the boxes is different in every situation. Below, you can see an example of how such an arrangement could look like.

**Example:** The four boxes are marked from left to right by Diamond, Heart, Spade, Diamond.



In this example, there are two boxes which are marked with the same symbol. However, the boxes on the most left and most right count as are different boxes.

*Only Experiment 2*

**Instructions for decision 1**
Although you choose a box in every situation, in some situations a box which was randomly chosen by the computer will be payoff relevant for you. This works in the following way:
After your decision, the computer draws one ball from the following urn in each situation:


If the blue ball that says "You" is drawn your own choice in decision 1 is relevant in this situation.
If the green ball that says "Computer" is drawn, the computer chooses one of the four boxes randomly (with equal probability of $\frac{1}{4}$) for you. This box will then be payoff relevant for you.
Your own decision is hence relevant with probability $\frac{1}{2}$ (=50%). The decision of the computer is relevant with probability $\frac{1}{2}$ (=50%).

**The decision of your matching partner**

**To determine whether you won the price, we always use the original decision of your matching partner. This also holds if the computer decides for you or the other participant.**
**To determine whether you won the price, we hence always use the original choice of your matching partner and, depending on the drawn ball, your decision or the decision by the computer.**

*Text in squared brackets is frame dependent. We show the opponent frame as example.*

**Instructions for decision 2**
In decision 2, your payoff also depends on your own decision and [on the decision of your matching partner. It will be the same matching parter, you already interacted with in decision 1.] We now explain decision 2 in detail.

**Decision 2**
Decision 2 refers always to a situation in which you already made decision 1. You will hence see the arrangement of boxes from the respective situation again. Again, the decision 1 [of your matching partner is relevant for you.] Decision 2 is about your assessment, [how your matching partner decided. We are interests in your assessment of the following question:]

[*See description of frames above*]

*Only Experiment 2*
[Please note that decision 2 is about the **actual** (human) decision of your matching partner and **not** about a possible computer decision.]

For every box, you can report your assessment [with what probability your matching partner chose the respective box]. You can enter the percentage numbers in a bar diagram. By clicking into the diagram, you can adjust the height of the bars. You can adjust as many times as you like, until you confirm. Since your assessments are percentage numbers, the bars have to add up to 100%. The sum of your assessment is displayed on the right. You can adjust this value to 100% by clicking. Or you enter the relative sizes of your assessments only roughly and then press the “scale” button. Please note, that because of rounding, the displayed sum ma deviate from 100% in some cases.
**On the next page, we explain the payoff of decision 2.**

*Text in squared brackets is frame dependent. We show the opponent frame as example.*

**The payoff in decision 2**
In this decision, you can either earn 0 or 7 points. Your chance of earning 7 points increases with the precision of your assessment. Your assessment is more precise, the more it is in line with [the decision behaviour of your matching partner. For example, if you reported a high assessment on the actually selected box, your chance increases. If your assessment on the selected box was low, your chance decreases.]
You may now look at a detailed explanation of the computation of your payment, which rewards the precision of your assessment.

**It is important for you to know, that the chance of receiving a high payoff is maximal in expectation, if you assess the behaviour of your matching partner correctly. It is our intention, that you have an incentive to think carefully about the behaviour of your matching partner. We want, that you are rewarded if you have assessed the behaviour well and made a respective report.**

Your chance will be computed by the computer-program and displayed to you later. At the end of the experiment, one participant of today's experiment will roll a number between 1 and 100 with dies. If the rolled number is smaller or equal to your chance, you receive 7 points. If the number is larger than your chance, you receive 0 points.

*Text in squared brackets is frame dependent. We show the opponent frame as example.*

**Payment of the assessments**
At the end of your assessment, you will receive the 7 points with a certain chance $(p)$ and with $(1-p)$, you receive 3 points. You can influence your chance $p$ with your assessment in the following way:

As described above, you will report an assessment for each box, on how likely [your matching partner is to select that box. One of boxes is the actually selected. At the end, your assessments are compared to the actual decision of your matching partner.] Your deviation is computed in percent.

Your chance $p$ is initially set to 1 (hence 100%). However, there will be deductions, if your assessments are wrong. The deductions in percent are first squared and then divided by two.

For example, if you place 50% on a specific box, but [your matching partner selects another box,] your deviation is equal to 50%. Hence, we deduct $0.50 * 0.50 * \frac{1}{2} = 0.125$ ( 12.5%) from $p$.

[For the box, which is actually selected by your matching partner, it is bad if your assessment is far away from 100%. Again, your deviation from that is squared, halved and deducted. For example if you only place 60% probability on the actually selected box, we will deduct $0.40*0.40*\frac{1}{2} = 0.08$ (8%) from $p$.]

With this procedure, we compute your deviations and deductions for all boxes.
At the end, all deductions are summed up and the smaller the sum of squared deviations is, the better was your assessment. For those who are interested, we show the mathematical formula according to which we compute the quality of your assessment and hence your chance $p$ of receiving 7 points.

$$p = 1 - \frac{1}{2}\left[\sum_i (q_{box_i,estimate} - q_{box_i,true})^2\right]$$

The value of $p$ of your assessment will be computed and displayed to you at the end of the experiment. The higher $p$ is, the better your assessment was and the higher your chance to receive 7 points (instead of 0) in this part. At the end of the experiment, the computer will roll a random number between 0 and 100 with dies. If this number is smaller or equal to $p$, you receive 7 points. If the number is larger than $p$ you receive 0 points.

**Summary**
**In order to have a high chance to receive the large payment, it is your aim to achieve as few deductions from $p$ as possible. This works best, if you have an good assessment of the behaviour of participant B and report that assessment truthfully.**

# Appendix E Experimental Instructions for Experiment 2-BOS

The instructions are translated from German. Boxes indicate consecutive screens shown to participants.

## Overview

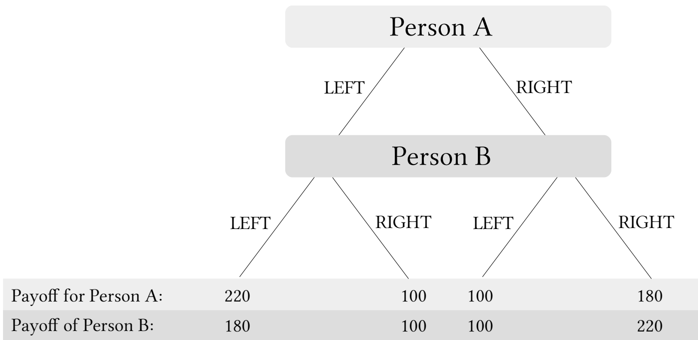
Welcome to this experiment.
For your participation in this experiment, you will be paid in cash at the end.
The payment amount will depend in parts on chance and in parts on your decisions as well as the decisions of others. Thus, it is important that you read the instructions carefully and understand them prior to the start of the experiment.
Today's experiment contains three parts. At the end, one of the parts will be randomly drawn and paid out.
Your payment amount results from the experimental currency units (ECU) from the randomly-drawn part, converted into Euros. All participants will be paid in private, so that other participants cannot see how much you earned.
The conversion rate from ECU to Euros is:

**1 Point = 1 Euro**

After the experiment is completed, there will be a short questionnaire. For completion of the questionnaire, you additionally receive 7 Euro.
You will shortly be shown the instructions for Part 1.

## Part 1: The underlying situation

We start by explaining the situation underlying Part 1, which is depicted in the figure shown below. There are two people involved, Person A and Person B. Both persons can choose between two options: LEFT or RIGHT. First, Person A makes his/her decision, then, Person B; however, Person B has to make his/her decision without knowing which choice Person A has made. If both persons choose LEFT, Person A receives 220 ECU and Person B 180 ECU. If both choose RIGHT, then Person B receives 220 ECU and Person A 180 ECU. If, in contrast, both choose different options, then both receive 100 ECU.




*Screen 1:*

## Overview of Part 1.

Part 1 contains 2 rounds. In round 1, you will be randomly assigned to a role, that of Person A or Person B. At the same time, you will be randomly assigned to another participants who thus will take on the opposite role. In round 2 you will then take on the other role and you will be randomly assigned to a new other participant. In case Part 1 turns out to be payoff relevant, you will be paid at the end for either round 1 or round 2.
In addition, we will be asking you in both rounds to provide an estimate of what [another participant/other participants] in this room will choose. If at the end, your decision in round 1 turns out to be payoff relevant, you will be paid in addition for the quality of your estimate in round 2. If, in contrast, your decision in round 2 turns out to be payoff relevant, you will be paid in addition for your estimate in round 1. In the following, we will explain how your payoff from the experiment is determined.

*Screen 2:*

## Payoff from the experiment

As you already know, the payoff-relevant part will be randomly drawn at the end of the experiment. In this, all three parts are equally likely and also will be the same for all participants of the experiment (i.e., if, e.g., Part 1 is paid out for you, then Part 1 will also be payoff-relevant for all other participants in the room). On the next page, you will find the description of how the payment for the quality of your estimate is calculated.

*Then, the information on the belief-elicitation incentives follows, analogous to Experiment 1-DISC.*