Contingent Belief Updating

Chiara Aina* Andrea Amelio† Katharina Brütt‡

September 24, 2024

Abstract

We study the impact of contingent thinking on belief updating. Engaging in contingent thinking calls for both processing hypothetical information and contrasting multiple contingencies during the belief-updating process. Our experimental findings show that contingent thinking leads to significant deviations from Bayesian updating. These deviations arise from the diminished perceived informativeness of hypothetical signals and the challenges posed by asymmetric signals, where comparing contingencies becomes more difficult. These results have implications for contingent planning, information acquisition, and information design.

Keywords: Belief Updating; Contingent Thinking; Experiment.
JEL Classification: C91; D83; D91.

*Universitat Pompeu Fabra and Barcelona School of Economics. E-mail: chiara.aina@upf.edu.
†Bocconi University. E-mail: andrea.amelio@unibocconi.it.
‡School of Business & Economics, Vrije Universiteit Amsterdam. E-mail: k.bruett@vu.nl.
We are grateful to Katie Coffman, John Conlon, Benjamin Enke, Christine Exley, Shengwu Li, Nick Netzer, Arthur Schram, Josh Schwartzstein, Joep Sonnemans, Jakub Steiner, Matthew Rabin, Chris Roth, Roberto Weber, Florian Zimmermann, as well as the seminar and conference participants at Cornell University, NYUAD, BSE Summer Forum, ECBE, EAYE Annual Meeting, and the ESA conference. We also thank all those who took the time to participate in our expert survey. The preregistration plan is available at `https://aspredicted.org/D2G_X81`. The experiment was approved by the Ethics Committee Economics and Business (EBEC) at the University of Amsterdam with reference number 20220303100340. Funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2126/1-390838866 and the Amsterdam Center for Behavioural Change is acknowledged.

1 Introduction

The role of beliefs in many settings of economic relevance is indisputable. A substantial body of research in economics and psychology has documented systematic deviations from the Bayesian benchmark. In these studies, participants report their updated beliefs after having observed additional information. However, numerous situations require individuals to proactively anticipate how expectations will evolve in response to diverse contingencies. Think about writing a contract, pricing a structured financial asset, or deciding to run a survey to collect new data. These scenarios demand that we incorporate a not-yet-observed piece of information into our belief system. Do we process the same information in the same manner in these circumstances? While, according to the Bayesian benchmark, the revision of beliefs should not depend on whether individuals engage with data contingently, this study shows that it does.

This paper experimentally studies whether and to what extent contingent thinking affects belief updating. To sharpen our question, we focus on the following working definition of contingent thinking: ahead of the resolution of some uncertainty, one reasons through the mutually exclusive potential realizations of such uncertainty (contingencies), assessing one's reaction to each potential realization.[1,2] Economics examines several situations that require the use of contingent thinking in belief updating. To formulate a strategy in sequential games, players often have to form beliefs contingently about the opponent's type or the state of nature. Also, decisions to acquire information, to experiment, or design information structures call for an ex-ante assessment of the value of information for different contingencies.

As an illustration, consider a doctor deciding whether to administer a test to a patient. The test produces an informative but noisy signal from which the doctor can learn about the patient's health. To make the decision, the doctor needs to anticipate how they would learn given each result, thereby engaging in contingent thinking. To do so, the doctor has to reason through both scenarios of a positive and a negative test result and update their beliefs for each contingency without having observed either. This is what we refer to as *contingent belief updating*, that is, assessing updated beliefs for all the possible signal realizations that could materialize. We distinguish this from what we call *conditional belief updating*: One observes a new piece of information and then assesses the updated beliefs only for that realized and relevant signal. Are beliefs assessed contingently the

[1] We assume contingencies to be known and foreseeable, ruling out concerns related to unawareness. While we believe this to be an important and interesting strand of literature (e.g., Schipper, 2022; Becker et al., 2020; Karni and Vierø, 2013, 2017, among many others), it is beyond the scope of this paper.

[2] A related but distinct concept to contingent thinking is *counterfactual thinking*. Following the prevalent definition in the psychology literature (see Kahneman and Tversky, 1982; Epstude and Roese, 2008; Byrne, 2016), counterfactual thinking refers to mental simulations of past events. Hence, the distinction between the two concepts lies in the object of the simulation, which concerns alternative versions of a realized event (counterfactual) as opposed to a potential future event (contingency). In some existing prominent works, this conceptualization seems to be less clear (e.g., Hoch, 1985); however, recent works in psychology embrace a clear-cut distinction between the two concepts (Pearl, 2009; Ferrante et al., 2012; Gerstenberg, 2022).

same as beliefs assessed conditionally? If not, would contingent belief updating help you form more accurate beliefs, or would this only lead to more biased beliefs?

Understanding the impact of contingent thinking on belief accuracy is important for three reasons. First, it provides an opportunity to deepen our understanding of the underlying factors contributing to why we observe biased beliefs. While much is known about biases in belief updating when individuals observe relevant and realized information, we cannot take for granted that the biases observed in this setting equally apply when reasoning through contingencies. This study offers valuable insights into the robustness of belief distortions in these settings. Second, this research question is economically relevant. On the one hand, if contingent thinking leads to less accurate beliefs, this is crucial in settings in which agents engage in contingent planning or information acquisition, such as in negotiating contracts or evaluating insurance plans. In the doctor's example, if beliefs updated conditionally differ from the ones assessed contingently, the ex-ante evaluation of the test might be misleading, resulting in either under- or over-testing. On the other hand, if contingent thinking proves effective in debiasing inaccurate beliefs, we would have an easily implementable, cheap, and portable debiasing mechanism to correct beliefs. This bears relevance across various domains to prevent over- or under-reactions to new information, ultimately improving economic outcomes. If this were the case in our previous example, the doctor would be better off sticking to how they evaluate the test results contingently rather than revising their beliefs upon observing the actual test result. Last, addressing this question is methodologically important. Framing the question differently, we investigate whether there is a systematic difference across beliefs elicited with the direct or strategy method. If this were the case, studies employing these methods should account for it in both the design and inference stages, ensuring accurate reporting and interpretation of the results.

We conduct an online experiment to investigate the effect of contingent belief updating. The experiment implements three between-subject treatments in the commonly used "balls-and-urns" updating exercise with binary state and signal. To investigate the underlying mechanisms, we employ two approaches. First, we identify two features of contingent belief updating that set it apart from conditional belief updating: (1) the hypothetical nature of the considered contingency (*hypothetical thinking*),[3] and (2) the consideration of all possible contingencies (*contrast reasoning*). Our treatments break down the effect of contingent thinking into these two components. The participants face contingent belief updating by employing the strategy method to elicit beliefs, while conditional belief updating can be induced by eliciting beliefs with the direct method. Both components of contingent thinking are present in the first, but absent in the second. Therefore, we introduce a third treatment that requires hypothetical thinking but

[3]There is a recent strand of literature in economics that focuses on the role of mental imagery, that is, *"representation that results from perceptual processing that is not triggered directly by sensory input"* (Stanford Encyclopedia of Philosophy). Dube et al. (2023), Ashraf et al. (2022), John and Orkin (2022), and Alan and Ertac (2018) show that mental imagery of future outcomes can lead to improvement in a wide range of economically relevant outcomes.

not contrast reasoning by eliciting posteriors conditional on one (random) hypothetical contingency. Second, we examine how the characteristics of the information structure and individual traits interact with the effect of contingent thinking. Participants face ten different signal-generating processes with different characteristics that could affect their updating, such as how diagnostic signals are (signal strength) and whether the different signals are equally diagnostic for different states (symmetric vs. asymmetric signal-generating processes). We measure the participant's capacity for cognitive reflection and their cognitive uncertainty.

The importance of studying the impact of contingent thinking on belief updating is emphasized by the fact that it is non-trivial even for experts to predict its directional effect. We asked a sample of academic experts in economics to predict how contingent belief updating affects belief distortions. We document significant heterogeneity in experts' expectations, with the majority believing biases would be unaffected or reduced if individuals update their beliefs contingently compared to conditionally. Our findings directly oppose the predictions of the experts we surveyed.

Overall, contingent thinking leads to more distortion in belief updating: compared to the Bayesian benchmark, we report both more biased beliefs in terms of the absolute distance and more underinference if beliefs are elicited contingently compared to conditionally. Contingent belief updating increases the absolute bias by one-third. In the doctor's example, this finding would suggest under-testing by an uninformed doctor. This effect seems to be entirely driven by hypothetical thinking rather than contrast reasoning. Indeed, the most striking insight emerging from our data is the harmful effect of hypothetical thinking. It leads to an increase of more than 50% in absolute bias and pushes participants to systematically underinfer more. We report how hypothetical thinking worsens a wide range of accuracy and consistency measures: not only are beliefs further from being Bayesian, but also, there is more noise in the reported beliefs and less consistency in how beliefs are updated across contingencies. The biasing effect of hypothetical thinking is more pronounced with stronger signals, and it also makes the task appear more challenging for participants.

Contrast reasoning compensates for the biasing effect of hypothetical thinking depending on the characteristics of the signal-generating processes. In particular, we report heterogeneous treatment effects by the symmetry of the signal-generating process. Our data show that contrast reasoning fully offsets the negative impact of hypothetical thinking when the signal-generating process is symmetric but has no effect when asymmetric. We further explain this sharp result by showing that the effectiveness of contrast reasoning decreases in a continuous measure of the symmetry of the signal-generating process. We can trace back the resulting increase in the bias to beliefs responding to asymmetric signals as if they were symmetric. As a consequence, contingent and conditional belief updating do not differ for symmetric signal-generating processes. In the example, the doctor's assessment of how their beliefs will evolve once exposed to the test's potential

outcomes is accurate if the false positive and false negative rates coincide. Finally, we find that individual measures of cognitive reflection and cognitive uncertainty do not mediate the ability to engage in either hypothetical thinking or contrast reasoning in this belief-updating task.

Our project speaks to several strands of literature. First, we contribute to the literature on biases in beliefs. There is ample evidence that, in particular, individuals underinfer from signals (Benjamin, 2019). The recent papers by Augenblick et al. (2021) and Ba et al. (2022) replicate this result, studying belief updating for several levels of signal diagnosticity, but also find that with weak signals, there is overinference. We purposefully exclude weak signals from our design to restrict our attention to underinference, allowing for a stronger identification of the effect. However, inspired by these studies, we employ several signal-generating processes that vary in signal strength to study how contingent belief updating is affected. Particularly close to our work, Gonçalves et al. (2024) report how people underinfer from retractions due to the increased complexity of retractions, which also require contingent reasoning. In contrast, we offer a direct test of the role of contingent reasoning in belief updating.

Second, there is a growing and recent body of literature in economics related to contingent thinking. These studies highlight the widespread challenges associated with contingent thinking (e.g., Li, 2017; Martínez-Marquina et al., 2019; Esponda and Vespa, 2014, 2023; Ngangoué and Weizsäcker, 2021). Our approach complements the existing literature on contingent thinking, recently surveyed by Niederle and Vespa (2023), as it differs in three key aspects from the most prominent papers. First, our focus lies on belief updating — processing of new information to report revised beliefs — rather than choosing an action — evaluating and comparing the implications of each alternative to implement the preferred one. Second, in these papers, agents are normatively expected to engage in contingent reasoning to solve the task at hand optimally. Instead, processing new information to update beliefs does not require thinking contingently.[4] Third, our approach involves participants reporting multiple contingency-specific guesses, either in the case where one contingency is observed (ex-post) or in the case there is uncertainty on the relevant realized contingency (ex-ante). In contrast, previous works focus on ex-ante decision-making, where contingent reasoning is instrumental in properly comparing the different contingency-specific consequences to choose the best course of action. Regardless of these differences, our paper and this literature document ways in which contingent thinking could impede payoff maximization, primarily rooted in the difficulty of considering hypothetical realizations, as discussed further in Section 5.

Last, this paper also contributes to the literature on elicitation methods. While most studies investigating biased beliefs employ the direct method to elicit beliefs, some few others adopt the strategy method (e.g., Esponda et al., 2020; Cipriani and Guarino, 2009;

[4]Moreover, in most of this literature, there is a (more) "relevant" contingency that participants may fail to pin down, leading to suboptimal behavior. In our study, all contingencies are relevant.

Toussaert, 2017; Agranov et al., 2020; Charness et al., 2021b; Ambuehl and Li, 2018).[5] Therefore, it becomes crucial to understand how to compare the results across methods of belief elicitation. The predominant focus of the literature on belief elicitation has been on the impact of payment schemes, rule complexity, and correspondence with actions (e.g., Charness et al., 2021a; Schlag et al., 2015; Schotter and Trevino, 2014). Despite a substantial body of research on the difference between direct and strategy methods for eliciting desired actions (for example, see Brandts and Charness, 2003; Brosig et al., 2003; Casari and Cason, 2009; Aina et al., 2020; and Brandts and Charness, 2011 for a review), to the best of our knowledge, our study is the first to demonstrate the that the method of belief elicitation will impact reported posteriors.

The rest of the paper is organized as follows: Section 2 describes our experimental design and data collection, Section 3 briefly describes the experts' predictions, Section 4 presents the results, and Section 5 discusses our findings.

2 Experimental Design

Studying how contingent thinking affects belief updating and the underlying mechanisms requires (i) a setting that prompts contingent thinking in belief updating, (ii) a treatment variation that disentangles the effects of hypothetical thinking and contrast reasoning, and (iii) a clean manipulation of characteristics of the signal-generating process.

To study belief updating, we employ the classic "balls-and-urns" updating exercise with a binary state and signal. The participants face two bags, A and B, which are equally likely to be selected, $\Pr(A) = \Pr(B) = 50\%$. Each bag has a total of either 80 or 60 balls.[6] Balls can be either blue or orange, and the participants know the distribution of the ball colors in the two bags. While the participants do not know which bag is randomly selected, the computer draws from the selected bag a ball whose color can be informative. The participant's task is to guess the probability of each bag being selected given the available information.[7]

2.1 Treatments

To manipulate whether participants engage in hypothetical thinking and contrast reasoning, the treatments change the method of belief elicitation. The treatments vary whether

[5]Kozakiewicz (2022) and Lilley and Wheaton (2024) use hypothetical signals to identify the effect of motivated reasoning on belief updating, while our research shows that there is a large difference in beliefs elicited for hypothetical signals and realized ones, even in a neutral setting.

[6]We do not use bags with a total of 100 balls to avoid that the heuristic answer (i.e., the probability of bag A after observing a blue ball is the number of blue balls in bag A) corresponds to the correct answer for the symmetric SGPs.

[7]We employ a version of this task in which participants are in control of each step: first, by clicking on 'Select the bag,' a bag is selected with a virtual coin flip; then, by clicking on 'Draw the ball,' a ball is drawn from the selected bag. We use graphical animations for the coin flip and the ball draw to create a realistic setting online and remind the participants of the basic structure of the task in each round.

Table 1: Treatments

		Contrast Reasoning	
		No	Yes
Hypothetical Thinking	No	Conditional	—
	Yes	One-Contingency	All-Contingency

the signal for which beliefs are reported has been observed (signal realization observed *vs.* hypothetical) and how many contingencies are considered (one *vs.* both signal realizations), as shown in Table 1. The three between-subject treatments, as detailed below, are summarized in Figure 1 and the corresponding choice interface is shown in Figure 2 (see Appendix C.2 for more details on the interfaces).

1. **Conditional**: The beliefs are elicited conditional on the realized signal. The participant observes the color of the drawn ball and is then asked to assess beliefs only conditional on that relevant contingency. This corresponds to the classic balls-and-urns task and what we refer to as *conditional belief updating*. It also corresponds to eliciting beliefs with the direct method.

2. **All-Contingency**: The beliefs are elicited conditional on both possible signal realizations. Before observing the color of the drawn ball, the participant is asked to report beliefs conditional on both cases on the same screen, in a randomized order: (1) the computer draws an orange ball, and (2) the computer draws a blue ball. Thus, participants consider two hypothetical contingencies with the possibility of comparing their beliefs conditional on one signal realization to their beliefs conditional on the other signal realization. After the beliefs are reported, the participants learn the color of the drawn ball. We refer to this as *contingent belief updating*. This features both *hypothetical thinking* and *contrast reasoning*. This treatment corresponds to a belief elicitation that employs the strategy method (as introduced in Mitzkewitz and Nagel, 1993).

3. **One-Contingency**: The beliefs are elicited conditional on only one possible signal realization. When participants have not yet observed the signal realization, they are asked to consider one of the following hypothetical cases: (1) the computer draws an orange ball, or (2) the computer draws a blue ball. Each case is chosen with equal probability in each round. As in *All-Contingency*, participants learn the color of the drawn ball after the belief elicitation. This treatment, therefore, requires to

engage in *hypothetical thinking*, but not *contrast reasoning*.[8]

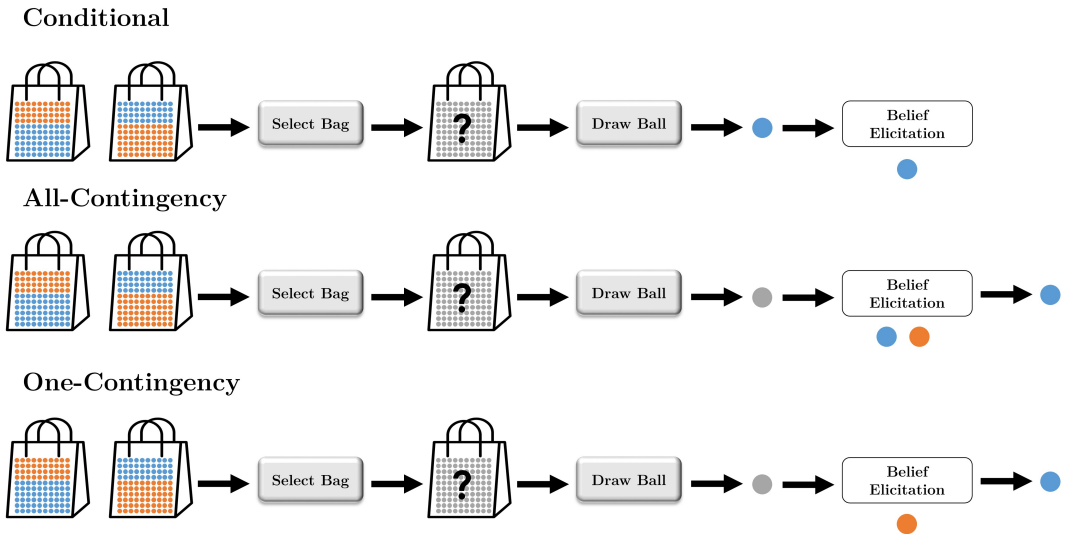


Figure 1: Task & Treatments

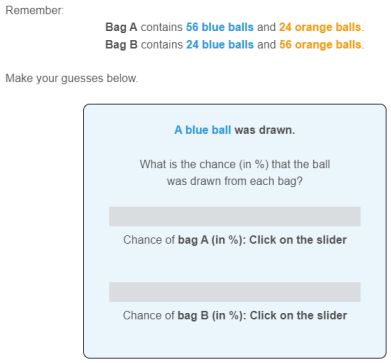
Notes. The figure illustrates the task timeline for each treatment. Treatments branch out after a ball is drawn from the selected bag. In *Conditional*, participants observe an animated colored ball being drawn, while in the other two treatments, the ball is uncolored with a question mark, indicating that its color remains unknown at this stage. Belief elicitation varies across treatments. In *Conditional*, participants are asked about their posterior given the observed drawn ball. In *All-Contingency*, participants are asked about their posteriors for both possible signal realizations. In *One-contingency*, participants are asked about their posterior for one of the possible signal realizations. After the belief elicitation stage, participants learn the color of the previously drawn ball in *All-Contingency* and *One-Contingency*.

2.2 Signal-Generating Processes

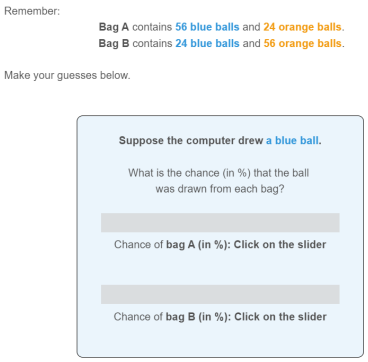
The task is repeated for ten rounds. In each round, participants face a different signal-generating process (hereafter, SGP). Figure 3 summarizes and illustrates the 10 SGPs used in this experiment in terms of their characteristics and induced Bayesian posteriors for both signals. In what follows, we refer to each SGP with the respective "$\Pr(blue|A) - \Pr(blue|B)$" as in Figure 3a. Each SGP specifies how diagnostic a ball of a specific color ball is for each bag. We measure the *signal strength of signal s for bag A* as

$$\lambda_s = \frac{\Pr(s|A)}{\Pr(s|B)}.$$

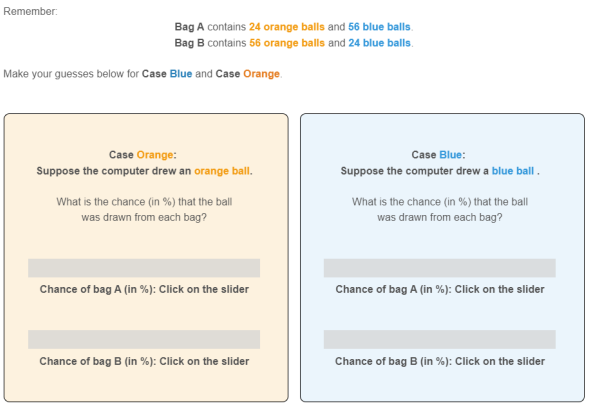
[8] Although other treatments could have been designed to disentangle the effect of hypothetical thinking and contrast reasoning, we found this version to be the cleanest for our purpose. For example, beliefs could have been elicited sequentially for each hypothetical signal realization. We discard this option as it might have triggered contrast reasoning over rounds. Alternatively, beliefs could have been elicited conditional on the observed signal realization for two identical but independent tasks on the same screen. This approach would trigger contrast reasoning whenever different signals were observed for the two independent tasks. We choose to avoid this treatment because participants may not easily understand the independence assumption.



(a) Treatment *Conditional*



(b) Treatment *One-Contingency*



(c) Treatment *All-Contingency*

Figure 2: Decision Interface by Treatment

Notes. The figure displays screenshots of decision interfaces for each treatment. Panel (a) presents the interface for the treatment *Conditional*, in the case where participants are asked to make a guess upon observing the drawing of a blue ball. Panel (b) presents the interface for the treatment *One-Contingency*, in the case where participants are asked to make a guess for the contingency in which the drawn ball was blue. Panel(c) presents the interface for the treatment *All-Contingency*.

If $\lambda_s = 1$, the signal is not diagnostic for either bag; however, if $\lambda_s > 1$ ($\lambda_s < 1$), the signal is more diagnostic for bag A (B) and λ_s measures by how much.[9] To compare signals more easily across SGPs, we consider the signal strength of each signal in terms of the bag for which the signal is more diagnostic: $\bar{\lambda}_s = \lambda_s$ if $\lambda_s \geq 1$ and $\bar{\lambda}_s = \lambda_s^{-1}$ otherwise. Varying signal strength within-subject over rounds allows us to investigate the mechanism along this dimension and the robustness of the effect of contingent thinking on belief updating.

We include both symmetric and asymmetric SGPs. An SGP is *symmetric* if the probability of drawing a blue ball from bag A is the same as the probability of drawing an orange ball from bag B. This implies that, with a symmetric SGP, examining only one bag suffices to obtain all the relevant information to determine the signal strength and, thus, to guess the posterior correctly. Moreover, for a symmetric SGP, the signal strengths of a blue ball and of an orange ball coincide, i.e., $\bar{\lambda}_{blue} = \bar{\lambda}_{orange}$. This simple relationship between signal strengths might facilitate contrast reasoning, leading to a heterogeneous effect of contingent thinking for symmetric and asymmetric SGPs.

To explore this further, alongside the binary classification, we introduce a continuous measure that quantifies the degree of asymmetry in an SGP by comparing how diagnostic signals are against each other. This measure captures the extent to which signals vary in their signal strength relative to their average signal strength:

$$\sigma\left(\bar{\lambda}_s, \bar{\lambda}_{s'}\right) = \frac{|\bar{\lambda}_s - \bar{\lambda}_{s'}|}{\bar{\lambda}_s + \bar{\lambda}_{s'}}.$$

For all symmetric SGPs, the ratio is zero, but positive for asymmetric SGPs. Moreover, the higher the ratio, the more asymmetric the signals are in their diagnosticity level.[10,11]

Lastly, some SGPs are *mirrored*, meaning that participants are exposed to the same SGP twice, inverting the distributions of balls in the bags and changing the number of balls in the bag. Throughout the experiment, we vary whether the total number of balls in the bags is 80 or 60. The mirrored SGPs are presented once with bags with 80 balls and once with 60 balls. We mirrored one symmetric SGP (15-85 and 85-15) and one asymmetric SGP (30-95 and 95-30). The reason why we include mirrored SGPs is twofold. First, we use them to check the consistency of reported beliefs given the same signal across rounds. This is a measure of how stable the deviations from Bayesian updating are within-task (*within-consistency*). Second, this allows us to better compare *Conditional* and *One-Contingency* to *All-Contingency*. When beliefs are elicited contingently, participants

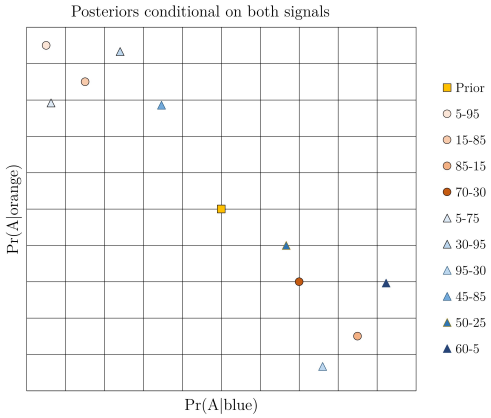
[9]Note that the Bayesian posterior for equal priors can be calculated as $\Pr(A|s) = \left(1 + \lambda_s^{-1}\right)^{-1}$.

[10]In perception studies (for a review in economics, see Woodford, 2020), the contrast between stimuli is quantified analogously to highlight the importance of relative rather than absolute differences in perceptual experience. In economics, salience theory (for a review, see Bordalo et al., 2022b) uses this functional form to measure the salience of certain attributes in decision-making, as it is a tractable version and satisfies the key properties of ordering and diminishing sensitivity.

[11]As there is no canonical way to measure the degree of asymmetry for an SGP, we consider this measure because it is simple and intuitive given our setting. In Appendix A.3.1, we replicate related findings using a conceptually similar but different measure: the distance between posteriors across signals.

Name	Pr(blue|A)	Pr(blue|B)	Symmetric	Mirrored	Total Balls
5-95	5%	95%	Yes	No	60
15-85	15%	85%	Yes	Yes	80
85-15	85%	15%	Yes	Yes	60
70-30	70%	30%	Yes	No	80
5-75	5%	75%	No	No	60
30-95	30%	95%	No	Yes	60
95-30	95%	30%	No	Yes	80
45-85	45%	85%	No	No	60
50-25	50%	35%	No	No	80
60-5	60%	5%	No	No	80

(a) Characteristics



(b) Bayesian Posteriors

Figure 3: Signal Generating Processes

Notes. Panel (a) summarizes the different characteristics of the SGP. Panel (b) illustrates the induced posteriors of the different SGPs graphically. The name of the SGP refers to the corresponding "$\Pr(blue|A) - \Pr(blue|B)$".

report their conditional beliefs on both signal realizations, while they report their beliefs only conditional on one signal in *Conditional* and *One-Contingency*. This allows us to study whether posteriors across signal realizations are consistent with the Bayes rule between signal realizations (*between-consistency*).

For the last task, we also elicited cognitive uncertainty (Enke and Graeber, 2023). For comparability, the last choice displays the same SGP for all participants — that is, 70-30; we randomize the order of the SGP for the remaining 9 choices. We pick 70-30 for this because this SGP is closest to the most widely used SGP (67-33) in this type of experiment (see meta-analysis by Benjamin, 2019).

2.3 Incentives

The belief elicitation was incentivized using the binarized scoring rule (Hossain and Okui, 2013): the closer the reported beliefs are to the realized state, the higher the probability of receiving the bonus of GBP 2.[12] One of the ten tasks is randomly selected for payment.

We design our incentives to (a) ensure incentive-compatibility for truthful reporting and (b) keep the strength of the incentives comparable in all three between-subject treatments. To achieve the first, we ensure that each contingency occurs with non-trivial probabilities for all SGPs (50-50 for symmetric SGPs, and at most 70-30 for asymmetric SGPs). To keep the strength of incentives comparable to *Conditional*, the incentivization requires minimal adjustments in *One-Contingency* and *All-Contingency*. In *All-Contingency*, par-

[12]Danz et al. (2022) show that providing complex details on the elicitation procedure might confuse participants and distort their incentives. Therefore, instructions clarify that "to maximize the chance of winning the bonus, it is in your best interest always to give a guess that you think is the true chance." This is also checked in one of the control questions. If interested, participants could read further about the details of this elicitation rule.

ticipants' beliefs are elicited for both contingencies, and the realized contingency determines which guess is relevant for the payment. Paying both belief elicitation tasks could have affected the participants' attention because of the difference in the magnitude of incentives across treatments, confounding our results. In *One-Contingency*, incentives are the same as in *Conditional* if the randomly-proposed contingency corresponds to the realized one; otherwise, the elicited guess is irrelevant for determining the bonus, and the participant receives a fixed payment of GBP 1.[13]

2.4 Logistics

The experiment was pre-registered on AsPredicted.[14] It was conducted on Prolific in March 2023, restricting the participant pool to workers located in the UK.The participants received a link to a Qualtrics survey that includes the instructions, choice tasks, cognitive uncertainty elicitation for the last choice, and a final survey — eliciting Cognitive Reflection Test (Frederick, 2005), Berlin numeracy task, demographics, and a short questionnaire. The average payment was 3.37 GBP, with an average duration of approximately 24 minutes.The participants earned GBP 2 for completing the study and could earn an additional bonus of GBP 2 depending on their performance in the tasks.

Instructions were split into two blocks, each followed by a set of control questions. The first block was the same for all treatments: it welcomed the participants, provided general information on the experiment, and explained the balls-and-urns task in detail. The second block focuses on the treatment-specific choice procedure and payment, and thus it varies by treatment. See Appendix C for the instructions, including control questions. Only participants who pass these questions are included in the analysis, as preregistered. A total of 525 participants completed the study, of which 86% passed the control questions about the experiment instructions (not statistically different between treatments: 88% in *Conditional*, 86% in *All-Contingency*, and 83% in *One-Contingency*). This leaves valid observations from 150 participants per treatment. In our final sample of 450 participants, 50% are female, 36% have low schooling ('High school' or lower educational level), and the median age is 37.

3 Expert Survey

To contextualize our findings, we elicit predictions from a sample of academic experts in economics that we considered knowledgeable about topics related to expectations or contingent thinking, before collecting the data. Answers to this expert survey were collected through the Social Science Prediction Platform; we report details of the data collection, survey, and results in Appendix B.

[13]While some guesses being payoff-irrelevant could lead to lower attention, we chose this incentivization as consistency of incentives across treatments is our priority. Furthermore, evidence in similar tasks has shown that the strength of incentives does not impact the belief accuracy (Enke et al., 2023).

[14]The preregistration plan is available at `https://aspredicted.org/D2G_X81`.

Our survey focuses on the comparison of the treatments *All-Contingency* and *Conditional* and documents significant heterogeneity in experts' opinions on the effect of contingent belief updating. Of 38 responses, 37% expected more accurate beliefs when they are elicited contingently compared to when they are elicited conditionally, 61% did not expect any significant difference between the two elicitation methods, and only one expert expected the opposite. The majority of the experts also do not expect any heterogeneous effects based on the characteristics of the SGPs or individual traits. We take this expert survey as evidence that experts believe that beliefs are not less accurate when assessed contingently.

4 Results

In this section, we first introduce our two main outcomes of interest, *bias* and *underinference*, and we provide an overview of the main treatment effects. We continue with a discussion of potential mechanisms, considering both characteristics of the SGPs and of the participants as drivers of the treatment effects.

4.1 Key Outcomes

We investigate the effects of conditional and contingent belief updating on two measures, both capturing distinct aspects related to deviations from Bayesian updating.

First, the bias is defined as the absolute distance between the reported posterior and the Bayesian posterior for each task. We focus on the absolute distance to determine whether individual guesses are systematically more accurate across treatments. In contrast, a directional measure assesses whether the average guess is more or less accurate, allowing individual biases to cancel out. Second, we consider how participants respond to the signal strength to capture directional deviations from Bayesian updating. We use the following model introduced by Grether (1980) that defines the posterior-odds ratio given equal priors as:

$$\frac{\Pr(A|s)}{\Pr(B|s)} = \left[\frac{\Pr(s|A)}{\Pr(s|B)}\right]^{\alpha} = \lambda_s^{\alpha}.$$

Deviations from $\alpha = 1$ capture a participant's distortion in how their beliefs respond to the signal strength. While Bayes' theorem prescribes $\alpha = 1$, *underinference* corresponds to $\alpha < 1$: the reported posteriors conditional on a signal are as if the signal strength is perceived as less diagnostic for bag A and more diagnostic for bag B than what it actually is. Symmetrically, $\alpha > 1$ corresponds to *overinference*: The signal strength is treated as more diagnostic for bag A and less for bag B than it actually is. Unlike the bias, α is a directional measure of deviations from Bayesian updating defined across SGPs. [15]

Our experiment replicates the deviations from Bayesian updating reported in the liter-

[15] We replicate our results using the alternative measure of underinference introduced in Ba et al. (2022) in Appendix A.2.1.

ature, both in terms of bias and underinference. Comparing the available data in the online appendix of Benjamin (2019) to our results in *Conditional* for comparable SGPs (equal prior, symmetric SGPs, including SGPs 60-40, 67-33, and 83-17), we find that the average bias of 5.9 percentage points for the most similar SGP in *Conditional* (70-30) aligns closely with the average bias of 6.7 percentage points in previous comparable studies.

In his meta-analysis, Benjamin (2019) estimates

$$\log \frac{\Pr(A|s)}{\Pr(B|s)} = \alpha \log \lambda_s + \beta \tag{1}$$

and finds strong evidence for underinference with $\hat{\alpha} = 0.86$ for incentivized similar tasks (equal prior, one observed signal, symmetric SGP).[16] In line with this, the estimated coefficient in *Conditional* for symmetric SGPs is also exactly $\hat{\alpha} = 0.86$.

4.2 Treatment Effect

First, we consider our main treatment effects for the two outcomes of interests. Figure 4a reports the average bias by treatment. Column I of Table 2 displays the corresponding results for OLS regressions of the bias on indicators for the different treatments.[17] Figure 4b shows the plot of the log posterior-odds ratio against the log signal strength. The slope captures the estimated underinference estimated across SGPs.[18] In Column I of Table 3, we show the results of regressing the log posterior-odds over the log signal strength interacted with treatment indicators.

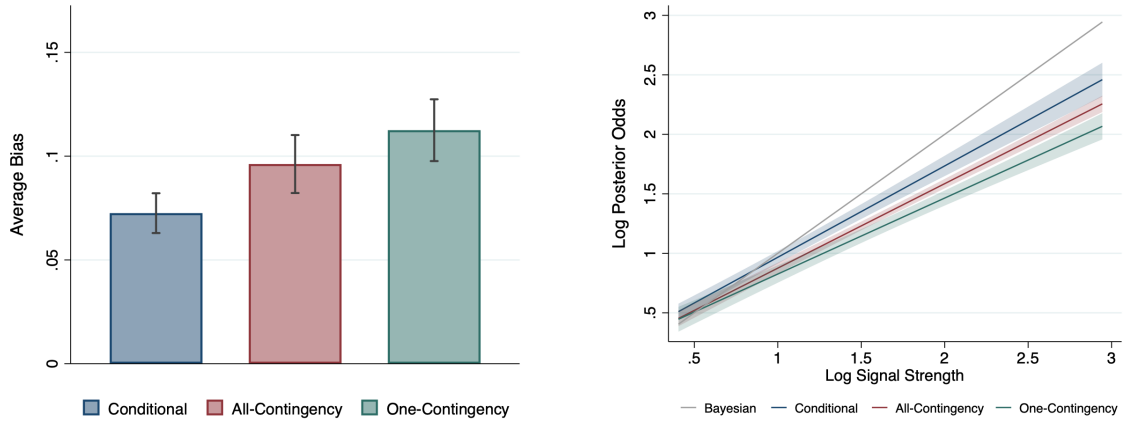
Finding 1. *Deviations from Bayesian updating are significantly larger if beliefs are updated contingently compared to conditionally.*

The treatment *All-Contingency* increases the bias compared to *Conditional*. The estimated baseline bias amounts to 7.2 percentage points in *Conditional*. We estimate that the average bias increases in *All-Contingency* by 2.4 percentage points, so by one-third,

[16] Augenblick et al. (2021) reports overinference from weak signals and underinference from strong signals for symmetric SGPs. Our symmetric SGPs are chosen such that their results would predict underinference. Also, there is some evidence of overinference for asymmetric SGPs and weak signals (for references see Benjamin, 2019). None of the signals in our asymmetric SGPs can be considered weak according to these standards. For these reasons, we do not carry out any analysis on the impact of signal strength on the level of underinference.

[17] Analyses in Table 2 include SGP fixed effects. Table A1 reproduces the analyses with SGP $\times$ contingency fixed effects. Contingency here refers to the observed signal realization in *Conditional* and to the contingency relevant for the belief elicitation in *One-Contingency* and *All-Contingency*. This does not change the interpretation of our results.

[18] Figure A2 shows the average bias by treatment for each SGP separately. See Figure A5 for an overview of the estimated degree of underinference by treatment and SGP. In all treatments, we observe underinference for most SGPs. For a more nuanced picture, Figure A6 provides an overview of the heterogeneity in the estimated degree of underinference by subject and treatment.



(a) Bias

(b) Underinference

Figure 4: Treatment Effect

Notes. Panel (a) shows the average bias defined as the absolute value of the difference between the posterior reported by participants and the normative (Bayesian) benchmark by treatment. Panel (b) shows the plot of the estimated relationship between the log posterior-odds ratio and the log signal strength by treatment as an illustration of underinference. Error bars in Panel (a) and shaded areas in Panel (b) indicate 95% confidence intervals, clustered at the individual level.

compared to *Conditional* ($p = 0.006$; Column I in Table 2).[19] We can, therefore, conclude that contingent belief updating increases the deviations from Bayesian updating.

We find directionally similar results in terms of underinference. Overall, there is strong evidence of underinference: the estimated coefficients $\hat{\alpha}$ are 0.76 in *Conditional* and 0.70 in *All-Contingency*, displaying significant deviations from the Bayesian benchmark of $\alpha = 1$ in both treatments ($p < 0.001$). This is reflected by the estimated log posterior-odds ratio being below the 45° line in Figure 4b. While the slope is visibly less steep in Figure 4b for *All-Contingency* than for *Conditional*, the estimated underinference in *All-Contingency* is not statistically different from the underinference in *Conditional* ($p = 0.243$; see the coefficient on 'Log Signal Strength × All-Contingency' in Column I of Table 3).

Next, we look at how the effect of contingent thinking can be explained by its decomposition into hypothetical thinking and contrast reasoning. Recall that comparing *One-Contingency* to *Conditional* allows us to identify the effect of purely hypothetical thinking without any opportunity for contrast reasoning. Instead, comparing *All-Contingency* to *Conditional* includes both hypothetical thinking and contrast reasoning.

Finding 2. *Hypothetical thinking is driving the biasing effect of contingent belief updating.*

Comparing the bias in *One-Contingency* to *Conditional*, we see a significant change of 4 percentage points ($p < 0.001$; Column I in Table 2), increasing the observed bias by more than 50%. The average bias in *All-Contingency* lies in between the bias in *Conditional*

[19]While the average bias is significantly different from zero in all treatments, 27% of the reported posteriors exhibit no bias. In particular, 25% in *Conditional*, 32% in *All-Contingency*, and 20% in *One-Contingency* of the reported guesses correspond to the correct Bayesian posterior. Figure A1 shows the cumulative distribution of this measure by treatments.

and in *One-Contingency*, even if the latter is not statistically different ($p = 0.118$; see the difference of the coefficients on 'All-Contingency' and 'One-Contingency' in Column I in Table 2). Therefore, we can attribute the entire increase in the bias induced by contingent thinking to hypothetical thinking.

Interestingly, treatment effects seem to be robust to learning throughout the experiment, as shown in Figure A3. We do not find evidence of learning over rounds in *Conditional* ($p = 0.650$). Instead, the average bias increases in each trial by 0.4 percentage points in *One-Contingency* ($p = 0.017$) and decreases by 0.3 percentage points in *All-Contingency* ($p = 0.021$). If anything, the treatment effect seems to strengthen throughout the experiment.

Turning to our second outcome measure, participants underinfer significantly more in *One-Contingency* ($\hat{\alpha} = 0.63$) than into *Conditional*, with $\hat{\alpha}$ decreasing by 12.9 percentage points ($p = 0.021$; see the coefficient on 'Log Signal Strength $\times$ One-Contingency' in Column I of Table 3). Hypothetical thinking thus pushes participants to systematically underinfer more. The level of underinference is also not statistically different between *One-Contingency* and *All-Contingency*, providing support for the argument that contrast reasoning does neither further increase nor decrease deviations from Bayesian updating.

Table 2: Bias

	I	II	III	IV
All-Contingency	0.024**	0.010	0.017*	0.011
	(0.009)	(0.011)	(0.008)	(0.009)
One-Contingency	0.040***	0.045***	0.014	0.039***
	(0.009)	(0.011)	(0.010)	(0.010)
Asymmetric		0.026*		
		(0.010)		
All-Contingency × Asymmetric		0.023*		
		(0.009)		
One-Contingency × Asymmetric		-0.008		
		(0.010)		
Log Signal Strength			0.013*	
			(0.005)	
All-Contingency × Log Signal Strength			0.003	
			(0.006)	
One-Contingency × Log Signal Strength			0.015*	
			(0.007)	
Degree of Asymmetry				0.027*
				(0.012)
All-Contingency × Degree of Asymmetry				0.041**
				(0.015)
One-Contingency × Degree of Asymmetry				0.003
				(0.018)
Constant	0.062***	0.068***	0.019	0.064***
	(0.008)	(0.009)	(0.013)	(0.006)
N	6000	6000	6000	6000
adj. R^2	0.024	0.026	0.028	0.019
Clusters	450	450	450	450

Notes. OLS estimates. Individual-level clustered standard errors. SGP fixed effects in columns I - III. The dependent variable is defined as the absolute value of the difference between the posterior reported by participants and the normative (Bayesian) benchmark; * p<.05, ** p<.01, *** p<.001.

Table 3: Underinference

	I	II	III
Log Signal Strength	0.768***	0.862***	0.874***
	(0.035)	(0.043)	(0.037)
All-Contingency	-0.034	0.034	-0.033
	(0.056)	(0.082)	(0.064)
One-Contingency	-0.012	0.118	0.005
	(0.071)	(0.116)	(0.082)
Log Signal Strength × All-Contingency	-0.058	-0.035	-0.007
	(0.046)	(0.057)	(0.050)
Log Signal Strength × One-Contingency	-0.129*	-0.184*	-0.137*
	(0.055)	(0.081)	(0.065)
Asymmetric		0.293***	
		(0.082)	
Log Signal Strength × Asymmetric		-0.127*	
		(0.056)	
All-Contingency × Asymmetric		-0.061	
		(0.101)	
One-Contingency × Asymmetric		-0.169	
		(0.143)	
Log Signal Strength × All-Contingency × Asymmetric		-0.071	
		(0.067)	
Log Signal Strength × One-Contingency × Asymmetric		0.071	
		(0.098)	
Degree of Asymmetry			0.849***
			(0.166)
Log Signal Strength × Degree of Asymmetry			-0.382***
			(0.103)
All-Contingency × Degree of Asymmetry			0.018
			(0.215)
One-Contingency × Degree of Asymmetry			-0.002
			(0.278)
Log Signal Strength × All-Contingency × Degree of Asymmetry			-0.151
			(0.128)
Log Signal Strength × One-Contingency × Degree of Asymmetry			0.016
			(0.171)
Constant	0.199***	-0.019	-0.044
	(0.044)	(0.066)	(0.051)
N	6000	6000	6000
adj. R^2	0.249	0.252	0.258
Clusters	450	450	450

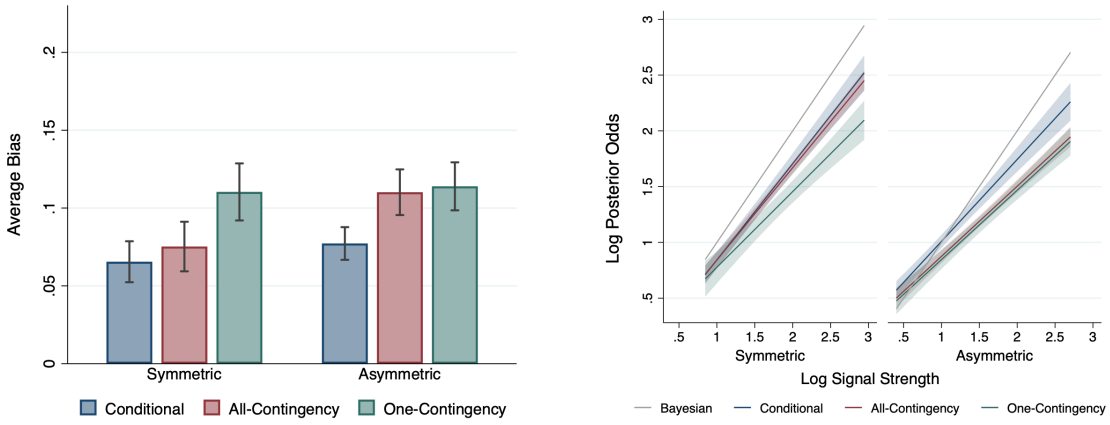
Notes. OLS estimates. Individual-level clustered standard errors. The dependent variable is defined as the logarithm of the ratio of the elicited posterior belief for each bag, for a given signal. The interactions of each treatment indicator and Log Signal Strength give the estimated underinference parameter α per treatment; * p<.05, ** p<.01, *** p<.001.

4.3 Mechanisms

We now further explore the mechanisms that drive the effect of contingent thinking on belief updating by highlighting first the role of the characteristics of the SGPs, then by looking closer at measures of consistency, and last by exploring the interaction with individual features.

4.3.1 Characteristics of SGPs

Looking at the features of the SGPs, we find that symmetry and signal strength differently impact hypothetical thinking and contrast reasoning.



(a) Bias by SGP Symmetry

(b) Underinference by SGP Symmetry

Figure 5: Treatment Effect by SGP Symmetry

Notes. Panel (a) shows the average bias defined as the absolute value of the difference between the posterior reported by participants and the normative (Bayesian) benchmark by treatment split by the symmetry of the SGP. Panel (b) shows the estimated relationship between the log posterior-odds ratio and the log signal strength by treatment as an illustration of underinference split by the symmetry of the SGP. Error bars in Panel (a) and shaded areas in Panel (b) indicate 95% confidence intervals, clustered at the individual level.

Symmetric vs. Asymmetric We begin our analysis of the mechanisms by looking at the heterogeneity of our treatment effects using the binary measure of symmetry, as defined in Section 2.2. Figure 5a provides an overview of the average bias of the posterior beliefs depending on whether the SGP is symmetric or asymmetric. Column II of Table 2 reports the difference-in-difference analysis of regressing the average bias on treatment indicators, a dummy indicator of whether the SGP is symmetric, and their interactions. Over all treatments, the average bias increases by 3.5 percentage points if signals are asymmetric ($p < 0.001$; Column II in Table 2). Hence, asymmetric SGPs clearly increase the difficulty of Bayesian inference. We document substantial heterogeneity in the treatment effect depending on the symmetry of the SGP.

Finding 3. *The impact of hypothetical thinking does not vary with the symmetry of the SGP. Contrast reasoning entirely offsets the effect of hypothetical thinking for symmetric SGPs, but not for asymmetric SGPs.*

For symmetric SGPs, the average bias increases by 4.5 percentage points if the participants consider one hypothetical contingency instead of observing the realized signal ($p < 0.001$; Column II in Table 2). However, we do not observe a significant increase in the average bias if beliefs are updated contingently compared to conditionally ($p = 0.354$; Column II in Table 2). In fact, we estimate that the posterior beliefs in symmetric SGPs are 3.5 percentage points more accurate in *All-Contingency* than in *One-Contingency* ($p = 0.005$; see the difference of the coefficients 'All-Contingency' and 'One-Contingency' in Column II in Table 2). By breaking down the effect of contingent thinking into hypothetical thinking and contrast reasoning, we thus observe that only hypothetical thinking further biases beliefs, but the presence of contrast reasoning fully compensates for this biasing effect for symmetric SGPs.

In contrast, our results for asymmetric SGPs show that the average bias is both significantly higher in the treatment *One-Contingency* and in the treatment *All-Contingency* than in the treatment *Conditional*. Hypothetical thinking with or without contrast reasoning increases the bias respectively by 3.3 and 3.6 percentage points, respectively (both $p < 0.001$; see the sum of the coefficients of the treatment indicators and their interactions with 'Asymmetric' in Column II in Table 2), so by more than 40%. The biases in these two treatments are indistinguishable for asymmetric SGPs ($p = 0.727$).

Therefore, these findings suggest that contrast reasoning only produces a debiasing effect for symmetric SGPs while exhibiting no impact for asymmetric SGPs. Furthermore, this insight indicates that Finding 1 summarizes a more nuanced picture. Recall that participants complete the task for 10 SGPs, of which 4 are symmetric and 6 are asymmetric. Given the heterogeneous effect by the SGP symmetry, we can infer that our main result about the harmful effect of contingent thinking on Bayesian updating is primarily driven by asymmetric SGPs, wherein contrast reasoning proves ineffective in mitigating the higher bias.

We report similar results also in terms of underinference, as illustrated by the log posterior-odds ratio plotted against the log signal strength for symmetric and asymmetric SGPs in Figure 5b. Interacting the variables to estimate the degree of underinference in Column I of Table 3 with indicators of the SGP symmetry in Column II of Table 3, we observe that, while hypothetical thinking increases the degree of underinference also if the SGP is symmetric ($p = 0.024$; see the coefficient on 'Log Signal Strength × One-Contingency' in Column II of Table 3), hypothetical thinking in combination with contrast reasoning only does so marginally for asymmetric SGPs ($p = 0.084$; see the sum of the coefficients on 'Log Signal Strength × All-Contingency' and 'Log Signal Strength × All-Contingency × Asymmetric' in Column IV of Table 3). Therefore, contrast reasoning reduces the degree of underinference if the SGP is symmetric but fails to do so if it is asymmetric.

Signal Strength Signal strength has a documented moderating effect on deviations from Bayesian updating (Augenblick et al., 2021). In Column IV of Table 2, we present the results of regressing the bias on indicators of the treatment, the SGP signal strength,

and their interactions. In line with the literature, we document a larger bias for stronger signals ($p = 0.011$). However, there is treatment-dependent heterogeneity. In *One-Contingency*, this effect is significantly stronger than in *Conditional* ($p = 0.039$; see Column IV of Table 2), suggesting that signal strength is an important driver of hypothetical thinking. Contrast reasoning has no such effect ($p = 0.741$; see the sum of the coefficients on 'One-Contingency' and 'All-Contingency' in Column IV of Table 2).

Finding 4. *Hypothetical thinking has a stronger effect on deviations from Bayesian updating for stronger signals.*

In the previous paragraphs, we show how the SGP's characteristics, such as symmetry and signal strength, influence our treatment effects. Next, we further explore their connection in two complementary ways.

Relative Signal Strength For symmetric SGPs, signals share the same signal strength, but for asymmetric ones, it is possible to distinguish between a weaker and a stronger signal. This might affect belief updating and, to explore this, we examine the role of relative signal strength. Formally, signal s is stronger than signal s' if $\bar{\lambda}_s > \bar{\lambda}_{s'}$. Intuitively, the stronger signal moves the updated beliefs further away from the prior compared to the weaker signal.[20]

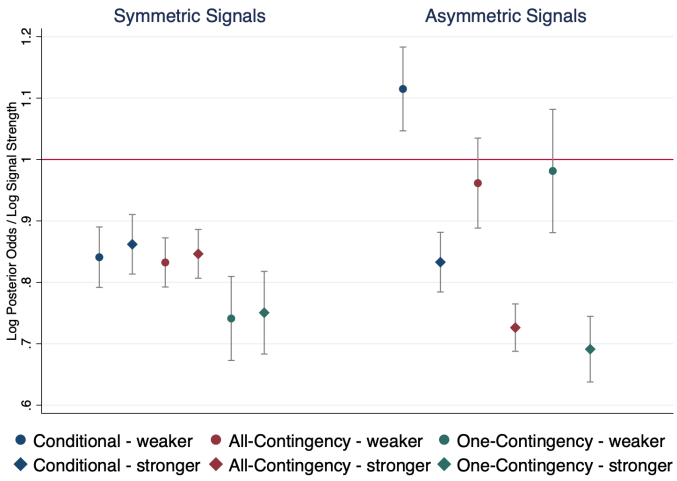


Figure 6: Underinference by Symmetry and Treatment for Stronger and Weaker Signals

Notes. Ratio of log posterior-odds and the log signal strength by treatment, symmetry of the SGP, and stronger vs. weaker signal; for symmetric SGPs, signals are equally diagnostic and we classified the blue ball as stronger arbitrarily. Error bars indicate 95% confidence intervals.

Figure 6 reports the average underinference levels by weaker vs. stronger signal, treatment, and SGP symmetry. For symmetric SGPs, signals are equally diagnostic and we should not observe any difference in the underinference level by signal within the same

[20] One can show that, for any asymmetric SGP, the stronger signal is the a priori less likely one, that is, $\bar{\lambda}_s > \bar{\lambda}_{s'}$ is equivalent to $\Pr(s) < \Pr(s')$. This allows us to offer another interpretation of these results based on the comparison between more likely or less likely signals.

treatment. Indeed, we do not observe that the participants react differently to the two signals. Turning to asymmetric SGPs, reported guesses exhibit a significantly higher degree of underinference for stronger signals. Across all treatments, we observe an underinference gap when comparing stronger and weaker signals: Relative signal strength matters for inference.[21] In fact, we find this underinference gap even at the individual level for *All-Contingency*. The average individual underinfers more when facing the stronger than when facing the weaker signal when reporting their guesses for the same SGP on the same screen (Figure A8).

We observe significant differences in the level of underinference across treatments. These differences follow the same pattern as in Finding 3, both for symmetric and asymmetric signals. Consider now only asymmetric SGPs, where relative signal strength is a meaningful measure. While underinference prevails for most treatments, for *Conditional*, weaker signals generate significant overinference. However, the underinference gap does not differ significantly across treatments. Table A4 reports the results of the regressions supporting these findings.

Degree of Asymmetry Finding 3 shows a sharp result: contrast reasoning fully offsets the biasing effect of hypothetical thinking for symmetric SGPs, but has no impact for asymmetric SGPs. To explore further the relation between contrast reasoning and symmetry, we turn to a more nuanced measure: degree of asymmetry. As formalized in Section 2, this quantifies the difference in signal strength between the two signals as a continuous measure of an SGP's asymmetry.

The degree of asymmetry captures when contrast reasoning is effective in reducing bias.[22] The coefficient of the interaction between the degree of asymmetry and *All-Contingency* is significantly positive ($p = 0.009$; Column IV of Table 2) and significantly larger than the coefficient of the interaction between the degree of asymmetry and *One-Contingency* ($p = 0.033$; Column IV of Table 2). Therefore, an increase in the degree of asymmetry decreases the effectiveness of contrast reasoning in counteracting the biasing effect of hypothetical thinking.

Next, we provide suggestive evidence regarding the possible mechanisms for these results. We consider three channels to explain why the higher degree of asymmetry may limit the ability of contrast reasoning to counteract the biasing effect of hypothetical thinking.

First, individuals may wrongly interpret an asymmetric SGP as if it were symmetric. For symmetric SGPs, the reported posteriors for the same bag across signals sum up to one because $\bar{\lambda}_{blue} = \bar{\lambda}_{orange}$ and, hence, $\Pr(A|b) = \Pr(B|o)$. This is not the case for

[21] This result builds and extends on the key result in Augenblick et al. (2021). Their analysis focuses on symmetric SGPs, where differences in signal strength can only be differences in absolute signal strength across different SGPs. Instead, we show that relative signal strength also drives the level of underinference by focusing on asymmetric SGPs where signals may have different strengths within the same SGPs.

[22] We only look at the effect of degree of asymmetry for bias, not underinference, as the effect of degree of asymmetry is hard to interpret for underinference, given that both measures are defined in terms of signal strength, as defined in Section 2.2.

asymmetric SGPs. The higher the degree of SGP's asymmetry, the more consequential this mistake is, leading to a higher bias. This may provide a reason why contrast reasoning becomes ineffective in reducing the bias for more asymmetric signals.

We classify pairs of guesses in *All-Contingency* as 'Treated as Symmetric' if there is an absolute distance between the reported posteriors $\Pr(A|b)$ and $\Pr(B|o)$ of at most one percentage point to detect the use of such a heuristic.[23]

Assuming that participants use this heuristic, we should observe (i) a substantial share of pairs of guesses in *All-Contingency* that are classified as 'Treated as Symmetric,' and (ii) that the bias increases more with the degree of asymmetry for those pairs of guesses than for those that do not fall under this classification. Both are indeed the case. Within all pairs of guesses for asymmetric SGPs in *All-Contingency*, 14.1% are classified as 'Treated as Symmetric.'[24] As Figure 7 illustrates, the increase in bias in *All-Contingency* with the degree of asymmetry can be attributed to choices classified as 'Treated as Symmetric.'

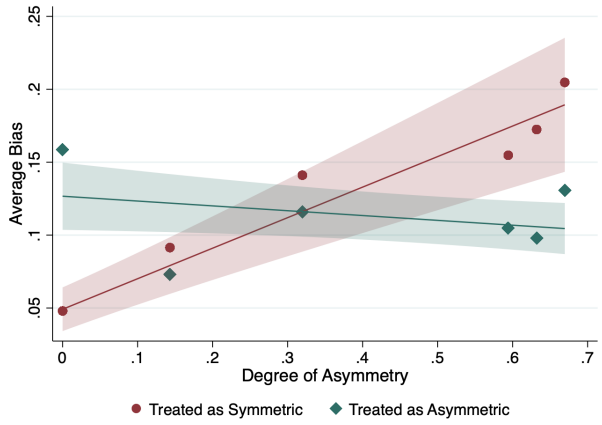


Figure 7: Treating an SGP as Symmetric and the Degree of Asymmetry

Notes. Average bias in *All-Contingency* by the degree of asymmetry for the pairs of guesses for asymmetric SGPs that are classified as 'Treated as Symmetric' and those that are not. Shaded regions indicate 95% confidence intervals.

The bias increases significantly more with the degree of asymmetry for guesses being classified as 'Treated as Symmetric' than for those that are not ($p < 0.001$; Column I in Table A2). Note that the share of guesses treating signals as symmetric decreases in the degree of asymmetry ($p < 0.001$). Nevertheless, we find that the entire increase in the bias with the degree of asymmetry is driven by the pairs of guesses classified as 'Treated as Symmetric.' ($p = 0.057$ for negative coefficient 'Degree of Asymmetry' in Column I in Table A2).

We explore two other potential sources of mistakes that may result in contrast reasoning being ineffective in counteracting hypothetical thinking for asymmetric signals. Both

[23]The results are similar if we use more or less strict classifications. We cannot use this classification in the other treatments, as we only elicit the posterior within one round for one contingency.

[24]The pairs of guesses for symmetric SGPs in *All-Contingency* classified 'Treated as Symmetric' is 75.3%. Figure A10 illustrates the scatter plot of these pairs of guesses by SGP symmetry.

involve errors in perceiving signal characteristics that arise when participants consider two contingencies simultaneously.

One possibility is that the participants report beliefs "as if" both signals are realized when asked for both contingencies on the same screen. By construction, for each SGP one signal is more diagnostic for bag A ($\lambda_s > 1$) and one signal is more diagnostic for bag B ($\lambda_{s'} < 1$). Thus, if participants misinterpret the treatment *All-Contingency* in this manner, beliefs would be reported as if they respond to a convex combination of the two signal strengths for bag A, λ_{blue} and λ_{orange}. This mistake would result in higher underinference in *All-Contingency* compared to *One-Contingency*, where hypothetical signals are considered separately. However, as we see in Section 4.2, this is not what we observe, ruling out this mechanism.

A third possibility is that participants mix the overall diagnosticity of the signals, perceiving a signal's diagnosticity as a convex combination of the two signal strengths, $\bar{\lambda}_{blue}$ and $\bar{\lambda}_{orange}$. This would imply an attenuated effect in *All-Contingency* compared to *One-Contingency*: There would be less underinference for the stronger signal and more underinference for the weaker signal in *All-Contingency* compared to *One-Contingency*. As shown in Figure 6 and Table A4, this is not the case.

Finding 5. *Contrast reasoning is less effective in reducing the bias for SGPs with a higher degree of asymmetry. This can be attributed to individuals who treat asymmetric SGPs as symmetric.*

4.3.2 Consistency Measures

Our analysis of the mechanisms continues by looking at the treatment effects on additional outcomes related to consistency.

Table 4: Consistency

	Δ Posteriors	Bayes-Inconsistent
All-Contingency	0.011	0.026
	(0.016)	(0.024)
One-Contingency	0.066**	0.081*
	(0.022)	(0.035)
Constant	0.112***	0.068**
	(0.015)	(0.021)
N	904	896
adj. R^2	0.016	0.006
Clusters	379	375

Notes. OLS estimates. Individual-level clustered standard errors. Symmetry SGP fixed effects. The dependent variable in Column I is the absolute difference in the reported posteriors for the same signal for mirrored SGPs, and in Column II a dummy taking value 1 if the vector of posteriors for mirrored SGPs is Bayes-inconsistent; * p<.05, ** p<.01, *** p<.001.

Within-Consistency Taking advantage of the mirrored SGPs, we investigate *within-consistency*: how stable the reported posteriors are within a task (beliefs elicited given the same signal for the same SGP). This measure allows us to evaluate whether the treatments have an important side effect: increasing the noise in how beliefs are updated.[25] Thus, examining this measure of consistency can provide valuable insights into the consequences of hypothetical thinking and contrast reasoning.

With this goal, our dependent variable ΔPosteriors is defined as the absolute difference between the posteriors for the probability of bag A given the same signal reported for two mirrored SGPs (see Appendix A.3.2 for details). While participants should report the same beliefs in both instances and this difference should be zero, our pooled data provides evidence of inconsistent beliefs for the same task: the average ΔPosteriors is 12 percentage points (statistically different from zero, with $p < 0.001$), with a median of 5 percentage points.[26]

Compared to *Conditional*, participants in *One-Contingency* are significantly more likely to be inconsistent ($p = 0.004$; Column I in Table 4). This is not the case in *All-Contingency* ($p = 0.477$; Column I in Table 4), where we observe higher levels of within-consistency than in *One-Contingency* ($p = 0.009$; see the difference of the coefficients of the treatment indicators in Column Δ Posteriors in Table 4). Therefore, hypothetical thinking leads to less within-consistent beliefs, while the presence of contrast reasoning counteracts this increase completely.

Between-Consistency So far, we have looked at measures of deviations from Bayesian updating given a signal realization. Next, we consider a way to categorize deviations from Bayesian updating by looking at the performance across contingencies: the consistency of the reported beliefs across signal realizations, given the same SGP (*between-consistency*). We investigate the impact of our treatments on this measure.

Bayes' rule prescribes that beliefs cannot be updated in the same direction for all signal realizations. Therefore, holding posteriors given both signal realizations either above or below the prior would be an extreme violation of Bayesian updating. For this analysis, we need for each participant the reported *vector of posterior beliefs*, that is, the posterior beliefs conditional on each signal realization given an SGP. We construct those thanks to our mirrored SGPs; see Appendix A.3.3 for details. Following Aina (2023), we say the reported vector of posteriors is *Bayes-inconsistent* if both posteriors are higher or lower than 50%. Bayes-inconsistency is an extreme form of deviation from Bayesian updating because not only are the posteriors different from the ones implied by the known SGP,

[25] To some extent, this measure is conceptually related to cognitive uncertainty under the assumption that participants are well-calibrated in assessing their own performance, which is not the case for belief-updating tasks (Enke et al., 2022). Also, since our measure of within-consistency and cognitive uncertainty are measured for different SGPs, they are not properly comparable.

[26] While on average beliefs are inconsistent within a task, a good portion of participants are perfectly consistent. Figure A9 shows the cumulative distribution of this measure by treatments. 30% are perfectly consistent in *All-Contingency*, 20% in *Conditional*, and 16% in *One-Contingency*.

but also it is impossible to find an SGP that would rationalize the reported vector of posteriors given the prior (Aina, 2023, Lemma 1). Bayes-inconsistency is quite rare: 6% in *Conditional*, 8% in *All-Contingency*, and 14% in *One-Contingency* in our mirrored SGPs.[27]

In support of our finding in Section 4.2, this analysis underlines the biasing effect of hypothetical thinking in the absence of contrast reasoning. In *One-Contingency*, we estimate that 8.1 percentage points more choices can be classified as Bayes-inconsistent ($p = 0.021$; Column II in Table 4). This is, even if only marginally significantly so, a larger increase than the statistically insignificant increase in *All-Contingency* ($p = 0.096$; see the difference of the coefficients of the treatment indicators in Column II in Table 4). Thus, there is suggestive evidence that contrast reasoning increases Bayes-consistency, while hypothetical thinking does the opposite.[28]

Taking together the evidence regarding within- and between-consistency, we can summarize the treatment effects of these additional measures as follows.

Finding 6. *Hypothetical thinking leads to more inconsistent belief updating both within a task and across contingencies. Due to the effect of contrast reasoning, the consistency of belief updating does not differ between contingent and conditional belief updating.*

4.3.3 Individual Measures

Finally, we examine the role of individual measures both for heterogeneous treatment effects and additional measures.

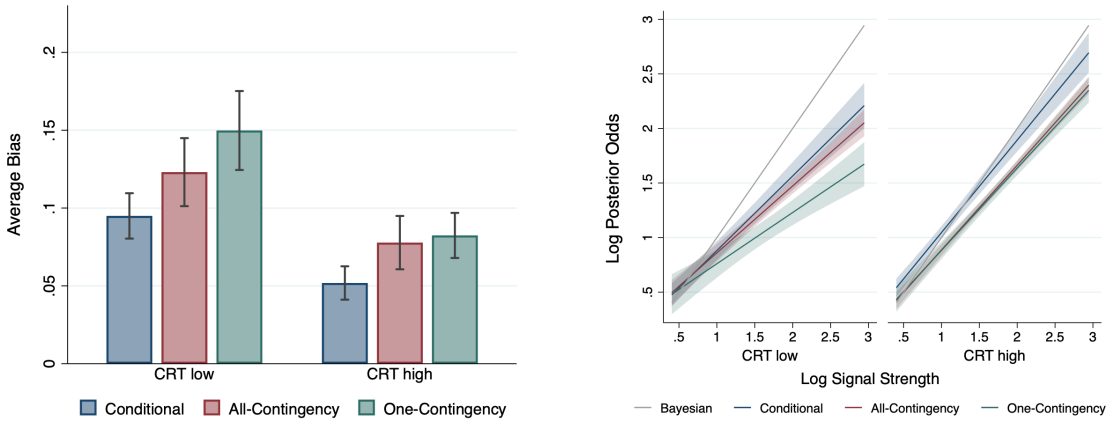
Cognitive Reflection Test We study the moderating effect of a participant's cognitive reflection capacity, as measured by the Cognitive Reflection Task (CRT), on our treatments. The CRT measures an individual's tendency to override intuitive responses and engage in reflective and analytical thinking (Frederick, 2005); it appears to correlate with mental heuristics also related to belief updating (Oechssler et al., 2009; Toplak et al., 2011; Hoppe and Kusterer, 2011; Augenblick et al., 2021).

We categorized participants who made one or no mistakes on the CRT as *high CRT* (56%), those who made two or more mistakes were categorized as *low CRT* (44%).[29] Figure 8a illustrates the average bias in posterior beliefs by treatment and CRT. In line with the existing literature, individuals classified as low CRT exhibit significantly higher biases, underlining that cognitive reflection captures a component relevant to

[27] For *All-Contingency*, vectors of posteriors are available for all SGPs: 11% are Bayes-inconsistent.

[28] We report analogous findings for a more nuanced measure of between-consistency: the squared distance between the reported and Bayesian vectors of posteriors. See Table A6 in the appendix.

[29] We modified the original version of the CRT, as reported in Appendix C.3, to avoid confounds in the event that subjects have previously been exposed to the classic version of the CRT. Out of the three questions, 26% of our participants made no mistakes, 30% made one mistake, 25% made two mistakes, and 19% made three mistakes. See Figure A4 in the appendix for an illustration of this heterogeneity using the full CRT scale (0-3) instead of the binary classification. The results are qualitatively comparable.



(a) Bias by CRT

(b) Underinference by CRT

Figure 8: Treatment Effect by CRT

Notes. Panel (a) shows the average bias defined as the absolute value of the difference between the posterior reported by participants and the normative (Bayesian) benchmark by treatment split by the participants' CRT. Panel (b) shows the estimated relationship between the log posterior-odds ratio and the log signal strength by treatment as an illustration of underinference split by the participants' CRT. Error bars in Panel (a) and shaded areas in Panel (b) indicate 95% confidence intervals, clustered at the individual level.

belief updating. If beliefs are elicited conditional on an observed signal as in *Conditional*, individuals with a high CRT are on average 4.3 percentage points closer to the Bayesian posterior ($p < 0.001$; Column IV in Table 2). Similarly, a high CRT implies lower levels of underinference ($p = 0.004$; Column III in Table 3).

While a high CRT is associated with a lower bias and underinference in all three treatments, CRT seems to have no effect on hypothetical thinking ($p = 0.165$; Column IV of Table 2) nor on contrast reasoning ($p = 0.282$; see the difference between 'All-Contingency $\times$ High CRT' and 'One-Contingency $\times$ High CRT' in Column IV in Table 2). Column III in Table 3 reports equivalent results for underinference.

Cognitive Uncertainty Next, we look at whether our treatments impact cognitive uncertainty, which is measured for the last task in the experiment. Enke and Graeber (2023) define cognitive uncertainty as "[...] *people's subjective uncertainty over which decision maximizes their expected utility*". They show that in a belief-updating setting, an increase in cognitive uncertainty is associated with a stronger bias. It is therefore relevant to assess to what extent cognitive uncertainty responds to hypothetical thinking and contrast reasoning in this setting.

Pooling all treatments, we find that an increase in cognitive uncertainty increases the bias ($p = 0.001$): the more uncertain individuals tend to report more biased posteriors, consistent with Enke and Graeber (2023). Across treatments, we find no difference. The average cognitive uncertainty in *Conditional* is not significantly different from the one in *One-Contingency* ($p = 0.306$) or *All-Contingency* ($p = 0.657$). This suggests that cognitive uncertainty is affected by neither hypothetical thinking nor contrast reasoning.

However, note that cognitive uncertainty was elicited only for the 70-30 SGP, for which we find no significant treatment effects.[30]

Measures of Difficulty In what follows, we consider two measures of difficulty across treatments: response time and self-reported degree of challenge in completing the tasks.[31]

Table 5: Individual Measures

	Time	Challenge	Bias	Log Odds
All-Contingency	18.719***	0.380*	0.028*	0.049
	(2.913)	(0.172)	(0.013)	(0.095)
One-Contingency	3.819	0.540**	0.055***	0.092
	(2.386)	(0.172)	(0.015)	(0.115)
High CRT			-0.043***	-0.001
			(0.009)	(0.087)
All-Contingency × High CRT			-0.002	-0.137
			(0.017)	(0.116)
One-Contingency × High CRT			-0.024	-0.162
			(0.018)	(0.142)
Log Signal Strength				0.683***
				(0.049)
Log Signal Strength × All-Contingency				-0.070
				(0.073)
Log Signal Strength × One-Contingency				-0.214*
				(0.084)
Log Signal Strength × High CRT				0.165*
				(0.067)
Log Signal Strength × All-Contingency × High CRT				-0.001
				(0.093)
Log Signal Strength × One-Contingency × High CRT				0.120
				(0.108)
Constant	33.330***	4.407***	0.085***	0.199**
	(3.179)	(0.122)	(0.010)	(0.067)
N	6000	6000	6000	6000
adj. R^2	0.037	0.016	0.053	0.262
Clusters	450	450	450	450

Notes. OLS estimates. Individual-level clustered standard errors. SGP fixed effects. The dependent variable in Column I is the response time measured in seconds, in Column II the perceived degree of being challenged, in Column III the absolute bias, and in Column IV the logarithm of the ratio of the elicited posterior belief for a given signal; * p<.05, ** p<.01, *** p<.001.

Response time is an important measure in economics because it can provide insights into the cognitive processes that underlie decision-making. An emerging strand of literature has been focusing on the role of response time and revealed preferences (e.g., Woodford,

[30]As discussed in Section 2.2, we elicit cognitive uncertainty only for an SGP, the most similar to the ones used in the literature. We also show in Section 4.2 how we replicate for this SGP in *Conditional* quantitative findings of previous studies. Running the OLS regressions of the bias on indicators of the different treatment effects (same as in Column I in Table 2) for each SGP, we find that 70-30 is the only SGP for which there is no treatment effect for either *All-Contingency* ($p = 0.728$) or *One-Contingency* ($p = 0.10$). For all other SGPs, at least one of the treatment effects is significant at the 5% level.

[31]In the final questionnaire, participants also answered an unincentivized question about how challenged they felt during the guessing tasks on a 7-point scale.

2014; Krajbich et al., 2015; Echenique and Saito, 2017; Alós-Ferrer et al., 2021; Schotter and Trevino, 2021). We regard the response time as a proxy of the indirect costs associated with the belief elicitation method, given the comparable strength of incentives across treatments. Longer response times may indicate that individuals are facing a more complex task, reflected in higher indirect costs.[32]

On average, the response time for each task is 27 seconds in *Conditional*, 46 in *All-Contingency*, and 31 in *One-Contingency*. We estimate that per elicitation task, the participants take more than 50% longer in treatment *All-Contingency* than in treatment *Conditional* ($p < 0.001$; Column 'Time' in Table 5). At the same time, the hypothetical nature of signals in *One-Contingency* appears not to decrease engagement with the task ($p = 0.110$; Column 'Time' in Table 5). This suggests that the longer response time is due to contrast reasoning, not hypothetical thinking.

The perceived level of challenge serves as a complementary measure to response time in assessing the difficulty in each treatment. Unlike for response time, the self-reported challenge level is significantly higher ($p = 0.002$; Column 'Challenge' in Table 5) in *One-Contingency* compared to *Conditional*. In other words, participants perceive a greater challenge when engaging in hypothetical thinking despite not dedicating significantly more time to solve each task. Interestingly, contrast reasoning does not increase the perceived level of challenge despite the longer response time and the higher computational complexity. If anything, the reported level is lower in *All-Contingency* than in *One-Contingency*, but not significantly so ($p = 0.351$).

5 Discussion

Our findings reveal a surprising effect of contingent thinking on how we process new information. Contrary to the majority of surveyed experts predicting an equal bias in conditional and contingent belief updating, our results indicate a different and more nuanced picture. Contingent belief updating can lead to less accurate beliefs than conditional belief updating. However, the effect is not uniform. We show how the effect varies depending on the characteristics of the signal-generating process. Our findings suggest that the effect is mediated by the complexity of the information structure (symmetry of the SGP) but not by one's ability to engage with it (performance in CRT).

To learn more about the mechanisms behind this finding, we decompose the effect of contingent thinking into hypothetical thinking and contrast reasoning using a treatment that requires engaging only in the former. On the one hand, our findings show a harmful effect of hypothetical thinking that is systematic across a wide range of measures of deviations from Bayesian updating. Thus, the results cast doubt on our ability to properly

[32]Taking more time to perform a task could also be due to the fact that the participants are engaging in more deliberate and reflective thinking. Indeed, participants pooled across treatments exhibit a lower bias when taking more time ($p = 0.033$). However, we cannot disentangle these two channels and also consider the higher engagement as an indirect cost.

process information that may only realize in a future scenario. This suggests that simulating a prospective scenario requires exerting mental effort. On the other hand, this data suggests that contrast reasoning can compensate to some extent for the negative consequences of hypothetical thinking. The range of this effect is broad: from fully compensating with symmetric SGPs, it continuously decreases in the degree of asymmetry of SGPs due to some individuals treating asymmetric SGPs as if they were symmetric.

These results have broad implications for situations that involve contingent thinking in belief updating, such as information acquisition, information design, and experimentation. Specifically, our findings offer an explanation for the "compression effect" in information acquisition documented in Ambuehl and Li (2018), where individuals undervalue informative sources due to underinference in belief updating. We provide a new perspective on this result: when individuals consider acquiring information, they also engage in contingent belief updating, which leads to underinference and can generate the compression effect. Additionally, our results open a promising avenue for future research regarding strategic interactions both theoretically and empirically. This is because contingent belief updating is often required to formulate a strategy in sequential games with uncertainty about players' types or unknown payoff-relevant variables.

Finally, we want to address similarities and differences in our results with the emerging literature on failures of contingent thinking. In the recent survey, Niederle and Vespa (2023) argue that there are failures of contingent thinking *"when an agent does optimize in a presentation of the problem that helps her focus on all relevant contingencies (i.e., contingencies in which choices can result in different consequences), but does not optimize if the problem is presented without such aids (i.e., standard representation)."* At first glance, it would seem that we report the opposite effect, but this is not the case. There are important differences in our research questions but similarities in the reported findings. As highlighted in the introduction, the main difference is not only the type of tasks — choosing an action vs. updating beliefs — but rather the type of suboptimal behavior studied and the overall problem structure. Suboptimal behavior in Martínez-Marquina et al. (2019) and Esponda and Vespa (2023) arise because participants should think contingently and fail to do so when making a choice ahead of the resolution of uncertainty, commonly implemented for all contingencies. Thus, individuals behave optimally when placed in the relevant contingency but struggle to determine the correct (common) action without knowing the realized contingency. Similarly, our paper also shows that beliefs are less biased when people observe the relevant contingency. However, we do not compare this to a setting where people choose an ex-ante action implemented across contingencies. Instead, we study how people determine their contingency-specific behavior. We find that people struggle when they have to update beliefs that may become relevant in a not-yet-observed contingency. So here, people are placed in a setting in which they have to think contingently, but doing so might bias how they would react if they were to observe the relevant contingency. Interestingly, a common aspect drives both suboptimal behaviors: biases related to thinking about hypothetical events.

Pitfalls of hypothetical thinking seem not to be limited to a specific type of task but rather to numerous instances of failures of contingent thinking (Esponda and Vespa, 2014; Ali et al., 2021; Farina and Leccese, 2024) and other relevant contexts (Paolacci and André, 2023; Gandhi et al., 2023). We show its relevance in a new setting, belief updating. Theoretical approaches have recently emerged to incorporate this bias in simulations of expected future utilities (Piermont and Zuazo-Garin, 2020; Piermont, 2021), to link associative memory to belief formation about novel risks (Bordalo et al., 2022a), to explore cognitive frictions that could explain differences in ex-ante and ex-post posteriors (Samuelson and Steiner, 2024; Bohren and Hauser, 2024). Also, Cohen and Li (2022) consider an extensive-form solution concept where players neglect the information from hypothetical events. These approaches can account for the biases introduced by hypothetical thinking. However, the effect of contrast reasoning is underexplored, both experimentally and theoretically. It would be valuable to develop formal models to incorporate both hypothetical thinking and contrast reasoning. As suggested by Bordalo et al. (2023), the framing of the inference problem can lead to observing different type of biased beliefs. A crucial challenge posed by formalize our findings is to first develop a convincing model that explains the differences in belief updating depending on the asymmetry of the signal-generating process.

One question that remains is whether the presence of contrast reasoning could extend beyond merely neutralizing hypothetical thinking and thus lead to more accurate beliefs in other contexts. Two potential avenues could address this. One approach is to explore this question in settings where contingencies are more concrete and familiar to the participants. The stylized and abstract setting of this study allows us to have a well-grounded benchmark in the literature and easily vary conditions over rounds; however, it might have also amplified the difficulty of imaging hypothetical contingencies. Another potential avenue is integrating contingent belief updating with nudging or training. For example, we could emphasize the importance of seriously imagining the proposed contingencies and encourage participants to contrast their answers across contingencies before proceeding. A novel paper by Ashraf et al. (2022) shows that the ability to imagine the forward-oriented scenario can be trained, and it is linked to improved economic outcomes. Enhancing this type of training to promote contrast reasoning might boost this effect further.

References

Agranov, Marina, Utteeyo Dasgupta, and Andrew Schotter (2020) "Trust me: Communication and competition in psychological games."

Aina, Chiara (2023) "Tailored Stories."

Aina, Chiara, Pierpaolo Battigalli, and Astrid Gamba (2020) "Frustration and anger in the Ultimatum Game: An experiment," *Games and Economic Behavior*, 122, 150–167.

Alan, Sule and Seda Ertac (2018) "Fostering patience in the classroom: Results from randomized educational intervention," *Journal of Political Economy*, 126 (5), 1865–1911.

Ali, S Nageeb, Maximilian Mihm, Lucas Siga, and Chloe Tergiman (2021) "Adverse and advantageous selection in the laboratory," *American Economic Review*, 111 (7), 2152–2178.

Alós-Ferrer, Carlos, Ernst Fehr, and Nick Netzer (2021) "Time will tell: Recovering preferences when choices are noisy," *Journal of Political Economy*, 129 (6), 1828–1877.

Ambuehl, Sandro and Shengwu Li (2018) "Belief updating and the demand for information," *Games and Economic Behavior*, 109, 21–39.

Ashraf, Nava, Gharad Bryan, Alexia Delfino, Emily A Holmes, Leonardo Iacovone, and Ashley Pople (2022) "Learning to See the World's Opportunities: The Impact of Imagery on Entrepreneurial Success," *London School of Economics & Political Science, Working Paper*.

Augenblick, Ned, Eben Lazarus, and Michael Thaler (2021) "Overinference from Weak Signals and Underinference from Strong Signals," *arXiv preprint arXiv:2109.09871*.

Ba, Cuimin, J Aislinn Bohren, and Alex Imas (2022) "Over-and Underreaction to Information," *Available at SSRN*.

Becker, Christoph K, Tigran Melkonyan, Eugenio Proto, Andis Sofianos, and Stefan T Trautmann (2020) "Reverse Bayesianism: Revising beliefs in light of unforeseen events."

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

Bohren, J Aislinn and Daniel N Hauser (2024) *The Behavioral Foundations of Model Misspecification: A Decomposition*.

Bordalo, Pedro, Giovanni Burro, Katherine B Coffman, Nicola Gennaioli, and Andrei Shleifer (2022a) "Imagining the future: memory, simulation and beliefs about COVID," *National Bureau of Economic Research Working Paper*.

Bordalo, Pedro, John J Conlon, Nicola Gennaioli, Spencer Yongwook Kwon, and Andrei Shleifer (2023) "How People Use Statistics."

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer (2022b) "Salience," *Annual Review of Economics*, 14 (1), 521–544.

Brandts, Jordi and Gary Charness (2003) "Truth or consequences: An experiment," *Management Science*, 49 (1), 116–130.

——— (2011) "The strategy versus the direct-response method: a first survey of experimental comparisons," *Experimental Economics*, 14 (3), 375–398.

Brosig, Jeannette, Joachim Weimann, and Chun-Lei Yang (2003) "The hot versus cold effect in a simple bargaining experiment," *Experimental Economics*, 6, 75–90.

Byrne, Ruth (2016) "Counterfactual Thought," *Annual Review of Psychology*, 67.

Casari, Marco and Timothy N Cason (2009) "The strategy method lowers measured trustworthy behavior," *Economics Letters*, 103 (3), 157–159.

Charness, Gary, Uri Gneezy, and Vlastimil Rasocha (2021a) "Experimental methods: Eliciting beliefs," *Journal of Economic Behavior & Organization*, 189, 234–256.

Charness, Gary, Ryan Oprea, and Sevgi Yuksel (2021b) "How do people choose between biased information sources? Evidence from a laboratory experiment," *Journal of the European Economic Association*, 19 (3), 1656–1691.

Cipriani, Marco and Antonio Guarino (2009) "Herd behavior in financial markets: an experiment with financial market professionals," *Journal of the European Economic Association*, 7 (1), 206–233.

Cohen, Shani and Shengwu Li (2022) "Sequential Cursed Equilibrium," *arXiv preprint arXiv:2212.06025*.

Danz, David, Lise Vesterlund, and Alistair J Wilson (2022) "Belief elicitation and behavioral incentive compatibility," *American Economic Review*, 112 (9), 2851–83.

Dube, Oeindrila, Sandy Jo MacArthur, and Anuj K Shah (2023) "A cognitive view of policing,"Technical report, National Bureau of Economic Research.

Echenique, Federico and Kota Saito (2017) "Response time and utility," *Journal of Economic Behavior & Organization*, 139, 49–59.

Enke, Benjamin, Uri Gneezy, Brian Hall, David Martin, Vadim Nelidov, Theo Offerman, and Jeroen van de Ven (2023) "Cognitive Biases: Mistakes or Missing Stakes?," *The Review of Economics and Statistics*, 105 (4), 818–832.

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

Enke, Benjamin, Thomas Graeber, and Ryan Oprea (2022) "Confidence, self-selection and bias in the aggregate."

Epstude, Kai and Neal Roese (2008) "The Functional Theory of Counterfactual Thinking," *Personality and Social Psychology Review*, 12, 168–92.

Esponda, Ignacio and Emanuel Vespa (2014) "Hypothetical thinking and information extraction in the laboratory," *American Economic Journal: Microeconomics*, 6 (4), 180–202.

——— (2023) "Contingent preferences and the sure-thing principle: Revisiting classic anomalies in the laboratory," *Review of Economic Studies*.

Esponda, Ignacio, Emanuel Vespa, and Sevgi Yuksel (2020) "Mental Models and Learning: The Case of Base-Rate Neglect."

Farina, Agata and Mario Leccese (2024) "Hiding a Flaw: A Lab Experiment on Multi-Dimensional Information Disclosur."

Ferrante, Donatella, Vittorio Girotto, Marta Stragà, and Clare Walsh (2012) "Improving the Past and the Future: A Temporal Asymmetry in Hypothetical Thinking," *Journal of Experimental Psychology: General*, 142.

Frederick, Shane (2005) "Cognitive reflection and decision making," *Journal of Economic Perspectives*, 19 (4), 25–42.

Gandhi, Linnea, Anoushka Kiyawat, Colin Camerer, and Duncan J Watts (2023) "Hypothetical nudges provide misleading estimates of real behavior change."

Gerstenberg, Tobias (2022) "What would have happened? Counterfactuals, hypotheticals and causal judgements," *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377 (1866), 20210339.

Gonçalves, Duarte, Jonathan Libgober, and Jack Willis (2024) "Retractions: Updating from Complex Information."

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

Hoch, Stephen J. (1985) "Counterfactual reasoning and accuracy in predicting personal events," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11, 719–731.

Hoppe, Eva I and David J Kusterer (2011) "Behavioral biases and cognitive reflection," *Economics Letters*, 110 (2), 97–100.

Hossain, Tanjim and Ryo Okui (2013) "The binarized scoring rule," *Review of Economic Studies*, 80 (3), 984–1001.

John, Anett and Kate Orkin (2022) "Can simple psychological interventions increase

preventive health investment?" *Journal of the European Economic Association*, 20 (3), 1001–1047.

Kahneman, Daniel and Amos Tversky (1982) *The Simulation Heuristic*, 201–208: Cambridge University Press.

Karni, Edi and Marie-Louise Vierø (2013) ""Reverse Bayesianism": A choice-based theory of growing awareness," *American Economic Review*, 103 (7), 2790–2810.

——— (2017) "Awareness of unawareness: A theory of decision making in the face of ignorance," *Journal of Economic Theory*, 168, 301–328.

Kozakiewicz, Marta (2022) "Belief-Based Utility and Signal Interpretation."

Krajbich, Ian, Björn Bartling, Todd Hare, and Ernst Fehr (2015) "Rethinking fast and slow based on a critique of reaction-time reverse inference," *Nature Communications*, 6 (1), 7455.

Li, Shengwu (2017) "Obviously strategy-proof mechanisms," *American Economic Review*, 107 (11), 3257–3287.

Lilley, Matthew and Brian Wheaton (2024) "Are Preconceptions Postconceptions? Evidence on Motivated Political Reasoning."

Martínez-Marquina, Alejandro, Muriel Niederle, and Emanuel Vespa (2019) "Failures in contingent reasoning: The role of uncertainty," *American Economic Review*, 109 (10), 3437–74.

Mitzkewitz, Michael and Rosemarie Nagel (1993) "Experimental results on ultimatum games with incomplete information," *International Journal of Game Theory*, 22, 171–198.

Ngangoué, M Kathleen and Georg Weizsäcker (2021) "Learning from unrealized versus realized prices," *American Economic Journal: Microeconomics*, 13 (2), 174–201.

Niederle, Muriel and Emanuel Vespa (2023) "Cognitive Limitations: Failures of Contingent Thinking," *Annual Review of Economics*, 15, 307–328.

Oechssler, Jörg, Andreas Roider, and Patrick W Schmitz (2009) "Cognitive abilities and behavioral biases," *Journal of Economic Behavior & Organization*, 72 (1), 147–152.

Paolacci, Gabriele and Quentin André (2023) "Probabilistic outcomes are valued less in expectation, even conditional on their realization," *Management Science*.

Pearl, Judea (2009) *Causality: Models, Reasoning and Inference*, USA: Cambridge University Press, 2nd edition.

Piermont, Evan (2021) "Hypothetical Expected Utility," *arXiv preprint arXiv:2106.15979*.

Piermont, Evan and Peio Zuazo-Garin (2020) "Failures of contingent thinking," *arXiv preprint arXiv:2007.07703*.

Samuelson, Larry and Jakub Steiner (2024) "Constrained Data-Fitters," *Yale University and University of Zurich, CERGE-EI, and CTS*.

Schipper, Burkhard C (2022) "Predicting the unpredictable under subjective expected utility."

Schlag, Karl H, James Tremewan, and Joël J Van der Weele (2015) "A penny for your thoughts: A survey of methods for eliciting beliefs," *Experimental Economics*, 18 (3), 457–490.

Schotter, Andrew and Isabel Trevino (2014) "Belief elicitation in the laboratory," *Annu. Rev. Econ.*, 6 (1), 103–128.

——— (2021) "Is response time predictive of choice? An experimental study of threshold strategies," *Experimental Economics*, 24, 87–117.

Toplak, Maggie E, Richard F West, and Keith E Stanovich (2011) "The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks," *Memory & Cognition*, 39 (7), 1275–1289.

Toussaert, Séverine (2017) "Intention-based reciprocity and signaling of intentions," *Journal of Economic Behavior & Organization*, 137, 132–144.

Woodford, Michael (2014) "Stochastic choice: An optimizing neuroeconomic model," *American Economic Review*, 104 (5), 495–500.

——— (2020) "Modeling imprecision in perception, valuation, and choice," *Annual Review of Economics*, 12 (1), 579–601.

A Appendix: Supplementary Analysis

A.1 Bias

Table A1: Bias with SGP×Contingency fixed effects

	I	II	III
All-Contingency	0.022*	0.010	0.026
	(0.009)	(0.011)	(0.013)
One-Contingency	0.038***	0.045***	0.054***
	(0.009)	(0.011)	(0.015)
Asymmetric		0.014	
		(0.012)	
All-Contingency × Asymmetric		0.020*	
		(0.009)	
One-Contingency × Asymmetric		-0.012	
		(0.010)	
High CRT			-0.043***
			(0.009)
All-Contingency × High CRT			-0.002
			(0.017)
One-Contingency × High CRT			-0.026
			(0.017)
Constant	0.057***	0.061***	0.080***
	(0.010)	(0.011)	(0.012)
N	6000	6000	6000
adj. R^2	0.027	0.028	0.056
Clusters	450	450	450

Notes. OLS estimates. Individual-level clustered standard errors. SGP×contingency fixed effects. Contingency refers to the realized signal in *Conditional* and to the relevant contingency in *One-Contingency* and *All-Contingency*. The dependent variable is defined as the absolute value of the difference between the posterior reported by participants and the normative (Bayesian) benchmark; * p<.05, ** p<.01, *** p<.001.

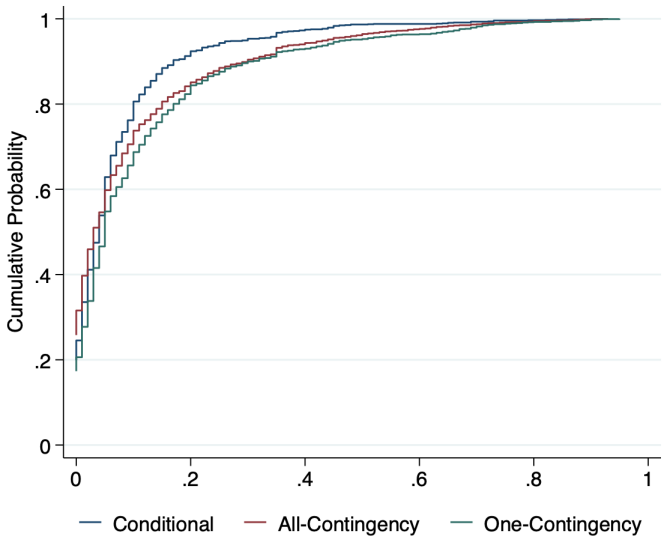


Figure A1: Cumulative Distribution of Bias

Notes. Cumulative distribution function of the bias by treatment.

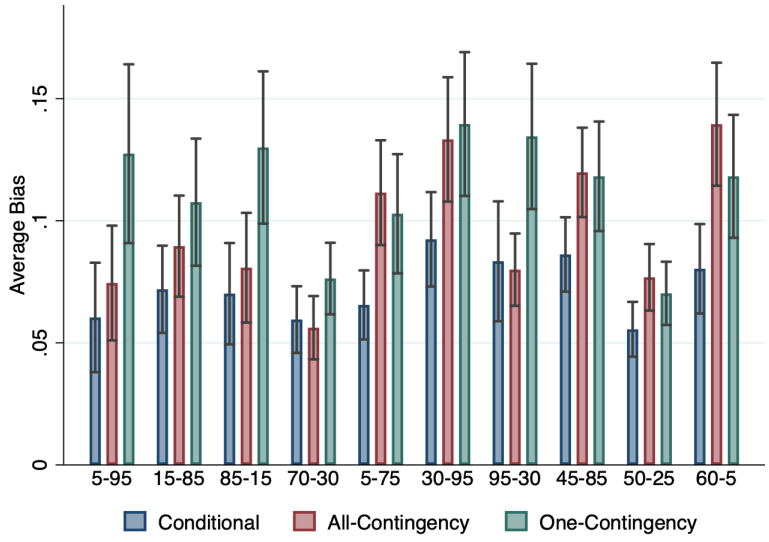


Figure A2: Treatment Effect in Bias by SGP

Notes. Each triplet of histograms represents the average bias by treatment and SGP. SGPs labels, reported on the x-axis, report the number of blue balls in the first and second bag, respectively (e.g., "5-75" indicated that for that SGP the first bag contained 5 blue balls and the second 75). Error bars indicate 95% confidence intervals.

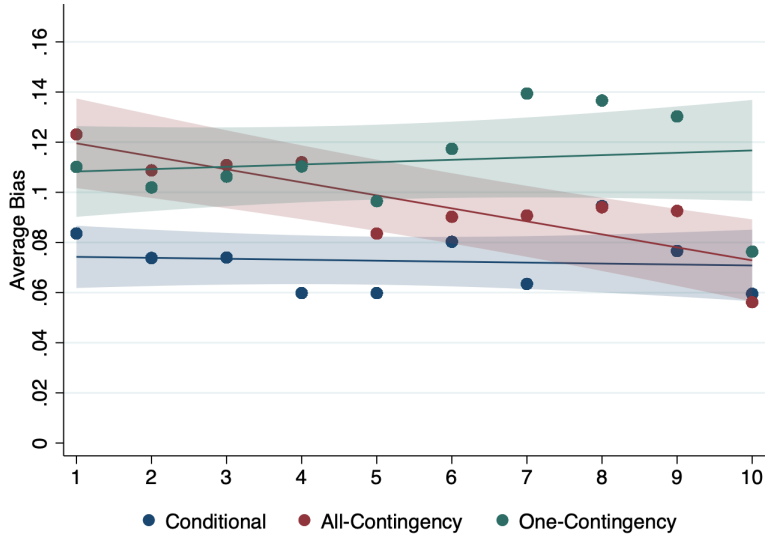


Figure A3: Trend in Average Bias by Trial

Notes. Each point represents the average bias by treatment and trial number. Lines indicate the linear predictions of the average bias by treatment, the shaded regions indicate 95% confidence intervals of these predictions.

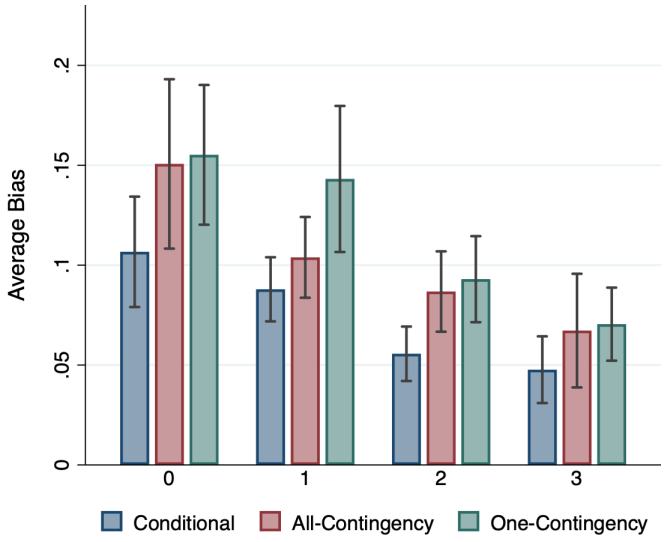


Figure A4: Treatment Effect in Bias by CRT scale

Notes. Each triplet of histograms represents the average bias by treatment and CRT level. Specifically, the latter is measured as the number of CRT questions correctly answered by participants, indicated on the x-axis label. Error bars indicate 95% confidence intervals.

Table A2: Bias and Treated as Symmetric

	I
Treated as Symmetric	-0.078***
	(0.013)
Degree of Asymmetry	-0.033
	(0.017)
Treated as Symmetric × Degree of Asymmetry	0.243***
	(0.036)
Constant	0.127***
	(0.012)
N	3000
adj. R^2	0.052
Clusters	150

Notes. OLS estimates. Individual-level clustered standard errors. The dependent variable is defined as the absolute value of the difference between the posterior reported by participants and the normative (Bayesian) benchmark; * p<.05, ** p<.01, *** p<.001.

A.2 Underinference

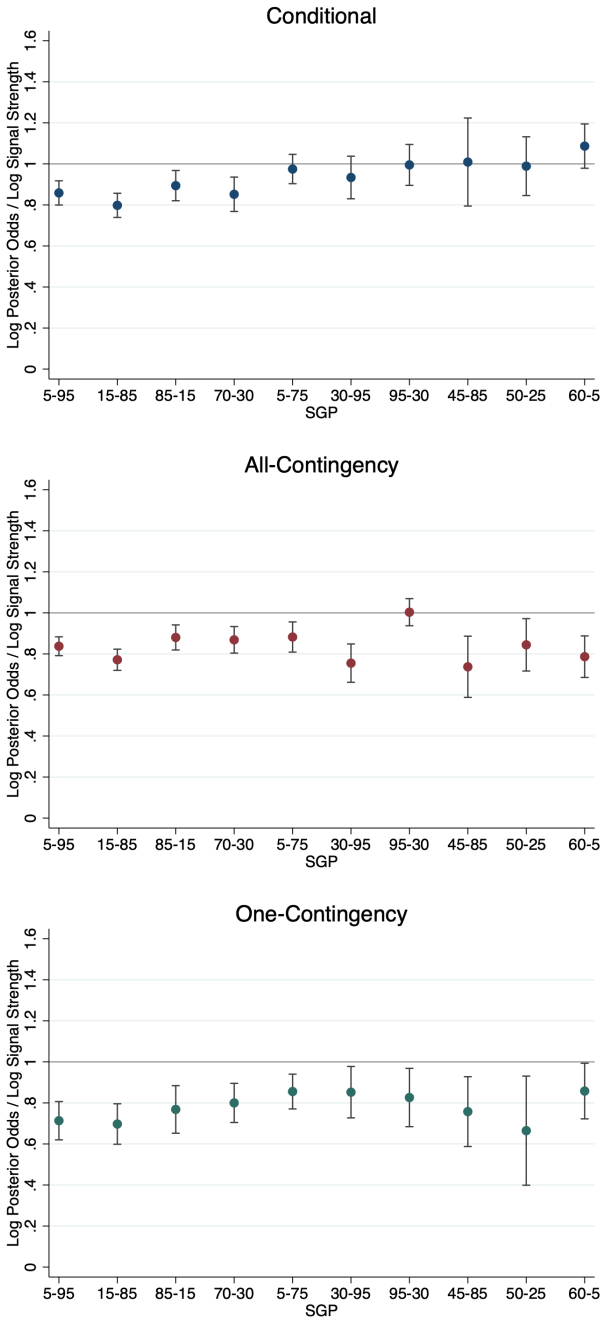


Figure A5: Underinference by SGP and Treatment

Notes. Each figure plots the estimated degree of underinference, measured as the average ratio of the reported log posterior-odds to the log signal strength for each SPG in a given treatment. The horizontal line at value one serves as the Bayesian benchmark: ratios below one indicate evidence of underinference, while ratios above one suggest evidence of overinference. Error bars indicate 95% confidence intervals.

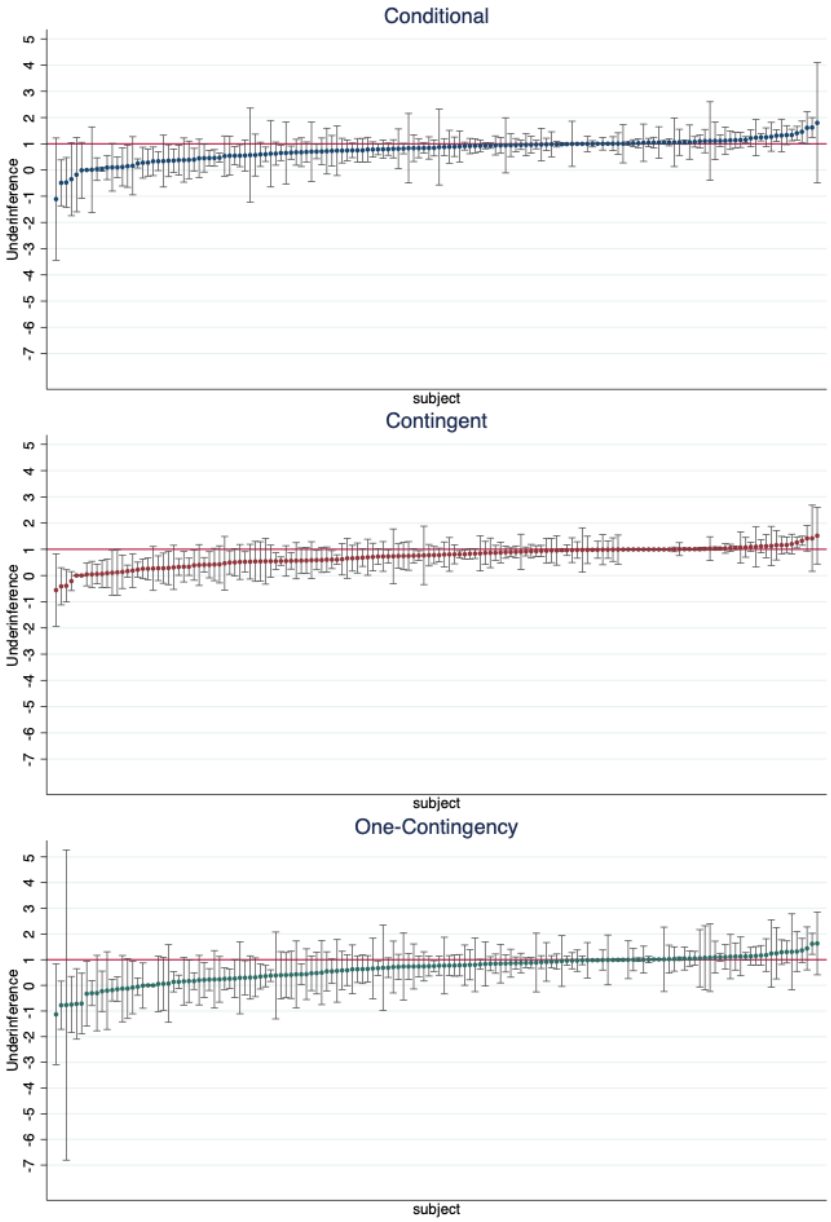


Figure A6: Underinference per Individual

Notes. Each figure plots the estimated degree of underinference, measured as the average ratio of the reported log posterior-odds to the log signal strength for each individual in a given treatment. The horizontal line at value one serves as the Bayesian benchmark: ratios below one indicate evidence of underinference, while ratios above one suggest evidence of overinference. Error bars indicate 95% confidence intervals.

A.2.1 Overinference as defined in Ba et al. (2022)

Ba et al. (2022) introduce an alternative measure of underinference, which we use as a dependent variable as a robustness check. For a contingency-specific guess $\hat{\Pr}(A|s)$, the ratio used is defined as:

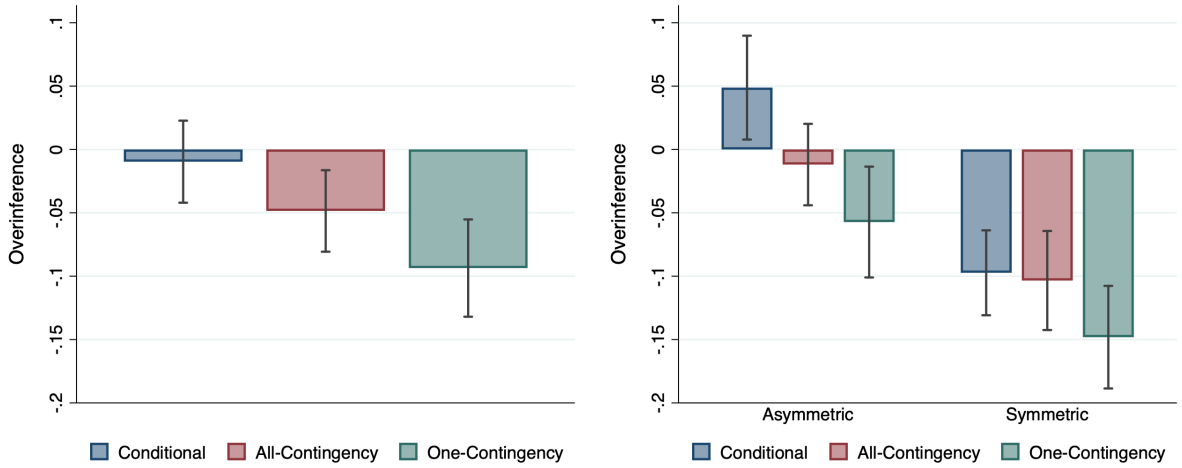
$$\frac{|\hat{\Pr}(A|s) - 0.5| - |\Pr(A|s) - 0.5|}{|\Pr(A|s) - 0.5|}$$

Table A3 replicate our main analysis using this ratio as a dependent variable.

Table A3: Overinference as defined in Ba et al. (2022)

	I	II	III
All-Contingency	-0.039	-0.006	-0.015
	(0.023)	(0.026)	(0.034)
One-Contingency	-0.084**	-0.051	-0.084
	(0.026)	(0.027)	(0.045)
Asymmetric		0.129***	
		(0.026)	
All-Contingency × Asymmetric		-0.055*	
		(0.026)	
One-Contingency × Asymmetric		-0.055	
		(0.028)	
High CRT			0.050
			(0.033)
All-Contingency × High CRT			-0.046
			(0.046)
One-Contingency × High CRT			-0.003
			(0.053)
Constant	-0.074***	-0.098***	-0.100***
	(0.018)	(0.019)	(0.027)
N	6000	6000	6000
adj. R^2	0.033	0.033	0.034
Clusters	450	450	450

Notes. OLS estimates. Individual-level clustered standard errors. The dependent variable is calculated as the difference of the absolute distances of (i) the elicited posterior and prior and, (ii) the Bayesian posterior and the prior, divided by the absolute difference of the Bayesian posterior and the prior; * p<.05, ** p<.01, *** p<.001.



(a) Overinference

(b) Overinference by Symmetry

Figure A7: Overinference (Ba et al., 2022) by Treatment and Symmetry

Notes. Panel (a) shows the average overinference, defined as as the difference of the absolute distances of (i) the elicited posterior and prior and, (ii) the Bayesian posterior and the prior, divided by the absolute difference of the Bayesian posterior and the prior, by treatment. Panel (b) shows this average overinference measure by treatment and symmetry of the SGP. Error bars indicate 95% confidence intervals, clustered at the individual level.

A.2.2 Relative Signal Strength

Next, we compare stronger and weaker signals within a specific SGP. For symmetric SGPs, signals are always equally strong because $\bar{\lambda}_s = \bar{\lambda}_{s'}$. Instead, for asymmetric SGPs, we say that a signal s is *stronger* than signal s' if $\bar{\lambda}_s > \bar{\lambda}_{s'}$.

Table A4: Relative Underinference

	I	II	III
Stronger Signal	-0.171***	0.021	-0.203***
	(0.033)	(0.035)	(0.055)
All-Contingency	-0.108*	-0.008	-0.075
	(0.048)	(0.043)	(0.086)
One-Contingency	-0.133*	-0.100	-0.103
	(0.054)	(0.051)	(0.096)
Stronger Signal × All-Contingency	0.036	-0.007	0.019
	(0.042)	(0.040)	(0.075)
Stonger Signal × One-Contingency	0.001	-0.012	-0.065
	(0.051)	(0.063)	(0.086)
Asymmetric		0.274***	
		(0.049)	
Stonger Signal × Asymmetric		-0.303***	
		(0.062)	
Stronger Signal × All-Contingency × Asymmetric		0.054	
		(0.079)	
Stronger Signal × One-Contingency × Asymmetric		0.003	
		(0.120)	
High CRT			0.108
			(0.067)
Stronger Signal × High CRT			0.063
			(0.067)
Stronger Signal × All-Contingency × High CRT			0.019
			(0.088)
Stronger Signal × One-Contingency × High CRT			0.104
			(0.102)
Constant	1.018***	0.841***	0.962***
	(0.033)	(0.031)	(0.057)
N	6000	6000	6000
adj. R^2	0.016	0.026	0.021
Clusters	450	450	450

Notes. OLS estimates. Individual-level clustered standard errors. The dependent variable is calculated as the ratio of the log posterior-odds and log signal strength for a given signal; * p<.05, ** p<.01, *** p<.001.

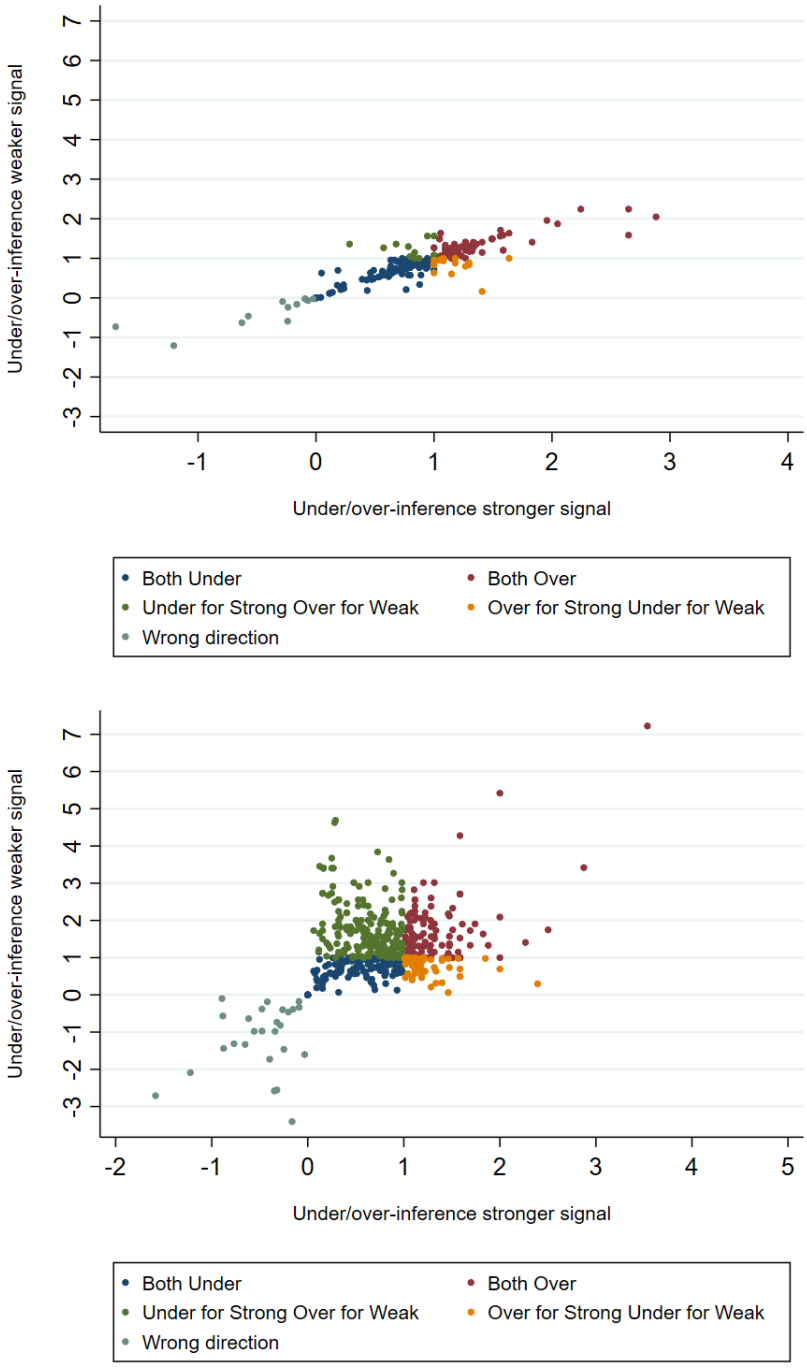


Figure A8: Relative Underinference Within Task: All-Contingency

Notes. Each point represents the estimated degree of underinference for the stronger (x-axis) and the weaker signal (y-axis), measured as the average ratio of the reported log posterior-odds to the log signal strength for each individual in a given task in All-Contingency; for symmetric SGPs, signals are equally diagnostic and we classified the blue ball as stronger arbitrarily.

A.3 Additional Measures

A.3.1 Degree of Asymmetry: Alternative Measure

This section proposes an alternative measure of degree of asymmetry and replicates the findings in the main text using this measure.

An alternative measure of continuous degree of asymmetry for an SGP can be quantified by the absolute distance between the posterior across signals and bags:

$$|\Pr(A|s) - \Pr(B|s')|.$$

This difference is always zero for symmetric SGPs, and positive for asymmetric SGPs. The larger the difference in posteriors the more dissimilar the Bayesian posteriors.

The two measures we consider to explore the role of degree of asymmetry are conceptually similar, but not the same. The main difference is that this measure captures also the role of the prior over states, even if this does not play a relevant role since we consider equal priors over states. Another difference is that the measure used in the text better captures relative differences rather than absolute ones.

Table A5: Bias and Underinference with Alternative Measure of Asymmetry

	I	II
All-Contingency	0.010	0.012
	(0.010)	(0.068)
One-Contingency	0.041***	0.044
	(0.011)	(0.089)
Alternative Degree of Asymmetry	0.098*	2.756***
	(0.041)	(0.554)
All-Contingency × Alternative Degree of Asymmetry	0.154**	-0.336
	(0.055)	(0.715)
One-Contingency × Alternative Degree of Asymmetry	-0.006	-0.435
	(0.063)	(0.921)
Log Signal Strength		0.886***
		(0.038)
Log Signal Strength × All-Contingency		-0.023
		(0.052)
Log Signal Strength × One-Contingency		-0.152*
		(0.069)
Log Signal Strength × Alternative Degree of Asymmetry		-1.268***
		(0.354)
Log Signal Strength × All-Contingency × Alternative Degree of Asymmetry		-0.408
		(0.441)
Log Signal Strength × One-Contingency × Alternative Degree of Asymmetry		0.218
		(0.592)
Constant	0.064***	-0.067
	(0.006)	(0.054)
N	6000	6000
adj. R^2	0.020	0.256
Clusters	450	450

Notes. OLS estimates. Individual-level clustered standard errors. The dependent variable is defined as the absolute value of the difference between the posterior reported by participants and the normative (Bayesian) benchmark in Column I and as the ratio of the log posterior-odds and log signal strength for a given signal in Column II; * p<.05, ** p<.01, *** p<.001.

A.3.2 Within-Consistency

To construct the within-consistency measure in our dataset, we proceed as follows.

First, for each pair of mirrored SGPs, all posteriors were reported in terms of one SGP (15-85 for symmetric and 30-95 for asymmetric). Second, we keep only the observation for which we can construct this measure. In *Conditional* and *One-Contingency*, the desired measure could only be constructed if the participant's posterior was elicited for the *same* signal for both mirrored SGPs (approximately in half of all cases, for each color of the ball). In *All-Contingency*, participants' beliefs are always elicited conditional on both signals for each SGP. Therefore, we keep 156 and 148 observations, respectively, in *Conditional* and in *One-Contingency*, and 600 in *All-Contingency*. Third, we calculate the difference between the posteriors conditional on the same signal. For any signal s and for any two mirrored SGPs M1 and M2, the dependant variable is defined as

$$\Delta\text{Posteriors} = |\Pr^{M1}(A|s) - \Pr^{M2}(A|s)|.$$

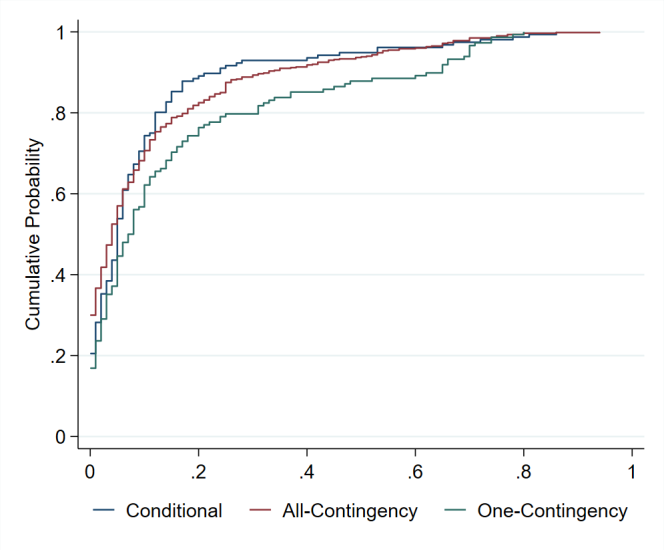


Figure A9: Cumulative Distribution of Δ Posteriors

Notes. Cumulative distribution function of the measure of within-consistency Δ Posterior by treatment.

A.3.3 Between-Consistency

To construct the between-consistency measure, we look at vectors of posteriors, that is, the reported posteriors conditional on both signal realizations: $(\Pr(A|blue), \Pr(A|orange))$. Given the method of belief elicitation, these are available for all SGPs in *All-Contingency*. For *Conditional* and *One-Contingency*, we construct the vectors of posteriors exploiting the mirrored SGPs as follows.

First, for each pair of mirrored SGPs, all posteriors were reported in terms of one SGP (15-85 for symmetric and 30-95 for asymmetric). This part overlaps with the construction of Δ Posteriors. Then, we keep only the observations of the participants whose posteriors were elicited conditional on the *different* signal realizations for the mirrored SGPs (around half of the times, for each color of the ball). Therefore, we have 144 and 152 observations, respectively, in *Conditional* and in *One-Contingency*, and 600 in *All-Contingency*.

Distance from Bayesian Vector of Posteriors We complement our between-consistency analysis by examining a more nuanced measure. Next, we calculate the squared distance between the reported vector of posteriors and the Bayesian one:

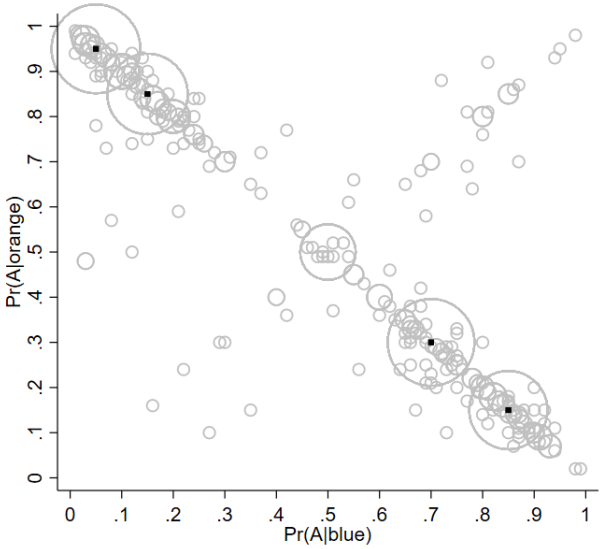
$$\text{Distance} = \sqrt{\Big(\Pr(A|blue) - \hat{\Pr}(A|blue)\Big)^2 + \Big(\Pr(A|orange) - \hat{\Pr}(A|orange)\Big)^2}.$$

The regression results presented in Table A6 align with the findings of Section 4.3.2.

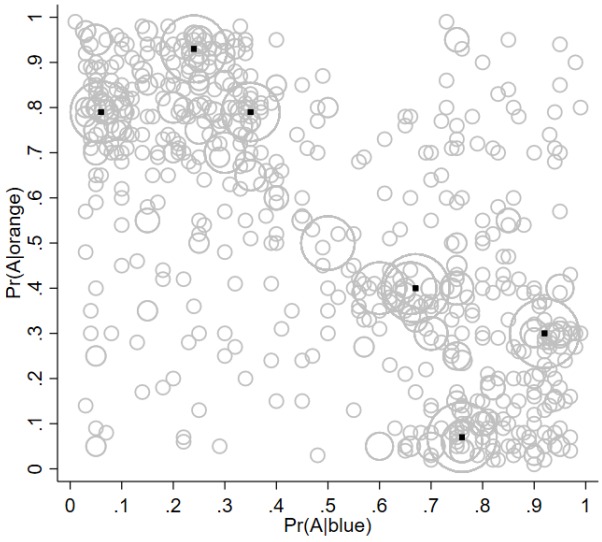
Table A6: Distance

	I	II
All-Contingency	0.022	0.009
	(0.019)	(0.014)
One-Contingency	0.067**	0.042*
	(0.025)	(0.021)
Constant	0.149***	0.116***
	(0.015)	(0.011)
N	896	817
adj. R^2	0.016	0.013
Clusters	375	360

Notes. OLS estimates. Individual-level clustered standard errors. SGP symmetry fixed effects. The dependent variable is defined as the squared distance between the reported posterior and the normative (Bayesian) benchmark. Column I includes all samples of vectors of posteriors, while Column II excludes vectors of posteriors that are not Bayes-consistent. * p<.05, ** p<.01, *** p<.001.



(a) Symmetric SGPs



(b) Asymmetric SGPs

Figure A10: Vector of Posteriors: All-Contingency

Notes. Each circle represents the frequency of the corresponding vector of posterior beliefs in a given task across SGPs: the bigger the circle, the higher the frequency. Each black square corresponds to the Bayesian vector of posteriors associated with an SGP.

B Appendix: Expert Survey

B.1 Survey Design & Data Collection

Our expert survey has three parts. First, we provide all relevant information on the experiment. The survey began with a short description of the goal of the study for which participants were asked to report predictions. After consenting to participate in our survey, we clarified that the experiment was already preregistered but not run yet; we informed the experts that the preregistration link was available at the end of the survey. Then, they read a detailed description of our experimental design. To keep the survey brief and focused on our main objective, we only describe two treatments: *Conditional* and *All-Contingency*. The survey participants could access further details on the design in linked documents, such as the instructions and control questions of these two treatments and information on the used SGPs. We also include information about the target sample, randomization, and incentives. Finally, we highlight as the key outcome of interest the bias as defined in Section 4.

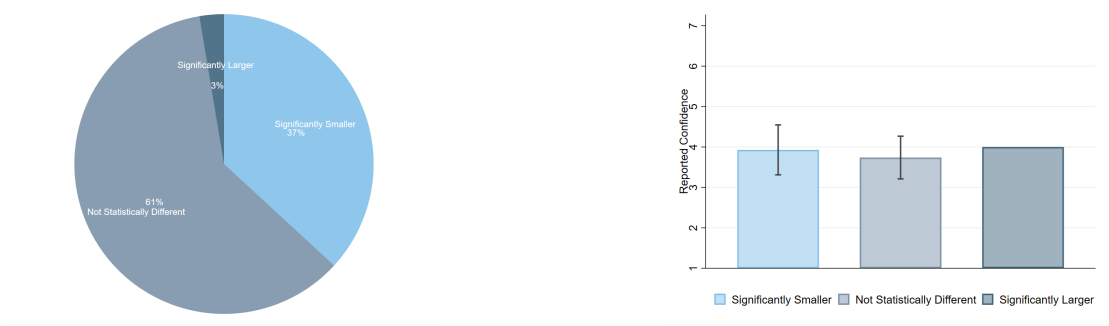
In the second part, we elicited the experts' predictions. This was followed by two sets of questions. First, we elicited the expected direction of the treatment effect: the participants reported whether they expected the bias in *Conditional* to be significantly smaller, higher, or not statistically significant than in *All-Contingency*. The participants also reported their confidence (1-7 scale) in their answers. Second, we elicited the participants' opinions on the heterogeneity of the treatment effect along two dimensions: CRT and the symmetry of SGPs. Also, for this set of questions, the participants reported their confidence in their previous answers (1-7 scale). Finally, the participants were asked how they classify their research (theoretical, experimental, and/or empirical). The pre-registration link was also available on the final screen.

The Qualtrics survey was distributed in February 2023 using the Social Science Prediction Platform (Study ID: sspp-2023-0007-v1) by invitation (the survey was not publicly accessible). We compiled a distribution list including researchers that we considered knowledgeable about topics related to expectations or contingent thinking for a total of 135 experts. We purposefully excluded colleagues who were aware of pilot results through conversations with us.

B.2 Predictions

Sample In total, we gathered 38 responses (28% completion rate). Our final sample includes 17 faculty members, 6 postdocs, and 12 PhD students (with 3 participants not reporting their position). 89% described their research as experimental, 29% as theoretical, and 26% as empirical (these categories were not mutually exclusive). 83% include behavioral economics as one of their main fields; other fields include experimental economics, microeconomics theory, game theory, development economics, and political economics, among others.

Main Prediction Figure B1a illustrates how experts expect the bias in *Conditional* to change compared to *All-Contingency*. Compared to Conditional, 14 participants predicted a significantly smaller bias in *All-Contingency*, and only one predicted a significantly higher bias in *All-Contingency*. 23 experts predicted no significant difference between *Conditional* and *All-Contingency*. These percentages do not vary much depending on the research field. Also, there does not seem to be a difference in confidence in the expected direction of the treatment, as shown in Figure B1b.



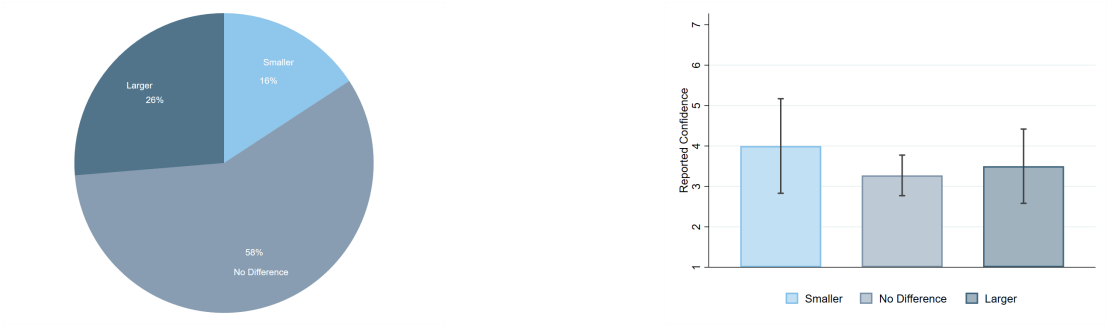
(a) Direction (b) Confidence

Figure B1: Main Prediction

Notes. Panel (a) shows the shares of experts predicting a significantly higher, significantly lower, and no significantly different bias in *All-Contingency* compared to *Conditional*. Panel (b) shows for each possible prediction the confidence of the experts in their answers on a Likert scale (1-7).

Heterogeneous Effect of SGP Symmetry In Figure B2a, we report the expectations of the change in the bias for symmetric SGPs compared to the change for asymmetric SGPs. 58% predicted no significant difference in the change in the bias between asymmetric and symmetric SGPs. 26% expects a significantly higher change in the bias and 16% expects a significantly lower change in the bias for asymmetric SGPs compared to symmetric SGPs. The predictions do not seem different by the expected treatment effect (Figure B3).

Heterogeneous Effect of CRT Figure B4a summarizes how participants expect the change in bias for individuals who score low on the CRT to vary compared to individuals who score high on the CRT. 55% predicted no significant difference in the change in the bias between individuals who scored low and high on the CRT. 29% expect a significantly smaller change in the bias, and 16% expect a higher change in the bias for individuals with high CRT scores compared to individuals with low CRT scores. The predictions do not seem different from the expected treatment effect (Figure B5).



(a) Direction (b) Confidence

Figure B2: Prediction about SGP Symmetry

Notes. Panel (a) shows the shares of experts predicting a significantly higher change in the bias, a significantly lower change in the bias, and no significantly different change in the bias for asymmetric compared to symmetric SGPs. Panel (b) shows for each possible prediction the confidence of the experts in their answers on a Likert scale (1-7).

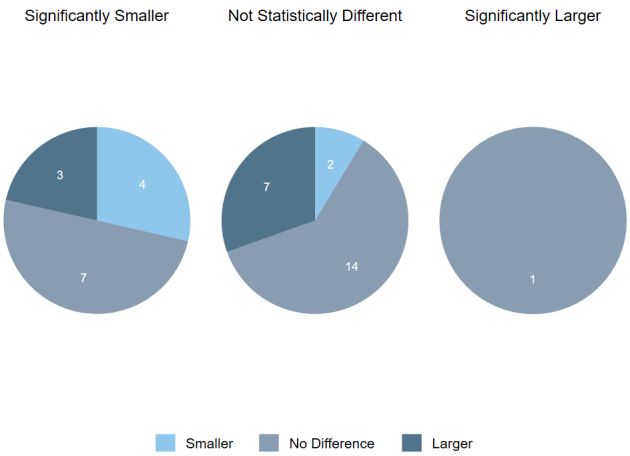
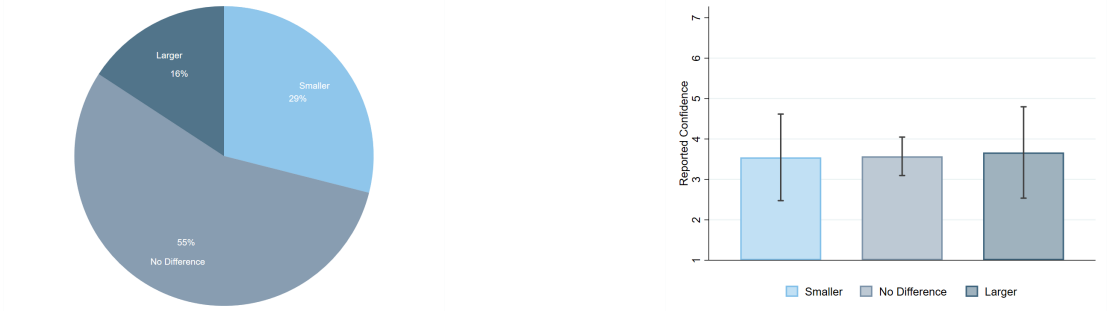


Figure B3: Prediction about SGP Symmetry, by Expected Treatment Effect

Notes. Shares of experts predicting a significantly higher change in the bias, a significantly lower change in the bias, and no significantly different change in the bias for asymmetric compared to symmetric SGPs by possible answers on the expected treatment effect.



(a) Direction (b) Confidence

Figure B4: Prediction about CRT

Notes. Panel (a) shows the shares of experts predicting a significantly higher change in the bias, a significantly lower change in the bias, and no significantly different change in the bias for individuals with high compared to low CRT. Panel (b) shows for each possible prediction the confidence of the experts in their answers on a Likert scale (1-7).

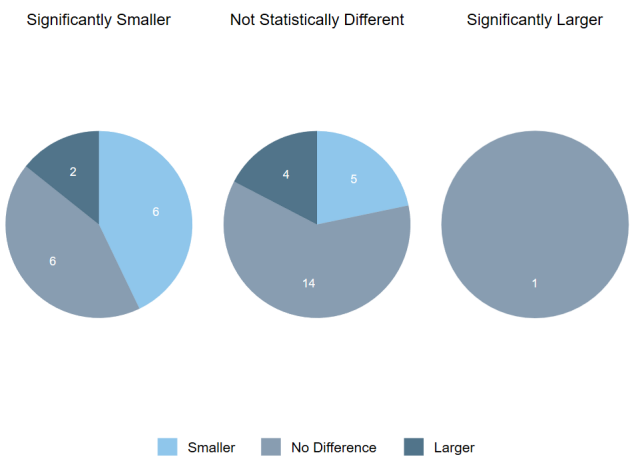


Figure B5: Prediction about CRT, by Expected Treatment Effect

Notes. Shares of experts predicting a significantly higher change in the bias, a significantly lower change in the bias, and no significantly different change in the bias for individuals with high compared to low CRT by possible answers on the expected treatment effect.

C Appendix: Experimental Instructions & Interface

C.1 Instructions

C.1.1 General Instructions

WELCOME!

Thank you for participating in this study. You are guaranteed to receive GBP 2 for completing the study. If you follow the instructions carefully, you may earn an additional bonus of GBP 2, as explained later. Your earnings will depend on your decisions and chance.

Please read the instructions carefully. **There will be two checks of your understanding of these instructions, for which you have three attempts**. If you provide three incorrect answers in either set of these questions about the instructions, you will not be eligible for a bonus payment. You always have to complete the study to receive the guaranteed payment.

There will be two parts to the experiment. The first part is the main part of the experiment and will take up most of the time. The second part will be introduced after you have finished the first part. In total, this study should take around 30 minutes.

PART ONE

In the first part, you will be asked to make a series of choices that can impact your bonus payment. Most of these choices will be of a similar format: You have to guess the chance that a bag was selected based on the available information.

In each task, you are asked to consider two bags, bag A and bag B. In each bag, there are several balls. The total number of balls can be either 60 or 80. The balls are either orange or blue. The number of orange and blue balls in each bag varies across tasks. You will be informed about the number of orange and the number of blue balls in each bag.

The task proceeds as follows:

- You start by clicking 'Select the bag'.
- The computer randomly flips a **fair coin** to select bag A or bag B. It is **equally likely** that the computer selects bag A or bag B.
- You do not know whether bag A or bag B was selected.
- When you click 'Draw the ball', the computer draws either an orange ball or a or blue ball from the selected bag.
- The computer draws one ball from the selected bag.

Your task is to guess the chance (in %) that the computer chose bag A or bag B.

You will repeat this task ten times. For each task, the computer selects a new bag and then draws a new ball from the selected bag. **So you should think about which bag was selected in each task independently of all other tasks.**

C.1.2 Control Questions 1

TESTING YOUR UNDERSTANDING OF THE INSTRUCTIONS

On the slider, please indicate the chance (in %) that a fair coin flip selects bag A.

Chance of **bag A (in %): Please click on the slider**

When you click 'Draw the ball', the computer draws a ball from the previously selected bag.

True

False

When you click 'Draw the ball', you know which bag was previously selected.

True

False

C.1.3 Conditional Instructions

YOUR CHOICE

The computer draws either an orange ball or a blue ball. You observe the color of the ball. **You will be asked to guess the chance that the ball was drawn from bag A or bag B.**

You make your guess by selecting the chance between 0% and 100%. Higher numbers mean that you think it is more likely that this bag was selected. The guess for the chance that the ball was drawn from bag A and the chance that the ball was drawn from bag B will automatically sum up to 100%.

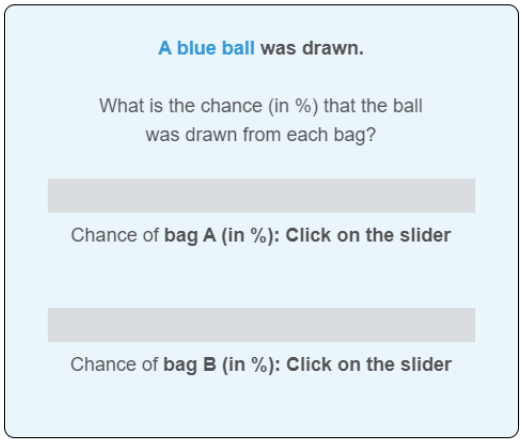
YOUR CHOICE: EXAMPLE

This is an example of the task. It is not relevant for your payment. Please familiarize yourself with the interface, then proceed with the instructions. **You can hover over the elements of the screen to see the explanations of each part of the screen.**

Remember:

Bag A contains **60 blue balls** and **40 orange balls**.
Bag B contains **40 blue balls** and **60 orange balls**.

Make your guesses below.



PAYMENT

For your bonus payment, one of the ten tasks will be randomly selected for payment. Your bonus payment will depend on your guesses in the selected task. Your guesses do not influence which task is selected for payment.

We have carefully chosen the payment rule such that you maximize the chance of winning a bonus of GBP 2 if you give your best guesses in all questions. To maximize the chance of winning the bonus, **it is in your best interest to always give a guess that you think is the true chance. The closer your guess is to the true chance, the higher is your probability of receiving the bonus.** If you are interested, further details on the payment are provided here.

▶ **Click here for further details**

C.1.4 All-Contingency Instructions

YOUR CHOICE

The computer draws either an orange ball (case orange) or a blue ball (case blue). You do not observe the color of the ball when making your guesses.

For each of the two possible cases (orange and blue), **you will be asked to guess the chance that the ball was drawn from bag A or bag B.**

For each case, **you make your guess by selecting the chance between 0% and 100%.** Higher numbers mean that you think it is more likely that this bag was selected. The guesses for the chance that the ball was drawn from bag A and the chance that the ball was drawn from bag B will automatically sum up to 100%.

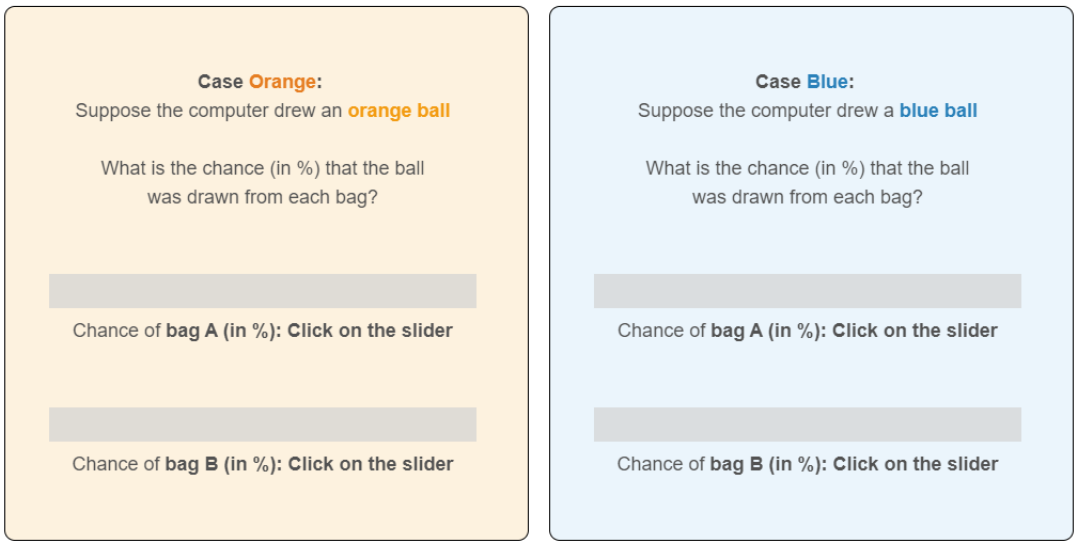
YOUR CHOICE: EXAMPLE

This is an example of the task. It is not relevant for your payment. Please familiarize yourself with the interface, then proceed with the instructions. **You can hover over the elements of the screen to see the explanations of each part of the screen.**

Remember:

Bag A contains **60 blue balls** and **40 orange balls**.
Bag B contains **40 blue balls** and **60 orange balls**.

Make your guesses below for **Case Blue** and **Case Orange**.



PAYMENT

For your bonus payment, one of the ten tasks will be randomly selected for payment. Your bonus payment will depend on your guesses in the selected task. Your guesses do not influence which task is selected for payment.

We have carefully chosen the payment rule such that you maximize the chance of winning a bonus of GBP 2 if you give your best guesses in all questions. To maximize the chance of winning the bonus, **it is in your best interest to always give a guess that you think is the true chance. The closer your guess is to the true chance, the higher is your probability of receiving the bonus.** If you are interested, further details on the payment are provided here.

▶ **Click here for further details**

You are asked about your guesses for case orange and case blue. Depending on the color of the ball drawn from the bag, only your guesses for that case will matter for your bonus payment. As you do not know the color of the ball when making your guess, it is therefore in **your best interest to give your best guesses for each case**.

As an example, imagine that the computer draws a blue ball. Then, only your guesses for case blue matter for your bonus payment.

C.1.5 One-Contingency Instructions

YOUR CHOICE

The computer draws either an orange ball (case orange) or a blue ball (case blue). You do not observe the color of the ball when making your guesses.

You will be asked to guess the chance that the ball was drawn from bag A or bag B for one of the two possible cases (orange or blue). It is equally likely that you will be asked about each case. This does not depend on the actual color of the ball drawn by the computer.

You make your guess by selecting the chance between 0% and 100%. Higher numbers mean that you think it is more likely that this bag was selected. The guess for the chance that the ball was drawn from bag A and the chance that the ball was drawn from bag B will automatically sum up to 100%.

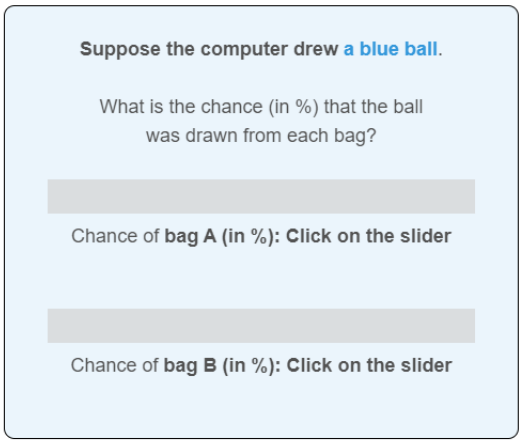
YOUR CHOICE: EXAMPLE

This is an example of the task. It is not relevant for your payment. Please familiarize yourself with the interface, then proceed with the instructions. **You can hover over the elements of the screen to see the explanations of each part of the screen.**

Remember:

Bag A contains **60 blue balls** and **40 orange balls**.
Bag B contains **40 blue balls** and **60 orange balls**.

Make your guesses below.



PAYMENT

For your bonus payment, one of the ten tasks will be randomly selected for payment. Your bonus payment will depend on your guesses in the selected task. Your guesses do not influence which task is selected for payment.

We have carefully chosen the payment rule such that you maximize the chance of winning a bonus of GBP 2 if you give your best guesses in all questions. To maximize the chance of winning the bonus, **it is in your best interest to always give a guess that you think is the true chance. The closer your guess is to the true chance, the higher is your probability of receiving the bonus.** If you are interested, further details on the payment are provided here.

▶ **Click here for further details**

You are asked about your guess for one case, either case orange or case blue. If the color of the ball drawn from the bag matches the case you considered **your guess matters for your bonus.** Otherwise, you will receive a fixed payment of GBP 1. As you do not know the color of the ball when making your guess, it is therefore in **your best interest to give your best guess.**

As an example, imagine that you are asked about case orange. If an orange ball was drawn, your guess matters for the bonus payment. If a blue ball was drawn, you receive the fixed payment.

C.1.6 Control Questions 2

TESTING YOUR UNDERSTANDING OF THE INSTRUCTIONS

The bonus payment will be implemented for one randomly selected task.

- True
- False

It is in your best interest to give your best guess of the chance that the ball was drawn from bag A or bag B.

- True
- False

We will ask you about the guess of the chance that the ball was drawn from bag A or bag B

- before you get to know the color of the ball.
- once you get to know the color of the ball.

C.2 Task Interface

C.2.1 Conditional

Part One: Task 1/10

Please click on the right arrow if you are ready to proceed to the next task.

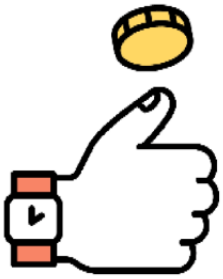
→

Bag A contains **9 orange balls** and **51 blue balls**.
Bag B contains **51 orange balls** and **9 blue balls**.

Next:
The computer randomly selects one bag by flipping a fair coin.

Select the bag

Flipping the coin to select the bag...



The coin was flipped and a bag was selected.

Next:
The computer will draw a ball from the bag that was previously selected.

Draw the ball

The computer draws a random ball from the bag that was previously selected...

Remember:

Bag A contains **9 orange balls** and **51 blue balls**.
Bag B contains **51 orange balls** and **9 blue balls**.

Make your guesses below.

A blue ball was drawn.

What is the chance (in %) that the ball was drawn from each bag?

Chance of **bag A (in %): Click on the slider**

Chance of **bag B (in %): Click on the slider**

C.2.2 All-Contingency

Part One: Task 1/10

Please click on the right arrow if you are ready to proceed to the next task.

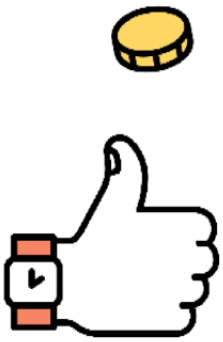
→

Bag A contains **42 orange balls** and **18 blue balls**.
Bag B contains **3 orange balls** and **57 blue balls**.

Next:
The computer randomly selects one bag by flipping a fair coin.

Select the bag

Flipping the coin to select the bag...

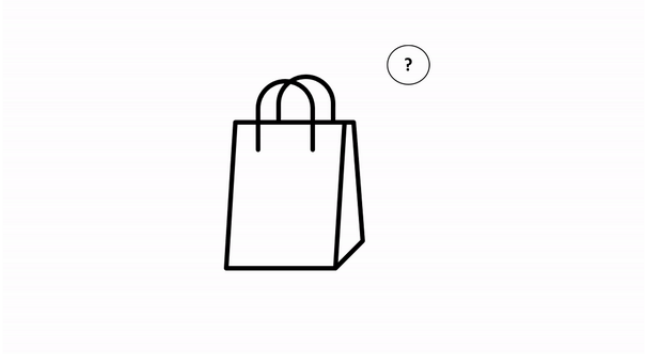


The coin was flipped and a bag was selected.

Next:
The computer will draw a ball from the bag that was previously selected.

Draw the ball

The computer draws a random ball from the bag that was previously selected...



Remember:

Bag A contains **42 orange balls** and **18 blue balls**.
Bag B contains **3 orange balls** and **57 blue balls**.

Make your guesses below for **Case Blue** and **Case Orange**.

Case Orange:
Suppose the computer drew an orange ball.

What is the chance (in %) that the ball was drawn from each bag?

Chance of bag A (in %): Click on the slider

Chance of bag B (in %): Click on the slider

Case Blue:
Suppose the computer drew a blue ball.

What is the chance (in %) that the ball was drawn from each bag?

Chance of bag A (in %): Click on the slider

Chance of bag B (in %): Click on the slider

An orange ball was drawn.

→

C.2.3 One-Contingency

Part One: Task 1/10

Please click on the right arrow if you are ready to proceed to the next task.

→

Bag A contains **68 orange balls** and **12 blue balls**.
Bag B contains **12 orange balls** and **68 blue balls**.

Next:
The computer randomly selects one bag by flipping a fair coin.

Select the bag

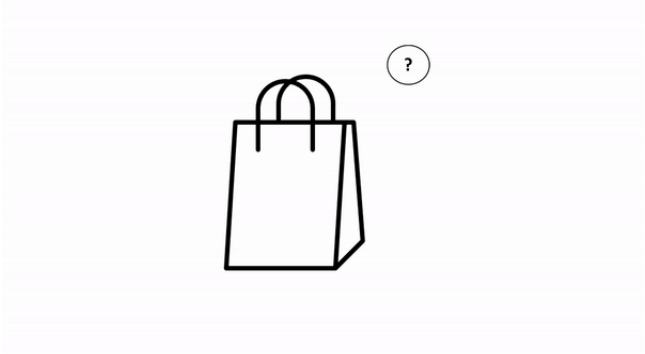
Flipping the coin to select the bag...

The coin was flipped and a bag was selected.

Next:
The computer will draw a ball from the bag that was previously selected.

Draw the ball

The computer draws a random ball from the bag that was previously selected...



Remember:

Bag A contains **68 orange balls** and **12 blue balls**.
Bag B contains **12 orange balls** and **68 blue balls**.

Make your guesses below.

Suppose the computer drew an orange ball.

What is the chance (in %) that the ball was drawn from each bag?

Chance of **bag A (in %): Click on the slider**

Chance of **bag B (in %): Click on the slider**

A blue ball was drawn.

→

C.3 Modified Cognitive Reflection Test

We modified the original version of the Cognitive reflection test (Frederick, 2005) to avoid previous experiences or cheating, asking the following three questions.

1. Milk and a cookie cost GBP 3.20 in total. Milk costs GBP 2 more than the cookie. How much does the cookie cost?
2. If it takes 50 workers 50 minutes to pick 50 apples, how long would it take 1000 workers to pick 1000 apples?
3. A runner doubles the number of kilometers he runs every month. After one year, he runs a marathon, 42 km. After how many months did he run a half marathon, 21 km?