How People Use Statistics

Pedro Bordalo John Conlon Nicola Gennaioli
Spencer Kwon Andrei Shleifer[1]

June 24, 2023

Abstract

We document two new facts about the distributions of answers in famous statistical problems: they are i) multi-modal and ii) unstable with respect to irrelevant changes in the problem. We offer a model in which, when solving a problem, people represent each hypothesis by attending "bottom up" to its salient features while neglecting other, potentially more relevant, ones. Only the statistics associated with salient features are used, others are neglected. The model unifies Gambler's Fallacy, its variation by sample size, under- and overreaction in inference, and insensitivity to multiple signals, all as a byproduct of selective attention. The model also makes new predictions on how controlled changes in the salience of specific features should jointly shape measured attention and biases. We test and confirm these predictions experimentally, including by measuring attention and documenting novel biases predicted by the model. Bottom-up attention to features emerges as a unifying framework for biases conventionally explained using a variety of stable heuristics or distortions of the Bayes rule.

[1] Saïd Business School, University of Oxford, Harvard University, Bocconi University and IGIER, Harvard University, and Harvard University. We are grateful to Ben Enke, Thomas Graeber, Alex Imas, Jesse Shapiro, and Josh Schwartzstein for helpful comments.

1. Introduction

Some of the most glaring judgment biases arise in statistical problems. When assessing flips of a fair coin, people tend to estimate a balanced sequence such as $hthtth$ to be more likely than $hhhhhh$. This striking phenomenon, called the Gambler's Fallacy, arises even though people *know* that each toss lands heads or tails with 50% probability, which implies that the two sequences are equally likely. People also make errors when updating beliefs based on a noisy signal. They underreact to the signal in some problems (Edwards 1968), but overreact in others (Kahneman and Tversky 1972). This is also striking: in these problems people *are told* numerical priors and likelihoods, and could compute the correct answer using the Bayes' rule.

Why do people make these systematic mistakes? And why are these mistakes unstable, changing from one problem to the next and across different versions of the same problem? To date, there is no answer to these questions. A large body of work formalizes specific biases such as the Gambler's Fallacy (Rabin 2002), sample size neglect in coin flips (Benjamin, Rabin, Raymond 2016), as well as base rate neglect (Grether 1980) and underreaction in inference (Enke and Graeber 2023), but does not connect biases across problems or in different versions of a problem.

We address these questions by proposing a new approach in which selective attention to some, but not all, features of a problem shapes how a decision maker represents and solves it. To see the idea, consider the famous duck-rabbit illusion, in which a drawing of an animal can be interpreted as either a duck or a rabbit. Some people attend to the beak and see a duck, others attend to the mouth and see a rabbit. One feature is attended to, the other neglected, so different people see a different animal. More controversially, some people see abortion as a human right, while others see it as a murder, but rarely both. Or consider sentencing a confessed bank robber (Clancy et al. 1981). Some judges focus on the defendant's age, others on whether he was armed, still others on his criminal record, leading to different sentences for the same crime under the same law. In bail decisions, some judges may even focus on hardly relevant aspects, such as whether a defendant is well groomed

(Ludwig and Mullainathan 2023). In all these examples, decision makers attend to some but not all features of the problem they face, leading to different representations and assessments.

We argue that the same logic is at play when people solve statistical problems, except here there is a correct answer. These problems are also characterized by many features, including the statistics that people are given. Depending on the hypotheses, only some features are relevant to correctly solve a problem. Critically, our decision maker, or DM, simplifies hypotheses "bottom up", by attending only to the features that are most salient to her. This causes her to neglect some relevant features, leading to error. For example, the DM may represent a sequence of coin flips by attending to each of its flips. Alternatively, she may attend to other salient features of the sequence: when judging the likelihood of $hthtth$ vs. $hhhhhh$ the DM may for instance find it striking that, the coin being fair, one sequence has no tails. As her attention is drawn to the share of heads, she may neglect individual flips, and represent $hthtth$ as a generic balanced sequence. She thus replaces the original question with the relative likelihood of obtaining a balanced sequence vs a sequence of all heads, leading to the Gambler's Fallacy. Bottom-up attention to features can generate a faulty representation of the problem, which leads to bias.

But what are the salient features in statistical problems? Psychologists have unveiled several drivers of salience. Following our prior work (Bordalo et al. 2022) we formalize two of them: contrast and prominence. In consumer choice, contrast means that the "price feature" is salient when it strikingly favors one good over another. Analogously, in a statistical problem a feature has high contrast if it strikingly favors one of two hypotheses. This is why when comparing $hthtth$ to $hhhhhh$, the share of heads is salient: obtaining a balanced sequence (considered as a set) is much more likely than obtaining an unbalanced one. Contrast depends on objective probabilities, so the model's predictions can be tested using controlled changes in the statistics of the problem. The second driver of salience, prominence, depends on non-statistical features of problems, for example on what jumps out visually or is otherwise top of mind. In consumer choice, making price or taxes more noticeable (Chetty et al. 2009) or cueing the high price of beer paid at a resort (Thaler 1985, Bordalo

et al. 2013) render the price feature salient. Analogously, in a statistical problem, a feature is salient if it is more visible, explicitly mentioned, or if hypotheses are cast in a context where this feature has been experienced as relevant. We do not measure changes in prominence directly, but its effects entail joint predictions on estimates and attention. In coin flips, making individual flips more prominent should reduce both the Gambler's Fallacy *and* measured attention to the share of heads.

We show that our approach unifies the Gambler's Fallacy and the insensitivity to sample size in judgments about i.i.d. draws, as well as overreaction, underreaction, and insensitivity to the weight of evidence (Griffin and Tversky 1992) in inference. It also yields three new predictions, which we test experimentally. First, beliefs often neglect relevant information. In inference problems, the features of the population or those of the signal may be salient, causing people to use either the base rate or the likelihood, but rarely both. Second, the distribution of answers in a given problem is multi-modal: different features are salient to different people, due to random variation in attention or to past experiences, who then use different statistics. We document empirically multimodality in both attention and estimates, with modes corresponding to partial use of information as reflected in the fact that many subjects anchor exactly to the base rate or to likelihood (see also Dohmen et al. 2009).

Our third and key prediction is that changes in the salience of a feature cause joint shifts in attention and in the distribution of answers. Guided by the model, we perform several experimental tests where we manipulate a feature's salience and generate shifts in both beliefs and measured attention. Consider inference problems. First, our model predicts that increasing the likelihood should raise the contrast induced by the signal and jointly boost attention to the signal *and* anchoring to the likelihood by neglecting the base rate. The data confirm this prediction. Second, our model predicts that describing more prominently the accuracy of the signal should also increase measured attention and anchoring to the likelihood. We show that these two mechanisms account almost fully for the dramatic shift in assessments from the balls and urns format (Edwards 1968), in which many people anchor to the base rate, to the formally identical "taxicabs" format (Kahneman and Tversky 1972), in which many people anchor to the likelihood. In addition to features, we also manipulate the salience

of a hypothesis by not mentioning its alternative in the question. This manipulation causes many subjects to estimate the prominent hypothesis by the product of its base rate and likelihood, which entails full neglect of the features of the other hypothesis. We show that such bottom-up neglect of a non-salient hypothesis throws new light on confirmation bias and casts doubt on the notion that human intuition is generally ecologically optimal (Gigerenzer and Hoffrage 1995).

Overall, a basic cognitive function, bottom-up attention to features, unifies biases typically attributed to heuristics such as availability, representativeness, or anchoring (Kahneman and Tversky 1972, Gigerenzer 1996). This mechanism yields a form of question substitution (Kahneman Frederick 2002) that explains why different people seem to be using different heuristics, why some observed biases are not associated with known heuristics, and why a specific bias becomes more prevalent when a feature becomes salient. The mechanism also highlights the importance of going beyond eliciting judgments, and directly measuring attention. We explore several ways to do so and find that they align with each other as well as with the model predictions.

Our approach connects to simplification and categorization by feature reduction, which is central in machine learning (Selfridge 1955, Guyon and Elisseeff 2003), neuroscience (Behrens et al 2018), and psychology (Kruschke 2008, Evers et al., 2021). In economics, neglect of information is often explained by rational inattention (Sims 2003, Woodford 2003, Woodford 2020, Gabaix 2019), or computational complexity (Enke, Graeber, and Oprea 2023). In our model, simplification is also central, but it occurs via bottom up attention to features, which explains why manipulating the salience of irrelevant features can strongly change biases. Rational inattention to statistics does not yield these framing effects.

The paper proceeds as follows. Section 2 presents new evidence that the distribution of answers in coin flip and inference problems is concentrated at specific modes, whose incidence changes with normatively irrelevant modifications. This evidence motivates our new approach. Section 3 introduces our model. Sections 4 and 5 develop and evaluate empirical predictions for coin flips and inference. Section 6 derives and tests other implications. Section 7 concludes.

2. Puzzles in famous statistical problems

In April 2023, we recruited participants online through Prolific to answer one "Gambler's Fallacy" problem and one "Inference" problem, in a random order at the beginning of the survey. They earned an additional bonus for each question if their answers were within 5 percentage points of the correct ones. Appendix A describes the experimental protocol and pre-registration.

In coin flips treatments, we told participants that we repeatedly flipped a fair coin n times. In 100 of these n-flip "sequences", the order of heads vs tails was H_1 or H_2. Participants were asked their best guess of how many were of each type. Panel A of Figure 1 reports the distribution of answers for $H_1 = hh$, $H_2 = th$, and Panel B the distribution for $H_1 = hhhhhh$, $H_2 = ththht$.

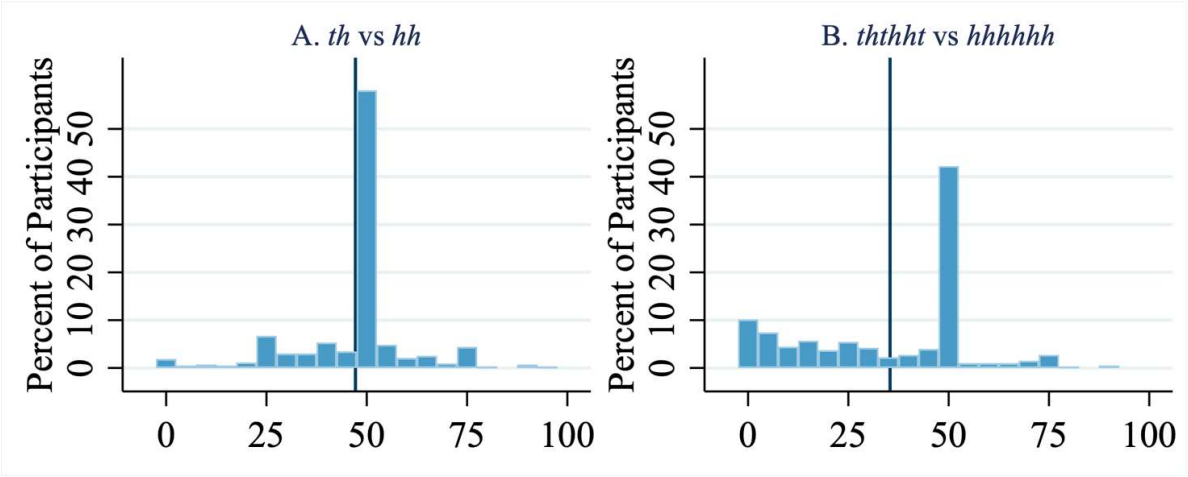

Figure 1. Each panel reports the distribution of estimated $\Pr(H_1|H_1 \cup H_2)$. Answers closer to 0 indicate higher probability of the balanced sequence H_2. The blue bar marks the mean answer.

As in previous studies (Benjamin 2019), the mean response is below 0.5, confirming the Gambler's Fallacy, the belief that a specific balanced sequence is more likely than an unbalanced one. There are, however, two new findings. First, the Gambler's Fallacy is much more severe when $n = 6$: the average probability estimate of H_1 drops from 47.2% in Panel A to 35.4% in Panel B ($p = 0.00$). Second, this occurs in part because about 14% of respondents *shift* from the 50% mode to answers around 5% (54.8% in panel A vs. 40.7% in panel B, $p = 0.00$). This instability is inconsistent with a mechanical, possibly heterogeneous, tendency to display the Gambler's Fallacy (Rabin 2002). It seems that when judging short sequences, many people attend to the fact that each

flip has a $50{:}\,50$ chance of h and t, but neglect this feature when the sequences are long. Why are different features neglected in the two experiments, where the correct answer is the same?

Consider next inference problems. In the "balls and urns" paradigm (Edwards 1968), urn A contains 80% green and 20% blue balls, urn B contains 20% green and 80% blue balls. A computer selects urn A or B with probabilities 25% and 75% respectively, and draws a ball from it. The ball is green. What is the probability that it was drawn from A vs. B? Kahneman and Tversky (1972) frame inference in a naturalistic context. There are two taxicab companies, the Blue and the Green, according to the color of the cabs they run. 25% of the cabs are Green, 75% are Blue. A cab is involved in a hit and run accident, and a witness reports the cab as Green. A test reveals that the witness can correctly identify each color cab with probability 80%. What is the probability that the errant cab is indeed Green vs. Blue?

In balls and urns formats, the mean answer exhibits underreaction (Benjamin 2019). In taxicabs and other so called "base rate problems" (Esponda et al. 2022), it exhibits overreaction. We run the two formats with identical statistical parameters with two sets of participants, which to our knowledge has not been done before. Using the Bayes rule, the correct answer is $\Pr(A|green) = \Pr(Green|green) = 0.57$ in both problems. The distribution of answers is reported in Figure 2.

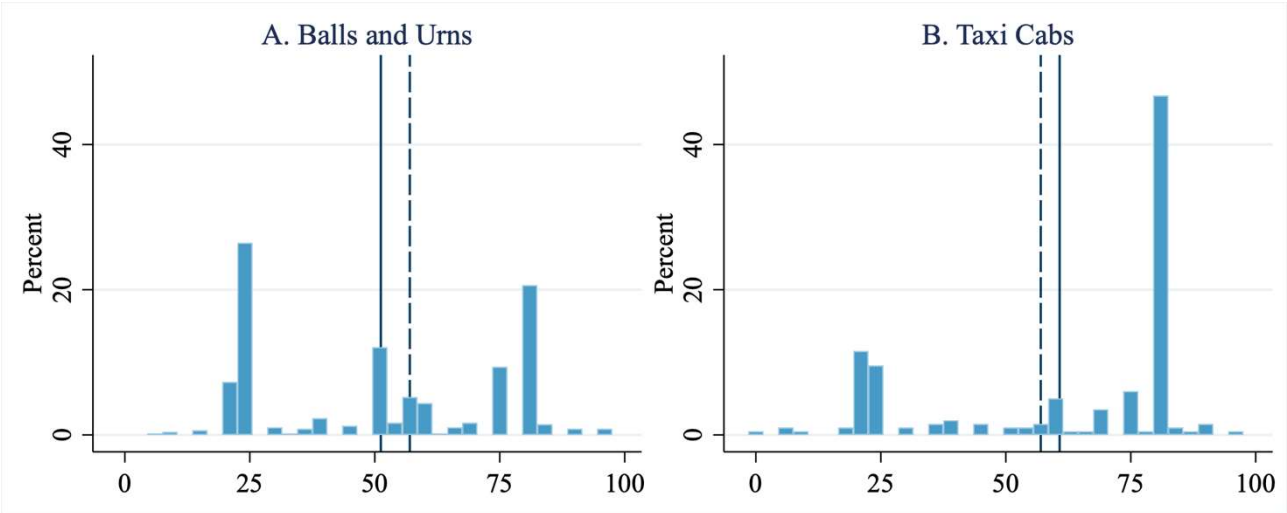

Figure 2. The left panel reports the distribution of $\Pr(A|g)$, the right panel of $\Pr(G|g)$. The solid line indicates the mean answer, while the dashed line indicates the Bayesian answer of 0.57.

Consistent with previous work, in balls and urns (Panel A) under-reaction to the data prevails on average: the mean answer (solid line) is 52%, lower than the correct answer (dashed line). There is however pronounced multi-modality: many answers cluster on the base rate 25%, the likelihood 80%, and 50%. Where do these different modes come from?

Crucially, there is dramatic instability: in the taxicab frame (Panel B) many more people anchor at or around 80%, so on average they over-react. Instability is inconsistent with a mechanical tendency toward base rate neglect (Edwards 1968, Grether 1980, Rabin 2002), with a shrinkage of beliefs to the prior due to noise (Woodford 2020, Enke and Graeber 2023), or even with fixed heuristics. Even answers previously attributed to epistemic uncertainty are unstable: the 50:50 mode essentially disappears when moving to taxicabs. The evidence is suggestive of selective attention. In balls and urns many people appear to neglect the color of the drawn ball, and answer with the base rate. In taxicabs, they instead neglect the baseline frequency of blue cabs, and answer with the likelihood. Why are different features neglected in different frames?

Figures 1 and 2 point to two challenges. First, summarizing beliefs in an experiment by the mean or modal response can be highly misleading in the presence of multimodality. In Figures 1 and 2 there is hardly anyone near the mean. This is dramatic in inference, where many people anchor to either the base rate or the likelihood, and fail to integrate them. In fact, experimental protocols that encourage participants to combine the two will fail to elicit what people do naturally: grasp at straws in a complex situation. Answers to standard statistical problems look like duck-rabbit or abortion.

Second, the sharp instability in the distributions of estimates across statistically equivalent problems shows that there are features of these problems other than statistical information that shape beliefs. The language of the question shapes the answer. This has key implications: under- and overreaction are not universal principles, but rather the result of whether in a particular setting relatively more people attend to the base rates (underreaction) or to likelihood (overreaction). To account for our findings, we need a new framework.

3. The Model

We present a model in which the patterns described in Section 2 arise from selective bottom-up attention to the features of the events of a problem. We first define a statistical problem and a rational solution to it. We next formalize the features of events and the role of bottom-up attention.

The formal structure of a statistical problem has three components: i) the sampling process, ii) the statistics, and iii) the hypotheses H_i, H_{-i}. The sampling process defines the set of possible outcomes, or sampling space Ω. Statistics are assigned to two kinds of events. The first are *unconditional* events $k_1 \subseteq \Omega$, of the kind "drawing k_1". Each such event is assigned a statistic π_{k_1}. The collection of such events, denoted by K_1, is a partition of Ω, i.e. $\sum_{k_1 \in K_1} \pi_{k_1} = 1$. Other events are *conditional*, they refine the partition of Ω. They are of the kind "drawing k_2 given k_1". A generic such event is denoted by $k_2|k_1 \subseteq k_1$ and assigned a conditional statistic $\pi_{k_2|k_1}$. The collection $K_2|k_1$ of such events form a partition of its parent k_1, with $\sum_{k_2 \in K_2|k_1} \pi_{k_2|k_1} = 1$ for all k_1. There is a total of $n \geq 1$ steps of conditioning, with the statistic corresponding to a generic step j event ($1 < j \leq n$) denoted by $\pi_{k_j|k_{j-1} \cdots k_1}$. We focus on the case in which statistics are probabilities. In Appendix B we show that the model also covers the case in which they correspond to absolute frequencies. Finally, hypotheses H_i, H_{-i} are events in Ω. We allow for $H_i \cup H_{-i} \subset \Omega$ which captures, among other things, inference problems: data provision restricts hypotheses to a subset of Ω. The statistical problem is solvable because the elementary events $\omega \in \Omega$ that constitute hypotheses are generated by a specific path of events $k_1, k_2|k_1 \ldots, k_n|k_{n-1} \ldots, k_1$ to which statistics are attached.

Consider the problems of Section 2. For sequences of two coin flips ($n = 2$) the sample space is $\Omega = \{(h,t),(t,h),(h,h),(t,t)\}$. The first flip defines two unconditional events $h_1 =$ "drawing h in the first flip" and $t_1 =$ "drawing t in the first flip", which are associated with statistics $\pi_{h_1} = \pi_{t_1} = 0.5$. The second flip defines the conditional events $h_2|k_1 =$ "drawing h in the second flip given k in the first" and $t_2|k_1 =$"drawing t in the second flip given k in the first". These events are assigned statistics $\pi_{h_2|k_1} = \pi_{t_2|k_1} = 0.5$ for $k_1 = h, t$. With i.i.d. draws, a step j event can be written

unconditionally as k_j, with associated statistics $\pi_{k_j} = 0.5$ for $k_j = h, t$. For inference, which has also two steps ($n = 2$), the sample space is $\Omega = \{(A, g), (A, b), (B, g), (B, b)\}$. The unconditional events consist of the "selection of urn" $U = A, B$, denoted by $k_1 = U$, and the conditional events consist of "drawing a ball of color k_2 from U", denoted $k_2|U$ for $k_2 = b, g$. Unconditional events are assigned base rates $\pi_A = 0.25$ and $\pi_B = 0.75$, and conditional events are assigned likelihoods $\pi_{g|A} = 0.8$ and $\pi_{b|A} = 0.2$ for urn A and $\pi_{g|B} = 0.2$ and $\pi_{b|B} = 0.8$ for urn B. Here the process is not i.i.d.

A rational solution consists of: a) expressing each hypothesis as a partition of the events about which statistics are provided, b) computing the probability of each hypothesis using these statistics, and c) normalizing the estimate if probabilities do not add to one, $H_i \cup H_{-i} \subset \Omega$. Sometimes different partitions of hypotheses exist, but they all lead to a correct answer.

We describe a decision maker, the DM, who solves the problem by attending to salient features of the hypotheses. In Section 3.1 we formalize the features of events. In Section 3.2, we formalize how selective attention shapes probability estimates by laying out the steps through which the DM solves a problem. The DM reaches the correct answer if she attends to the relevant features, but commits errors if not. Section 3.3 formalizes the DM's bottom-up attention to features: salience can cause the DM to consider too few relevant features, or to focus on irrelevant ones. Following the psychology of bottom up attention, we formalize two drivers of salience: contrast and prominence. Section 3.4 describes how to apply the model in the lab and offers guidance on field applications.

3.1 The Features of Events

Each event $\omega \in \Omega$ is described by $F > n$ features, collected in vector $f(\omega) = (f_1, f_2, \ldots, f_F)$. The first n features $f_1, \ldots, f_n$ identify the unconditional and conditional events $k_1, k_2|k_1, \ldots$ that must occur for ω to happen, from the coarsest k_1 to the finest $k_n|k_{n-1} \ldots k_1$. We call these features "statistical", because the *true probability* of each such event, denoted by $\Pr(f_j)$, $j \leq n$, is given by a statistic. With two coin flips the statistical features are f_1 ="first flip is k_1" and f_2 = "second flip is

k_2" with true probabilities $\Pr(k_1) = \pi_{k_1} = 0.5$ and $\Pr(k_2) = \pi_{k_2} = 0.5$. In ball and urns, they are f_1 ="select urn k_1" and f_2 = "draw a ball of color k_2 from k_1", whose true probabilities $\Pr(k_1)$ and $\Pr(k_2|k_1)$ are the base rate of urn k_1 and the likelihood of k_2 in k_1, respectively.

The remaining features $f_{n+1}, \dots, f_F$ of ω are not directly tied to statistics, and we call them "ancillary". Like statistical features, each ancillary feature captures a property of the event and hence an equivalence class to which it belongs. In coin flips, one such feature is a sequence's "share of heads", which we denote by $sh \in [0,1]$. It identifies the class of sequences having the same share of heads as ω. This is a key feature for it shapes the similarity of a sequence to the coin that generated it: (h,t) is similar to coin because its 0.5 share of heads is what a fair coin tends to produce. Longer sequences have more features, e.g. (h,t,h,t,h,t) is "alternating", and (t,t,t,h,h,h) is "sorted".

In inference, an important ancillary feature is whether the signal realization is similar to the process that generated it, in the sense of coming from the hypothesis for which this signal is most likely. In Section 2, urn A is 80% green and urn B is 80% blue. Thus, a green signal is similar to A, not to B, and vice-versa for blue. We call "match" the feature taking value $m = 1$ if the color of the ball is similar to the urn, and $m = 0$ otherwise. This feature defines two equivalence classes: events (A,g) and (B,b) form the class of signal realizations similar to the hypothesis, $m = 1$, while events (A,b) and (B,g) form the class of dissimilar ones, $m = 0$.

The share of heads, sh, in coin flips, and match, m, in inference are key and intimately related features: they capture the similarity of events and their data generating process. They are thus linked to KT's "representativeness" heuristic: an event is representative of a statistical process if it resembles salient features of the latter. In our model, though, there are no stable heuristics. There are many features, including the statistical ones tied to specific sampling steps, and others capturing different properties, such as the similarity of an event to the statistical process described in the problem. These features "compete" for DM's attention, shaping representations and biases.

To simplify the analysis, we focus on the case with $F = n + 1$: each $\omega \in \Omega$ is described by the n statistical features set by the problem plus an ancillary one, sh in coin flips and m in inference. The restriction to one ancillary feature may reduce the model's explanatory power, but buys us parsimony and does not affect our core predictions. In Section 3.4 we discuss the selection of features, in both experimental and field contexts, which are important to apply the model.

3.2 Attention to Features, Representation and Solution

Given the statistical problem and its features, the DM: 1) constructs a simplified feature-based representation of the hypotheses based on selective attention, 2) computes of the probability of these representations using the statistics, and 3) normalizes the estimate. Denote by $\alpha_j \in \{0,1\}$ the DM's attention to feature $j = 1, \dots, O$, where $\alpha_j = 1$ if feature j is attended to and $\alpha_j = 0$ if not. The attention profile is $\alpha = (\alpha_1, .., \alpha_{n+1})$. For simplicity, the DM attends either to statistical or ancillary features, not to the mixtures of the two. We also assume that the DM can attend to at most K features, $\sum_j \alpha_j \leq K$, which captures the well-established fact that attention is limited. Denote the set of feasible attention profile by A_K. Selective attention then distorts representations as follows.

Task 1 *(Selective Attention). At attention profile $\alpha \in A_K$ the DM simplifies the feature vector $f(\omega)$ of each event $\omega \in H_i$ in the hypothesis as $\tilde{f}_\alpha(\omega) = \left(\tilde{f}_{\alpha,1}, \dots, \tilde{f}_{\alpha,n+1}\right)$, where:*

$$\tilde{f}_{\alpha,j} = \begin{cases} f_j & if \quad \alpha_j = 1 \\ \varphi & if \quad \alpha_j = 0 \end{cases}. \tag{1}$$

Hypothesis H_i is then represented as $R_\alpha(H_i) = \bigcup_{\omega \in H_i} \tilde{f}_\alpha(\omega)$.

The DM replaces the value of each unattended feature in $f(\omega)$ with "φ", meaning that this feature is not used to describe events. Consider a coin flip problem in which the DM evaluates $H_1 = (h, h)$ vs $H_2 = (h, t)$. If she attends to individual flips, neglecting the share of heads, she represents H_1 as "first head and then head", $R_\alpha(H_1) = (h_1, h_2, \varphi)$, and H_2 as "first head and then tail", $R_\alpha(H_2) = (h_1, t_2, \varphi)$. If instead she attends to the share of heads, neglecting individual flips, she

represents H_1 as "share of heads is 1", $R_\alpha(H_1) = (\varphi, \varphi, 1)$, and H_2 as "share of heads is 0.5", $R_\alpha(H_2) = (\varphi, \varphi, 0.5)$. The DM describes the hypotheses differently when she attends to different features of events. Attention to features then shapes her use of statistics in Task 2.

Task 2 (*Simulation*). *For each* $\tilde{f}(\omega) \in R(H_i)$, *let* $\Pr(\tilde{f}_j)$ *denote the true probability of event* $\tilde{f}_j$ *in* $\tilde{f}(\omega)$, *with the convention* $\Pr(\varphi) = 1$. *The DM simulates* H_i *as:*

$$\Pr\big(R(H_i)\big) = \sum_{\tilde{f}(\omega) \in R(H_i)} \Pr(\tilde{f}_1) \cdot \Pr(\tilde{f}_2) \cdots \Pr(\tilde{f}_{n+1}). \tag{2}$$

The DM computes the probability of the features-events he attends to. If she attends to more than one statistical feature, for each vector $\tilde{f}(\omega) \in R(H_i)$ she computes their joint probability $\Pr(\tilde{f}_r \cap \ldots \cap \tilde{f}_s)$ by multiplying their probabilities. She then sums the products across all vectors. A DM attending to individual flips simulates $H_1 = (h, h)$ and $H_2 = (h, t)$ by multiplying the 0.5 statistic attached to these features, $\Pr\big(R_\alpha(H_1)\big) = \pi_{h_1} \cdot \pi_{h_2} = (0.5)^2$ and $\Pr\big(R_\alpha(H_2)\big) = \pi_{h_1} \cdot \pi_{t_2} = (0.5)^2$. If instead the DM attends to the share of heads, she simulates the same hypotheses by computing the true probability of sh, simulating $R_\alpha(H_1) = (\varphi, \varphi, 1)$ by the probability of obtaining only heads $\Pr(sh = 1) = (0.5)^2$, and the other hypothesis $R_\alpha(H_2) = (\varphi, \varphi, 0.5)$ by the probability of obtaining a balanced sequence $\Pr(sh = 0.5) = 2 * (0.5)^2$. Different representations focus the DM on different features of hypotheses, leading to different simulated probabilities.

The DM then reaches the final estimate by normalizing the simulated probabilities.

Task 3. *(Normalization). The DM computes the probability of* H_i *as:*

$$\Pr(H_i; \alpha) = \frac{\Pr\big(R_\alpha(H_i)\big)}{\Pr\big(R_\alpha(H_i)\big) + \Pr\big(R_\alpha(H_{-i})\big)}. \tag{3}$$

Normalization is only material when the simulated probabilities do not add to one, which is the case in our running example. A DM attending to individual flips estimates the relative probability of $H_1 = (h, h)$ vs $H_2 = (h, t)$ by normalizing the identical $(0.5)^2$ simulations of the two hypotheses, yielding $\Pr(H_1; \alpha) = 0.5$. This DM does not commit the Gambler's Fallacy. A DM instead attending to the share of heads erroneously simulates H_2 with the broad equivalence class of balanced

sequences yielding, after normalization, $\Pr(H_1; \alpha) = 1/3$. This DM commits the Gambler's Fallacy. This bias is due to the fact that she represents hypotheses using the wrong feature: the share of heads.

In general, the DM is biased whenever she attends to the wrong features.

Proposition 1 *(Rationality). For a given statistical problem, there exists a set of event-specific attention vectors* $\alpha^*(\omega) = (\alpha_1^*, \dots, \alpha_{n+1}^*)$, $\omega \in H_i \cup H_{-i}$, *containing at least one zero such that a DM simplifying features according to* $\alpha^*(\omega)$ *in Equation (1) implements the Bayes' rule.*

It is always possible for our DM to reach the correct solution. To do so, she needs to simplify events by focusing on features that are relevant to the problem while neglecting features that are not relevant to it. With the correct simplification strategy in Equation (1), the structure of Tasks 1, 2 and 3 guarantees a correct solution. As we show in the proof, the minimum number of relevant features of a problem can be easily found: they identify a coarsest partition of the hypothesis in terms of events whose probability can be computed.[2] In our example, there is a unique partition of the hypotheses H_1 and H_2, constituted by the atoms (h, h) and (h, t), respectively. These atoms are identified by their first and second flip. The share of heads is instead not relevant to *this* problem because the class of events having $sh = 0.5$ includes both (h, t) and (t, h), so it does not represent a partition of H_2. This is why the DM reaches a correct solution of this problem when she attends to the first and second flip and she instead commits the Gambler's Fallacy when she attends to the share of heads. Our theory of judgment biases is squarely based on this idea: erroneous simplification of features based on selective attention.[3] But what shapes attention? We address this question next.

Section 3.3 Bottom-up Attention to Features

There is a consensus in psychology that selective attention is based on two mechanisms: top down and bottom-up. Top-down attention reflects motivational factors such as the relevance of a

[2] For instance, hypothesis $H = \{(h, t), (h, h)\}$ can be perfectly represented by the feature h_1 = "first flip is h", $R(H) = (h_1\varphi, \varphi)$, which identifies the coarsest partition and neglects the other two features, leading to a correct simulation.

[3] Another attention limit implicitly imposed in Task 1 compared to the rational benchmark in Proposition 1 is that the DM does not select an event-specific attention vector, $\alpha(\omega) = \alpha$ for all ω. This limit does not play a role in our analysis.

stimulus to the goals of the DM. Rational inattention models formalize this idea (Sims 2003; Gabaix 2019, Woodford 2003, 2020; Khaw et al., 2021).[4] Bottom-up attention reflects an extent of involuntary focus on salient stimuli (BGS 2012, 2013, 2022, Li and Camerer 2022). At the extreme, a stimulus that is entirely irrelevant for our goals can draw attention if it is very salient, as when a stain on the wall draws our attention and distracts us from watering a plant.

Section 2 suggests the importance of bottom-up forces. Different people use different statistics despite having the same incentives for accuracy: they do not choose the "most accurate" statistics for a given attention limit K, as for instance is implied by models of sparsity (Gabaix 2014). When thinking about a problem with many features, attention may be driven bottom up to some of them, even if irrelevant, causing neglect of others. Systematic instability from irrelevant changes that cause a feature to become salient is the result, also inconsistent with goal-optimal attention.

This approach yields a new perspective on biases. In standard models, including those of perceptual noise (e.g. Khaw, Li, and Woodford 2022), people "know" the Bayes rule but distort true probabilities in a stable way. In our model bias instead arises because, due to bottom-up attention, people answer a question different from the original one (Kahneman and Frederick 2002). Biases, i.e., systematic errors, arise in the representation of hypotheses, not in numerical computations.

Our new predictions follow from regularities in bottom-up attention. While there is no complete theory, two factors are known to be important: contrast and prominence. Contrast means that a stimulus is more salient if it strongly differs from the background (e.g. the stain has a different color than the wall). Prominence means that the stimulus is more salient if it is located in the center of the visual field or more top of mind (e.g. if the stain is by the light switch, or if one is reminded of it). In both cases, salience depends on context.

We formalize these forces using salience theory (BGS 2012, 2013, 2022), which models how the salient features of goods, e.g. quality or price, affect valuation and choice. In statistical problems,

[4] Attention may be distorted due to mis-specified models (Schwartzstein 2014), but even in this case it reflects goals.

salience is a property of representations $R_\alpha(H_i), R_\alpha(H_{-i})$, which are shaped by the attention vector α. Consider first the contrast induced by α. In BGS, an attribute such as price is contrasting when it sharply favors one of the goods. In a statistical problem we likewise say that attending to a feature induces contrast if it sharply favors one hypothesis over the other. Formally, the contrast of α is:

$$C(\alpha) = \frac{\left|\Pr\left(R_\alpha(H_i)\right) - \Pr\left(R_\alpha(H_{-i})\right)\right|}{\Pr\left(R_\alpha(H_i)\right) + \Pr\left(R_\alpha(H_{-i})\right)}. \quad (4)$$

The numerator in (4) captures the extent to which the representation favors one hypothesis over the other. The denominator captures diminishing sensitivity, as in BGS (2012, 2013). To illustrate, when assessing (h,h) vs (h,t), the contrast induced by the share of heads, $\alpha = (0,0,1)$, is given by $|\Pr(sh = 1) - \Pr(sh = 0.5)|/(\Pr(sh = 1) + \Pr(sh = 0.5)) = 1/3$. Importantly, contrast is shaped by the objective parameters of the problem. In coin flips, for example, it is shaped by the 0.5 probability of a head and the sequence length $n = 2$. In inference, by the base rate and the likelihood. In our experiments, we manipulate contrast by changing statistics.

Consider prominence next. In BGS (2022), as in Chetty et al (2009), an attribute, such as the price or sales tax, is more salient if it is more visible to the consumer. Analogously, in a statistical problem a feature is more salient if it is more visible or explicit in the description. The prominence of a feature can vary across people with different experiences, and may even vary for a person over time. For instance, demand for insurance increases after floods because the recent experience makes this risk top of mind (Slovic, Kunreuther, and White 1974). Something analogous occurs in statistical problems, based on the description context. Describing an inference problem in a courtroom setting may increase the prominence of the signal because people have prior experiences reacting to witness reports in court. This renders the signal salient even if its statistical informativeness is unchanged.

In our experiments we propose several treatments that intuitively increase the prominence of specific features, especially by making them more visible/explicit in the description, but we do not measure prominence directly. We thus introduce prominence as a latent variable, and show that it creates a systematic association between attention to features and beliefs, which we can empirically

validate. We formalize the prominence of feature j by a scalar P_j. Aggregating across features yields the prominence of an attention profile. We consider the simplest possible specification, and define the prominence $P(\alpha)$ of an attention vector α as the average prominence of its features:

$$P(\alpha) = \frac{\sum_j \alpha_j P_j}{\sum_j \alpha_j}. \quad (5)$$

Equation (5) captures two important aspects of attention. First, making a feature more prominent affects all attention profiles that use it: increasing P_j increases the salience of all representations with $\alpha_j = 1$. In balls and urns, making the act of "drawing a green ball from U" more prominent cues all representations that use this feature, including in conjunction with the "selection of U" feature. Second, there is interference: if a DM attends to feature j', increasing the prominence P_j of feature j is less impactful, because the DM's attention is divided. Interference favors sparse representations. We see the duck or the rabbit, but not both at once.

The salience of attention profile α increases in its contrast $C(\alpha)$, prominence $P(\alpha)$, and also an individual specific random (extreme value) term ϵ_α. This term captures not only stable individual differences due different past experiences, but also transient fluctuations in attention. To simplify, we formalize salience as additive in these terms.

Salience and Attention. *The DM uses attention profile $\alpha \in A_K$ that maximizes total salience:*

$$\alpha = argmax_{\tilde{\alpha} \in A} \; C(\tilde{\alpha}) + P(\tilde{\alpha}) + \epsilon_{\tilde{\alpha}}. \quad (6)$$

The idiosyncratic noise $\epsilon_{\tilde{\alpha}}$ generates heterogeneity in the attention profiles and beliefs across participants. Thus, the distribution of estimates is multinomial, with the relative share on each mode pinned down by the relative salience of features. Because $\epsilon_{\tilde{\alpha}}$ varies across people, the theory predicts an individual level association between attention and probability estimates. As one feature becomes more salient, there is an aggregate reallocation of attention to that feature and of estimates to the corresponding mode. For simplicity, in Sections 4 and 5 we assume that the attention limit is not binding: $K \to \infty$. We study the interaction of K with salience in Section 6.2.

3.4 Applying the Model

Our DM solves a statistical problem by attending to its salient features and forming simplified representations $R_\alpha(H_i)$ and $R_\alpha(H_{-i})$ of hypotheses. To apply and test our model generally, the analyst must specify and measure two objects: features and attention. Some features are directly given by the statistics of the problem: the 50:50 outcomes of individual coin flips, the base rate (of urn selection) and likelihood (of drawing a color) in inference. Ancillary features are not directly specified, but often capture broader properties of events. We identified two such features based on the similarity of an event to its data generating process, following the work on the representativeness heuristic (Kahneman Tversky 1972). In more complex problems, there may be more ancillary features that shape beliefs, just like there are many non-hedonic yet salient features, such as advertising and broader context, that shape consumer choice. These features can be empirically discovered, for instance by asking people a rationale for their choices or by using algorithms.[5] Discovering and specifying features is thus the key first step.

Once some explanatory features are identified, the model can be tested by studying how behavior, captured by the estimate $\Pr(H_i; \alpha)$, and measured attention α jointly shift when one feature becomes more salient. There is no universally accepted best practice in measuring attention, but several approaches are available. Eye tracking (Reutskaja et al 2011) is often used to capture visual attention, but for our purposes we need to measure a more semantic kind of attention: the reliance on a feature when solving a problem. We offer three approaches to such measurement, each outlined in our pre-registration. First, after participants solve the statistical problem, we ask them, "Could you describe to us in your own words how you came up with your answer to the previous question?" We then use a language model to code these responses according to whether the participant appeared to be paying attention to specific features (see the Appendix for details). Second, after the free-response,

[5] For example, Kleinberg, Liang, and Mullainathan (2017) use algorithms to detect predictable patterns people use when producing random looking sequences, which can help identify features of the data that people associate with randomness. In a field setting, Kleinberg et al (2018) find that judges underperform algorithms in identifying defendants who will commit crime on bail, and tend to be more lenient if the defendant is well groomed (Ludwig and Mullainathan 2023). This feature was discovered via machine learning, rather than what the analyst specified ex ante.

a multiple-choice question asks participants to select from a list the features they attended to. Third, we ask respondents to rate the *similarity* between events and infer attention from these ratings. The connection between similarity and attention to features is well established (e.g. Tversky and Gati 1982, Nosofsky 1988): people judge two objects to be more similar when they attend to features the two objects share.[6] We then assess whether different measures yield comparable results.

To summarize, to apply our model to a general setting, one needs to specify a) the key features of the problem, and b) how partial attention to them maps to beliefs ($\Pr(H_i;\alpha)$). Our model then generates two new, testable predictions. First, beliefs should be correlated with measured attention, which by Equation (6) entails multimodality. Second, treatments that change the relative salience of features, Equations (4) and (5), should generate corresponding shifts in beliefs and attention. In Sections 4 and 5, we apply this method to coin flips and inference, respectively.

4. Salience, Multimodality and Instability in Gambler's Fallacy

We show that, applied to coin flips, our model yields the multimodality and instability in the distribution of estimates in Section 2 and new predictions, which we test, on how changes in the description of the problem affects measured attention to features and probability estimates.

The Problem and its Features. Recall from Section 3 that $\Omega \equiv \{h,t\}^n$, where n is the number of flips. The feature vector of a sequence, $f(\omega) = (f_1, \dots, f_{n+1})$, is pinned down by n statistical features each corresponding to individual flips $f_i = h_i, t_i$ for $i \leq n$, and by the ancillary feature "shares of heads", $f_{n+1} = sh$, which is the share of heads in ω. The DM assesses the relative likelihood of two sequences: H_1 vs. H_2, where the former is unbalanced ($sh = 1$), and the latter is balanced ($sh = 0.5$). Each hypothesis has its feature vector $(f_1, \dots, f_n, sh)$.

[6] In a classic example, Tversky (1977) showed that Austria was deemed similar to Hungary when geography is salient and hence attended to, but similar to Sweden when political alignment is salient and hence attended to. Formally, under attention profile α the similarity between two events ω_1 and ω_2 could be written as:

$$S(\omega_1, \omega_2; \alpha) = 1 - \sum_j w_j d_j,$$

where d_j takes value 1 if the two events differ along feature $j = 1, \dots, F$ and zero otherwise, while $w_j = \alpha_j / \sum_k \alpha_k$ captures the DM's attention to feature j relative to the other features she attends to.

Attention and Representation. A DM attending to all statistical features, individual flips, while ignoring the share of heads has attention vector $\alpha_n = (1,1,..,0)$, and represents the generic hypothesis by $R_{\alpha_n}(H_i) = (f_1, \dots, f_n, \varphi)$. This DM behaves rationally: by Equation (2) she simulates the generic hypothesis by $\Pr\left(R_{\alpha_n}(H_i)\right) = (0.5)^n$, which yields – after normalization – the correct estimate $\Pr(H_1|\alpha_n) = 0.5$. The rational estimate is also reached by a DM only attending to $r < n$ flips, who simulates $\Pr\left(R_{\alpha_r}(H_i)\right) = (0.5)^r$, which is again identical across hypotheses. By contrast, a DM attending to the share of heads of a sequence has attention $\alpha_{S,n} = (0, \dots 0,1)$ and represents hypotheses as $R_{\alpha_{S,n}}(H_i) = (\varphi, \dots, \varphi, sh)$. By (2), then, her simulation is $\Pr\left(R_{\alpha_{S,n}}(H_i)\right) = \Pr(sh)$, the probability that a sequence has the same share of heads of the hypothesis. This is a large equivalence class for the balanced sequence H_2 and a small one for H_1, causing Gambler's Fallacy.

Endogenous Attention and Estimates. To determine the distribution of estimates in an experiment, we must describe the attention profile of different DMs. By Equation (6), attention depends on prominence, contrast, and a random term. Denote by P the scalar prominence of each individual flip relative to sh. Denote by $C\left(\alpha_{S,n}\right)$ the contrast of $\alpha_{S,n}$, which depends on length n.

Proposition 2 *A share $\mu\left(\alpha_{S,n}\right)$ of DMs attends to the share of heads and commits the Gambler's Fallacy, estimating the relative probability of the unbalanced sequence as:*

$$\Pr\left(H_1; \alpha_{S,n}\right) = \frac{\binom{n}{n/2}}{1 + \binom{n}{n/2}} < 0.5. \tag{7}$$

The remaining DMs attend to a subset of flips and answer $50{:}\,50$.

There are two modes for beliefs: one at 50% and another in Equation (7) below 50%. DMs who focus on individual flips make a correct judgment. DMs who focus on the share of heads instead commit the Gambler's Fallacy. These DMs represent the problem as "drawing *any* balanced sequence

vs. a fully unbalanced one", which inflates the likelihood of the former.[7] The prediction here is that a DM's measured attention to the share of heads will positively correlate with her tendency to commit the Gambler's Fallacy. The next testable predictions concern instability in attention and estimates.

Corollary 3 *The Gambler's fallacy becomes more severe as the length* n *of the sequences increases. The share* $\mu(\alpha_{S,n})$ *of DMs who attend to the share of heads and commit the Gambler's Fallacy increases in sequence length* n *and decreases in the prominence of individual flips* P.

As n increases the Gambler's Fallacy gets stronger for two reasons. First, because the number of balanced sequences grows with n while that of unbalanced sequences stays constant, the relative number of the latter sequences in Equation (7) falls. This effect appears in Figure 1, where the distribution of answers shifts to the left for longer sequences. Second, as n rises the salience of sh increases, because $C(\alpha_{S,n}) = \left[\binom{n}{n/2} - 1\right] / \left[\binom{n}{n/2} + 1\right]$ rises with n. This causes a shift in attention to the ancillary feature, increasing $\mu(\alpha_{S,n})$. When comparing two long sequences such as $hthtth$ and $hhhhhh$, the DM cannot avoid thinking how much harder it is, with a fair coin, to get a long streak of heads compared to a $50{:}\,50$ outcome. The share of heads sticks out as a focal representation, and for many DMs replaces the original question. This effect appears in Figure 1 as a collapse in the 50:50 mode. Corollary 3 also predicts that, conversely, raising the prominence of individual flips should draw attention away from sh, and reduce the incidence of the Gambler's Fallacy.

We next test the model's predictions. First, we study multimodality: the individual level link between attention to sh and the Gambler's Fallacy in Proposition 2. Second, we study instability: whether attention to sh and the Gambler's Fallacy increase when sh becomes more salient, as in Corollary 3. To do so we run several treatments, listed in Table 2. After eliciting participants' estimates, we measure free-response and direct-elicitation proxies for attention to features. The

[7] In Section 6 we show that the attention limit qualifies this result: when $K < \infty$ and $n > 2$ several modes of the kind in (7) arise, some of which exhibit a more severe form of the Gambler's Fallacy than others.

features include: 1) the share of heads, 2) whether the final flip is heads or tails, and 3) anything else. For a subset of participants, later in the survey we also elicit perceived similarity between the two judged sequences. We allow "similarity" to be fully subjective, without encouraging participants to consider any particular feature. Intuitively, if the DM attends to the share of heads rather than to individual flips, similarity should be lower: the sequences in fact sharply differ along sh, while they share several flips.[8] We thus interpret low similarity as a proxy for attention to the share of heads. We later present evidence corroborating this idea.

Multimodality in Attention and Estimates. First, we document multimodality in attention and probability estimates within each treatment. Pooling across all treatments and adding treatment fixed effects, we run OLS regressions of a respondent level indicator for whether a respondent commits Gambler's Fallacy (i.e., report a belief of less than 50 out of 100 for the unbalanced sequence) on indicators for directly elicited and free response attention to share of heads (Column 1), on the perceived similarity between sequences (Column 2), and on all three attention proxies (Column 3).

	Dependent Variable: Commit Gambler's Fallacy		
	(1)	(2)	(3)
Directly Elicited Attention to Share	0.169***		0.180***
	(0.017)		(0.032)
Free-Response Attention to Share	0.082***		0.091***
	(0.017)		(0.032)
Similarity between Judged Sequences		-0.062***	-0.066***
		(0.021)	(0.020)
Treatment FEs	Yes	Yes	Yes
N	2855	846	846
R^2	0.110	0.088	0.134

Table 1. Correlating measures of attention with the Gambler's Fallacy. Table shows OLS regressions where the dependent variable is an indicator whether the participant judged the unbalanced sequence to be less likely than the balanced sequence. Similarity measure is normalized (within sequence lengths) to have a mean of 0 and standard deviation of 1. *** indicates statistical significance at the 1% level.

[8] Using the similarity function in footnote 5, if the DM attends to all individual flips the similarity between the balanced and the unbalanced sequence is 0.5, if she attends to the share of heads it is zero.

We see that a participant's measured attention to the share of heads is positively correlated with her tendency to commit the Gambler's Fallacy, which is also negatively correlated with her perceived similarity between the two sequences. Note that each measure of attention has predictive power conditional on the others. These results suggest that attention to the share of heads plays an important role in shaping whether a respondent commits the Gambler's Fallacy.

Instability in Attention and Estimates. Our treatment variations, summarized in Table 2, intend to shift the salience of features. These include the treatments in Section 2, $H_1 = hh$ vs. $H_2 = th$ and $H_1 = hhhhhh$ vs. $H_2 = ththht$, denoted by T_2 and T_6 (for the length of the sequences). By Corollary 3, when moving from T_2 to T_6 the contrast of sh and hence attention to it should increase. To vary prominence, we introduce two new treatments. In T_{full} subjects estimate $H_1 = hhhhhh$ vs. $H_2 = hhhhht$. In T_{last} we tell subjects that the first five flips were $hhhhh$ and ask them to estimate whether the final flip was heads or tails. The treatments are equivalent but T_{last} is intended to make the last flip more prominent (and, thus, the share of heads less salient). By Corollary 3, it should then reduce attention to sh and the Gambler's Fallacy.

Treatment	N	Summary	Purpose
T_2	434	Balanced vs unbalanced 2-flip sequences	Compare to T_6
T_6	405	Balanced vs unbalanced 6-flip sequences	Increase contrast of share compared to T_2
T_{full}	1038	Ask about full 6-flip sequences $H_1 = hhhhht$ vs $H_2 = hhhhhh$	Compare to T_{last}
T_{last}	978	Ask about final flip. in 6-flip sequences i.e.. P(h vs $t \mid hhhhh$)	Increase prominence of final flip compared to T_{full} (and thereby reduce attention to share heads

Table 2. Treatments manipulating salience in the gambler's fallacy problem.

We already saw in Figure 1 that increasing the length of sequences increases the incidence of the Gambler's Fallacy. Consider next the T_{last} vs T_{share} treatments. Figure 4 shows that, consistent with Corollary 3, in T_{last} the mean relative likelihood of H_1 is significantly higher (49.3 vs 44.4 out

of 100, $p < 0.01$), driven in part by an increase in the mode at $50:50$ compared to T_{full} (68% vs 54% of participants, $p < 0.01$).

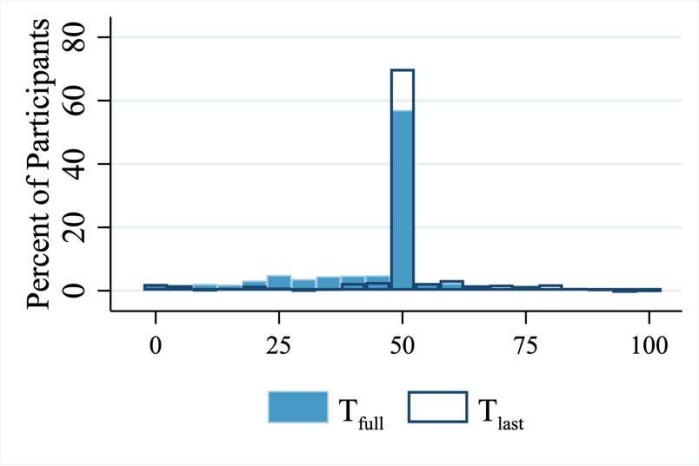

Figure 4. Making last flip more prominent reduces the Gambler's Fallacy. This figure reports the distribution of estimated Pr($ththht \mid ththht\ or\ hhhhh$). Answers closer to 0 indicate higher probability of the balanced sequence.

We next test whether our treatments have effects on our proxies for attention that mirror their effects on beliefs. Figure 5 plots the share of participants in each treatment who commit the Gambler's Fallacy along with the share attending to the share of heads according to the direct-elicitation proxy (Panel A) and the free-response proxy (Panel B). We find a positive correlation in both panels. The correlation is only significant for the free-response measure, since direct elicitation fails to detect greater attention to sh in T_6 than in T_2 (but it correctly detects greater attention to sh in T_{full} than in T_{last}).[9] Reassuringly, the free response measure, which is based on subjects' reasoning, detects model-consistent instability in attention across the four treatments. Multimodality and instability in coin flips are closely associated with bottom-up attention as predicted by our model.

[9] In direct elicitation, attention to sh is not significantly different across T_2 and T_6 (and in fact goes slightly in the wrong direction, 65.7% vs 62.0%, p = 0.27). One explanation is that when $n = 2$ even a respondent focusing on individual flips has in mind that (h, t) is balanced. In the free response measure attention to sh is 46.4% in T_6 and 40.8% in T_2 (p=0.10).

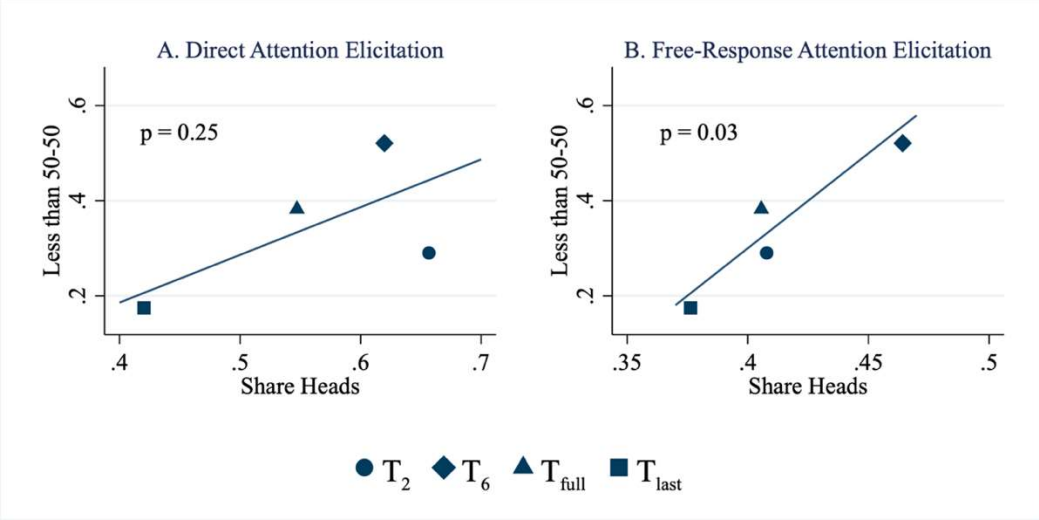

Figure 5. Treatment effects in Gambler's Fallacy and attention. The x-axis is the fraction of participants in each treatment that attend to share heads according to our direct-elicitation (Panel A) and free-response (Panel B) attention measures. The y-axis is the fraction of participants across treatments who judge the balanced sequence to be more likely than the unbalanced sequence.

We conclude by corroborating the connection of similarity judgments and attention to the share of heads and the Gambler's Fallacy. To this end, we added two treatments in a random order at the end of the experiment. In $Probability_n$, participants rated the probability of multiple randomly generated n-flip sequences. In treatment $Similarity_n$, they rated the similarity of *pairs* of n-flip sequences. The sequence length n was randomized across participants to be either 2, 4, or 6. For $n = 2$ ($n = 4$), participants rated all four (sixteen) possible sequences and two (eight) non-overlapping pairs. For $n = 6$, they rated 16 randomly selected sequences and non-overlapping pairs.[10] The similarity measure between the two judged sequences in the Gambler's Fallacy problem described above came from answers in $Similarity_n$.

Figure 3 plots the average stated frequency of a target sequence against its average judged similarity to other sequences, for $n = 2$ (Panel A) and $n = 6$ (Panel B) (see the appendix for the corresponding figure for $n = 4$), with lighter dots indicating more balanced sequences. In both cases, the correlation between judged frequency and average similarity is positive and statistically significant ($p < 0.05$ for both panels). Importantly, this relationship is largely driven by the share of heads: more balanced sequences are deemed to be both more likely and more similar to the average

[10] When calculating average similarity, we correct for the fact that some pairs of sequences were more likely to be selected.

sequence. Indeed, controlling for the share of heads removes any significant correlation between similarity and frequency (see Appendix A).

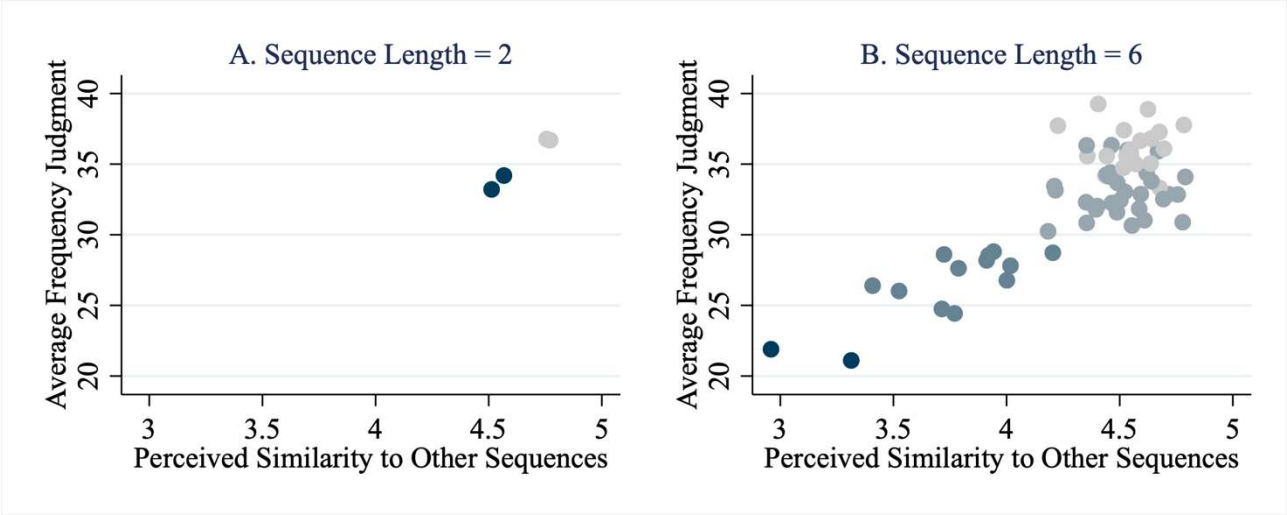

Figure 3. Average judged similarity to other sequences predicts frequency judgments. Lighter dots indicate more balanced sequences, indicating that share heads drives both measures. Frequency judgments are expected number of sequences out of 100 (Panel A) or 1000 (Panel B).

These results corroborate our mechanism: the Gambler's Fallacy arises because, when seeing $hthhtt$ vs $hhhhhh$ many people attend to the share of heads and simulate a specific balanced sequence using superficially similar balanced sequences, neglecting differences in individual flips. Attention-driven representations explain why similarity and probability go hand in hand. In their analysis of human inference, Kahneman and Tversky (1972) famously showed that the perceived similarity between the description of a person called Tom and a librarian correlates with the judged probability that Tom works as a librarian, causing neglect of the low base rate of this occupation. Our model suggests the following explanation: when thinking about Tom, people attend to his described features – "a meek and tidy soul" – and easily simulate a librarian, neglecting many non-salient features that may cause Tom to land in a different job. Both similarity judgments and probability assessments are driven by partial attention to features.

5. Salience, Multimodality and Instability in Inference

We show that bottom-up attention to the features associated with the base rate or the likelihood accounts for the coexistence of under and over-reaction in Figure 2, and that exogenously increasing

contrast or prominence of different features explains instability, establishing a systematic association between measured attention and beliefs.

The Problem and its Features. In balls and urns, $\Omega \equiv \{(A,g),(A,b),(B,g),(B,b)\}$, the statistical features are $f_1 =$ "select urn U" ($U = A,B$) and $f_2 =$ "draw color c from urn U" ($c|U, c = g,b, U = A,B$). As discussed in Section 3, we also define the ancillary "match" feature m, which is 1 for (A,g) and (B,b) and zero otherwise. The DM is asked to estimate the probability of urn A vs B after a green signal. The urn-U hypothesis, $H_U = (U,g)$, has feature vector $(U,c|U,m)$, where m is 1 for H_A and zero for H_B. As in Section 2, we assume that urn A is less likely to be selected and mostly green ($\pi_A < \pi_B, \pi_{g|A} = \pi_{b|B} = q > 0.5$), and that the Bayesian answer is $\beta > 0.5$.

Attention and Representation. We consider five attention profiles $\alpha = (\alpha_U, \alpha_{c|U}, \alpha_m)$. First, a DM attending to both statistical features, $\alpha_\beta = (1,1,0)$, represents the generic hypothesis H_U, $U = A,B$, as first selecting the urn and next drawing a green ball from it, $R_{\alpha_\beta}(H_U) = (U,g|U,\varphi)$. This DM then simulates the hypothesis as $\pi_{g|U}\pi_U$ and obtains, after normalization, the Bayesian answer, $\Pr(H_A;\alpha_\beta) = \beta$. The Bayes' rule is recovered with full attention to relevant features.

Under the other four attention profiles, the DM is biased. A DM attending to urn selection and neglecting the drawing of a color, $\alpha_{BR} = (1,0,0)$, represents the problem as "what is the probability that *a ball* comes from A vs B?", formally $R_{\alpha_{BR}}(H_U) = (U,\varphi,\varphi)$. This DM simulates each hypothesis using its base rate, which yields the answer $\Pr(H_A;\alpha_{BR}) = \pi_A$. A DM attending only to drawing a green ball from U, $\alpha_c = (0,1,0)$, represents the problem as "what is the probability of drawing a green ball from A vs B?", formally $R_{\alpha_c}(H_U) = (\varphi,c|U,\varphi)$. This DM simulates each hypothesis using its likelihood $\pi_{g|U}$ which, in this symmetric case, yields the final estimate $\Pr(H_A;\alpha_c) = q$. A DM attending to the ancillary "match" feature, $\alpha_m = (0,01)$, represents the problem as "what is the probability that the ball is similar to the urn?", formally $R_{\alpha_m}(H_U) = (\varphi,\varphi,m)$. This DM simulates H_A as $\pi_{g|A}\pi_A + \pi_{b|B}\pi_B$, which also yields the likelihood

$\Pr(H_A; \alpha_m) = q$. In these cases, bias takes the form of the DM anchoring to only one statistic in the problem, the base rate or the likelihood.

Finally, DMs who attend to none of the features $\alpha_0 = (0,0,0)$ represent the problem as "what is the probability that one hypothesis vs another is true?". These DMs neglect both the fact that urns are selected with different probabilities and that they have different color compositions. They think "a green ball could come from either urn" and report $50{:}\,50$.[11] This behavior does not reflect a sophisticated reaction to epistemic uncertainty, but rather the fact that no feature is salient. When a feature becomes salient, anchoring to 50:50 should decline, as we saw in Figure 2.

Endogenous Attention and Estimates. Attention depends on contrast $C(\alpha)$, which is computed using the statistics, on systematic prominence $P(\alpha)$, and on the individual specific random term. We denote by P_l the prominence of feature $l = U, c|U, m$. Proposition 4 collects the results above.

Proposition 4 *A share* $\mu(\alpha_\beta)$ *of DMs attends to both statistical features,* α_β*, and gives the correct answer,* $\Pr(H_A; \alpha_\beta) = \beta$*. A share* $\mu(\alpha_{BR})$ *of DMs attends only to urn selection,* α_{BR}*, anchoring to the base rate* $\Pr(H_A; \alpha_{BR}) = \pi_A$*. Shares* $\mu(\alpha_c)$ *and* $\mu(\alpha_m)$ *of DMs attend to the color of the ball or to "match",* α_c *and* α_m *respectively, and anchor to the likelihood* $\Pr(H_A; \alpha) = q$*. The remaining DMs neglect all features and answer* $\Pr(H_A; \alpha_0) = 0.5$.

The model predicts a systematic relationship between measured attention to features and the probability estimate. Within an experimental treatment, this relationship yields the multi-modality observed in Figure 2 due to the random component in attention. This mechanism can be tested by correlating at the individual level attention to features and estimates. Critically, the dependence of the shares $\mu(\alpha)$ on systematic drivers of salience, contrast and prominence, yields predictions on the effect of changing the statistics or context of the problem.

[11] Here, no attention to features can capture literally no attention, as well as the possibility that the DM attends to the problem but is unable to use any feature in the representation of the hypotheses. For example, the DM's attention may jump between "urn selection" and signal, because they favor different hypotheses. In this case, the DM may answer 50-50 because no specific feature (or subset of features) stands out, which could verbally be described as "could be either".

Corollary 5 *The ratio* $[\mu(\alpha_c) + \mu(\alpha_m)]/\mu(\alpha_{BR})$*, which describes the share of DMs attending to signal or match vs. urn selection, as well as the share of answers at the likelihood vs. the base rate, increases with: 1) Contrast of color, i.e. the likelihood* q*, and 2) Prominence of color,* $P_{g|U}$*, or of match,* P_m*. The relative share of Bayesian answers* $\mu(\alpha_\beta)/\mu(\alpha_{BR})$*, is insensitive to* P_m*.*

The model points to two drivers of instability. Contrast implies that making the signal more informative boosts attention it receives and anchoring to the likelihood (the opposite occurs if the base rate becomes more extreme). Prominence implies that changes in context drawing attention to the signal or to its similarity to the hypothesis also increase anchoring to the likelihood, even if normatively irrelevant. The model also predicts that the relative share of people at the Bayesian mode should not change when the "match" feature is made more prominent.

Corollary 5 explains the instability in Figure 2 when switching from balls and urns to taxicabs. Changing the context of inference from balls and urns to taxicabs makes the signal or the match feature more prominent. In taxicabs the signal is the witness' report, "green", while the match feature is the witness' accuracy, namely whether the report is correct or not. Due to personal or fictional experiences, both features of the witness are arguably highly prominent in a courtroom context.

We next test our model systematically. We first study multimodality in Proposition 4 by connecting attention and estimates at the individual level. We then study instability in attention and estimates in Corollary 5 by running several treatments, listed in Table 4.

Multimodality in Attention and Estimates. First, we document multimodality in attention and probability estimates within each treatment. The large majority of answers is anchored to one of the modes in Proposition 4 (ranging from 68.2% to 78.2% of all answers depending on treatment). To check whether measured attention correlates with estimates, we run OLS regressions of an indicator for whether participants anchor at a given mode (base rate, likelihood, the Bayesian answer, and 50-

50) on indicators for measures of attention to its associated feature profile as well as treatment fixed effects.[12] We pool all the inference treatments (listed in Table 4), and include treatment fixed effects.

	(1) Base Rate	(2) Likelihood	(3) Bayes	(4) 50%
Directly Elicited Attention				
Only Urn	0.418***			
	(0.022)			
Only Color/Match		0.408***		
		(0.023)		
Only Urn and Color			0.128***	
			(0.026)	
Nothing				0.166***
				(0.041)
Free-Response Attention				
Only Urn	0.169***			
	(0.022)			
Only Color/Match		0.121***		
		(0.027)		
Only Urn and Color			0.110***	
			(0.026)	
Nothing				0.054***
				(0.011)
Treatment Fes	Yes	Yes	Yes	Yes
N	2061	2061	2061	2061
R^2	0.296	0.256	0.069	0.052

Table 3. Multimodality in attention and in estimates. The dependent variable is whether participants' answers were the base rate (column 1), the likelihood (column 2), within 5 percentage points of the Bayesian answer (column 3), or 50-50 in the inference problem (column 4). All regressions include treatment fixed effects. Robust standard errors in parentheses. *** indicates statistical significance at the 1% level.

Table 3 shows that measured attention profiles strongly predict estimates. In Column 1, directly reporting attending to only the urn feature is associated with being 41.8 percentage points more likely to anchor to the base rate. Conditional on that, free-response attention to urn increases that probability by 16.9 percentage points. Thus, both measures appear to contain information on how respondents think about the problem. The same positive correlation between measured attention and estimates arises for the other modes, corroborating the key prediction of Proposition 4.

[12] For balls-and-urns problems, the list of features includes 1) the probability the computer would choose Jar A vs Jar B, 2) whether the drawn ball was green or blue, 3) whether the drawn ball matched many balls in the jar it came from, and 4) none of the above. For taxicabs, analogous options appeared about the cab companies and the witness report. When deriving the model's predictions, we assume the DM either attends only to (a subset of) the statistical features or only to the ancillary features. Here we assume that statistical features take precedence when participants report paying attention to both statistical features and the ancillary feature. That is, we treat such participants as if they only paid attention to the statistical features they report attending to.

Attention and Instability in Estimates. We develop several treatments, summarized in Table 4, that vary the contrast or prominence of particular features. In the baseline balls and urns and taxicab treatments in Section 2, T_B and T_C respectively, the base rate (for Jar A and green cabs, respectively) is 0.25, and the likelihood is 0.80. We introduce four additional treatments. The "less extreme" likelihood treatment T_{LE} is identical to T_B except the base rate is 0.15 and the likelihood is 0.70. The "more extreme" likelihood treatment T_{ME} is identical to T_{LE} except that the likelihood is 0.90. By Corollary 5, moving from T_{LE} to T_{ME} should increase contrast of the likelihood (or match), boosting attention to these features compared to urn selection and anchoring to q.

We next test the relationship between T_B and T_C. Treatment T_H is nearly identical to T_B but increases the prominence of the match feature "visually", by describing it explicitly in the problem. To do so, we describe the composition of the two urns in terms of the accuracy of the match between the color of a ball and the urn it comes from.[13] In turn, treatment T_U is nearly identical to T_C except that we modify the description of the witness as follows (in italics) "the court found that the witness was *very unreliable: he was* able to identify each color correctly only about 80% of the time, and confused it with the other color about 20% of the time." We also modify the description of the base rate as follows: "*the large majority* of cabs in the city—75% to be exact—are Blue, while the remaining 25% are Blue." Compared to T_C, these wordings reduce the prominence of the witness' report and increase that of base rates. Crucially, the underlying statistics are unchanged.

[13] The question includes the following text: "Imagine two jars filled with marbles, the "Blue Jar" and the "Green Jar". Each jar contains some blue marbles and some green marbles. A computer randomly chooses a jar and draws a marble from it. With probability 25% it chooses the Green Jar, and with probability 75% it chooses the Blue Jar. The computer then records the color of the jar and of the marble. Finally, it puts the marble back and shakes the jar to shuffle its contents. After repeating this procedure many times, we observed the following. For each jar, the marble matched the color of the jar it came from about 80% of the time. About 20% of the time, it was the opposite color."

Treatment	Base Rate	Likelihood	N	Summary	Purpose
T_B	0.25	0.80	480	Balls and urns: baseline	Compare to T_B
T_C	0.25	0.80	199	Taxicabs: baseline	Compare to T_U
T_{LE}	0.15	0.70	497	Balls and urns: less extreme likelihood	Compare to T_{ME}
T_{ME}	0.15	0.90	487	Balls and urns: more extreme likelihood	Increase contrast of likelihood compared to T_{LE}
T_H	0.25	0.80	202	Balls and urns: highlight match	Increase prominence of match compared to T_B
T_U	0.25	0.80	196	Taxicabs: undermine witness's report	Decrease (increase) prominence of report/match (company) compared to T_C

Table 4. Treatments manipulating salience in inference problems.

We first study the instability in estimates across treatments and then link it to the instability in measured attention. Consider contrast first. Figure 6 reports treatments T_{LE} and T_{ME}, where the likelihood varies from 0.7 to 0.9. The distribution of answers in Panel A shows that, consistent with increased contrast drawing attention toward the likelihood's associated features, more participants anchor to the more likelihood in T_{ME} than in T_{LE} (from 15.5% to 22.8%, $p = 0.00$). Anchoring to the base rate also drops (from 32.8% to 23.4%, $p = 0.00$), with little effect on the mass near (i.e., within 5 percentage points of) the Bayesian answer (from 12.1% to 9.2%, $p = 0.15$).[14] As shown in Panel B, the implied treatment effect on the relative share of answers at the likelihood or Bayes vs. the base rate is also consistent with Corollary 5.

[14] Changing the likelihood also changes the correct answer. In the Appendix, we describe a sharper test in which the contrast of the ball's color increases in a spurious way, keeping the correct answer the same. To do so, we describe urns using absolute rather than relative frequencies (i.e, the number of blue vs green balls in each), so that across treatments urns have the same share of green and blue balls but different absolute numbers. Consistent with the model's prediction, when the absolute difference in the number green balls increases, overreaction becomes more common.

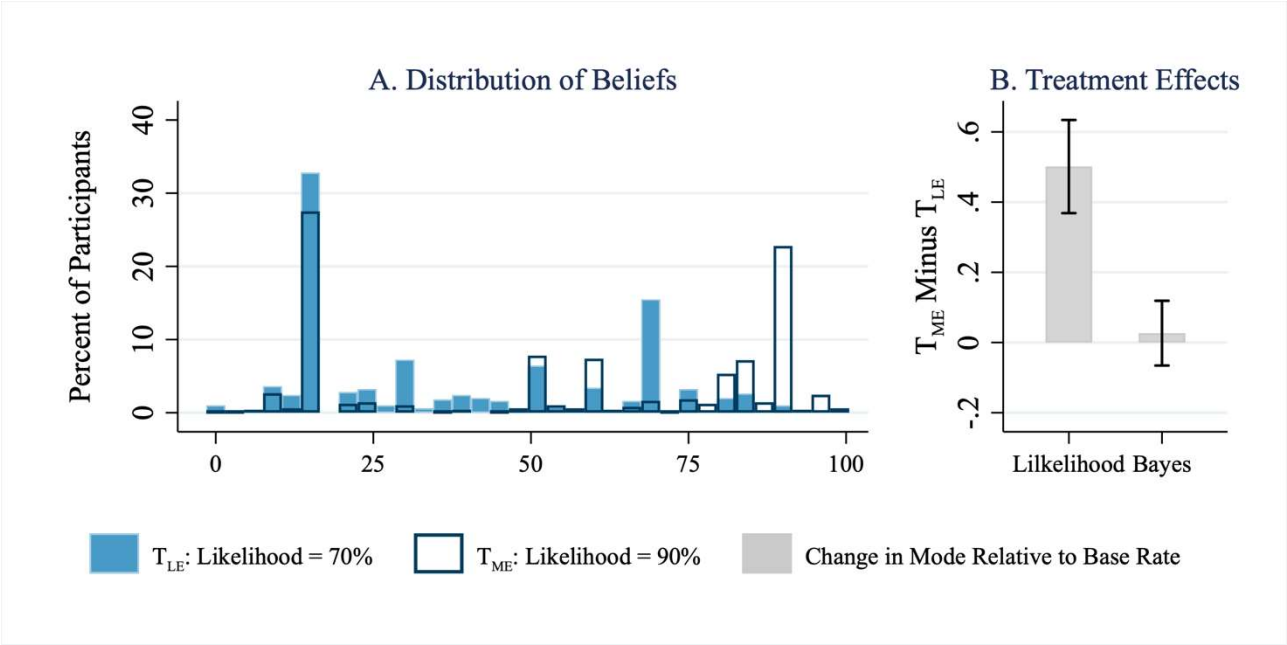

Figure 6. A more extreme likelihood increases anchoring to itself. Panel A shows the distribution of beliefs about $\Pr(A \mid g)$. Panel B shows treatment effects on the fraction of participants who anchor to the likelihood or Bayesian mode divided by the fraction who anchor to the base rate. Whiskers show +/- one standard error.

Consider prominence next. Figures 7 and 8 reconcile the balls and urns and cabs experiments. In Figure 7, Panel A shows that T_H, which adopts the statistics and balls-and-urns setting of T_B but explicitly describes the match feature, dramatically increases the share of participants who anchor to the likelihood, in absolute terms (Panel A) and relative to the base rate (Panel B), in line with Corollary 5. There is also a modest reduction in the relative prevalence of the Bayesian answer.

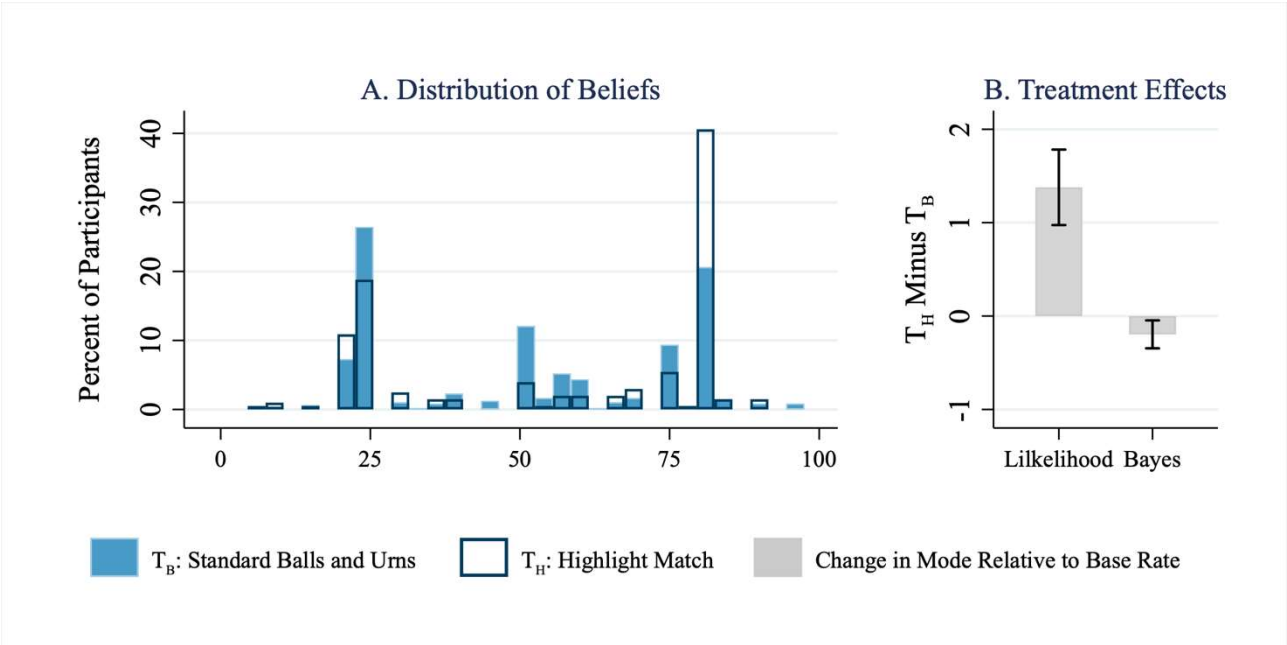

Figure 7. Highlighting the "match" feature boosts anchoring to the likelihood. Panel A shows the distribution of beliefs about $\Pr(A \mid g)$. Panel B shows treatment effects on the share of participants who anchor to the likelihood or Bayesian mode divided by that who anchor to the base rate. Whiskers show +/- 1 standard error.

Figure 8 compares T_C to T_U, which seeks to decrease the prominence of the witness's report and increase the prominence of the base-rate. Panel A shows that this manipulation starkly decreases the share of participants who answer with the likelihood (from 40.6% to 31.1%, $p < 0.01$), and increases, though not significantly, in the share of participants answering with the base rate (from 9.5% to 14.3%, $p = 0.15$). The changes in the share of respondents anchoring to the likelihood vs the base rate, reported in Panel B, are again consistent with Corollary 5.

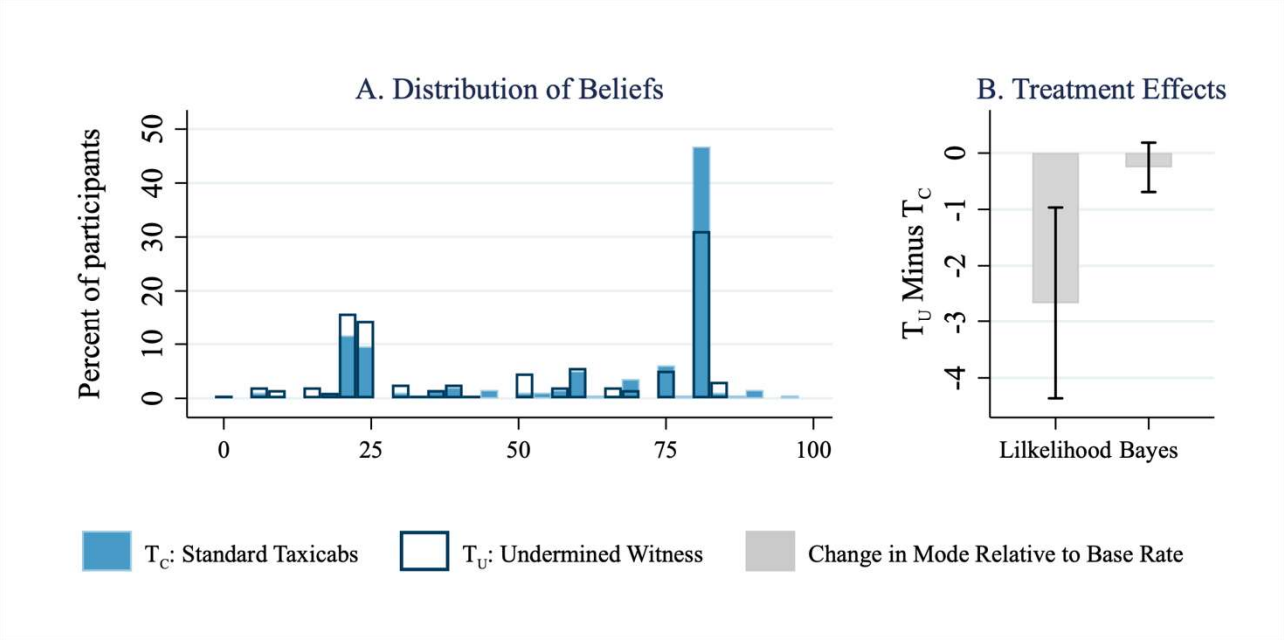

Figure 8. Increasing the relative prominence of the signal reduces anchoring to the likelihood. Panel A shows the distribution of beliefs about $\Pr(G \mid g)$. Panel B shows treatment effects on the share of participants who anchor to the likelihood or Bayesian mode divided by that who anchor to the base rate. Whiskers show +/- one standard error.

Changes in the contrast or prominence of features thus appear to be associated with the predicted changes in the prevalence of different estimates. Do the same treatments change attention to features in the direction predicted by Corollary 5? Figure 9 plots on the x axis the share of participants paying attention to color, match, or both, relative to those attending to urn selection. It plots on the y axis the share of participants anchoring at the corresponding likelihood and Bayes modes relative to those anchoring to the base rate. Panel A reports the results using the direct elicitation measure, Panel B using the free response measure. Both measures of attention show strong positive relationships for the base rate, likelihood, and Bayesian modes. Consistent with Corollary

1, increasing the likelihood from T_{LE} to T_{ME} increases attention to color or match and anchoring to the likelihood. Highlighting the match feature in T_H strongly boosts attention to the same feature and anchoring to the likelihood compared to baseline balls and urns T_B. Finally, undermining the witness in T_U increases relative attention to the base rate and anchoring to it. Taken together, treatment effects on measured attention and beliefs appear tightly linked, in line with Corollary 5.

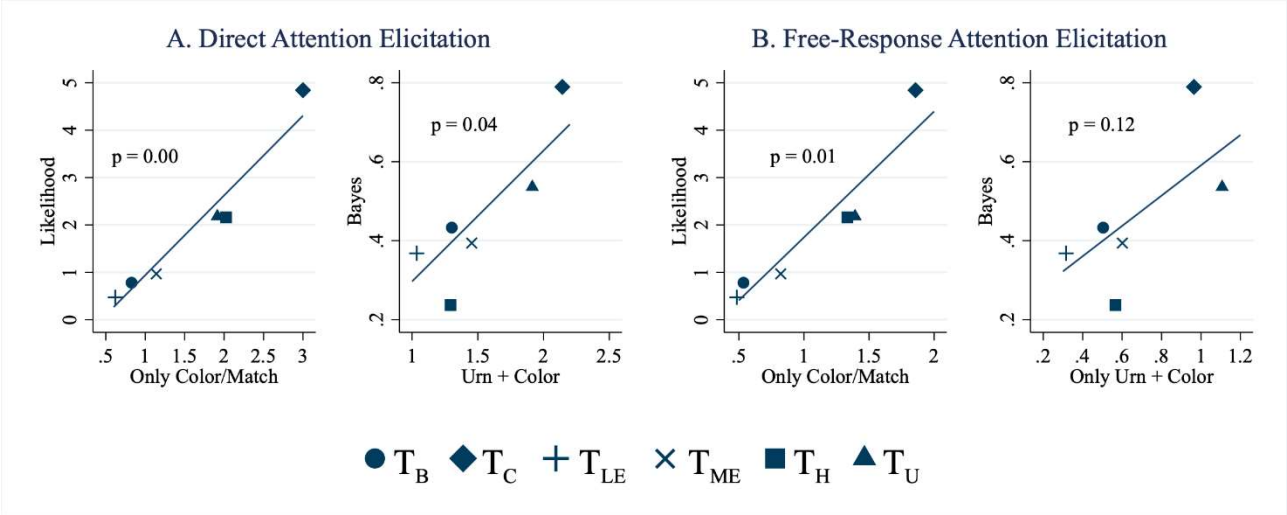

Figure 9. Correlating treatment effects on beliefs and attention. The x-axis is the fraction of participants in each treatment attending to color and/or match (left figure within each panel) and to urn + color (right figure with each panel) divided by the fraction attending only to urn according to our direct-elicitation (Panel A) and free-response (Panel B) attention measures. The y-axis is the fraction of participants in each treatment who anchor to the likelihood (left figure within each panel) or close to the Bayesian answer (right figure within each panel) divided by the fraction who anchor at the base rate.

Model Estimation. Figure 9 shows the correspondence between beliefs and measured attention, but attention is itself an outcome of the primitives of the model, namely the contrast and prominence of features. We next formally estimate a full likelihood model. This allows us to infer the latent cognitive primitives from the estimates and assess whether the theoretical attention allocation predicted by the model matches measured attention. By estimating the cognitive primitives (contrast and prominence) directly, we can also isolate precisely how a given treatment influences the salience of a feature. This allows us to highlight two additional findings. First, the treatment-level prominence of the ancillary feature ("match") should be associated only with increases in measured attention to "match" itself and not to Bayes. Second, the model estimates can inform us how much of the shift in measured attention is due to contrast across all treatments.

Due to the model's multinomial structure, the share of estimates at a given mode $e = Bayes, Likelihood$, relative to that at the base rate in Corollary 5 is given by:

$$\ln \frac{\mu(\alpha_e)}{\mu(\alpha_{BR})} = (P_e - P_U) + \beta\,[C(\alpha_e) - C(\alpha_{BR})] \qquad (8)$$

where $(P_e - P_U)$ is the prominence of attention profile α_e, while the second term is the contrast of α_e, all relative to urn selection. Contrast is pinned down by the statistics of the problem. In our baseline formulation, $\beta = 1$, but we test whether $\beta > 0$. The constant in the above regression captures the relative prominence of e. We estimate these coefficients by maximum likelihood (details in appendix C). Figure 10 plots on the x axis model-implied salience and on the y axis measured attention to the same feature profile.

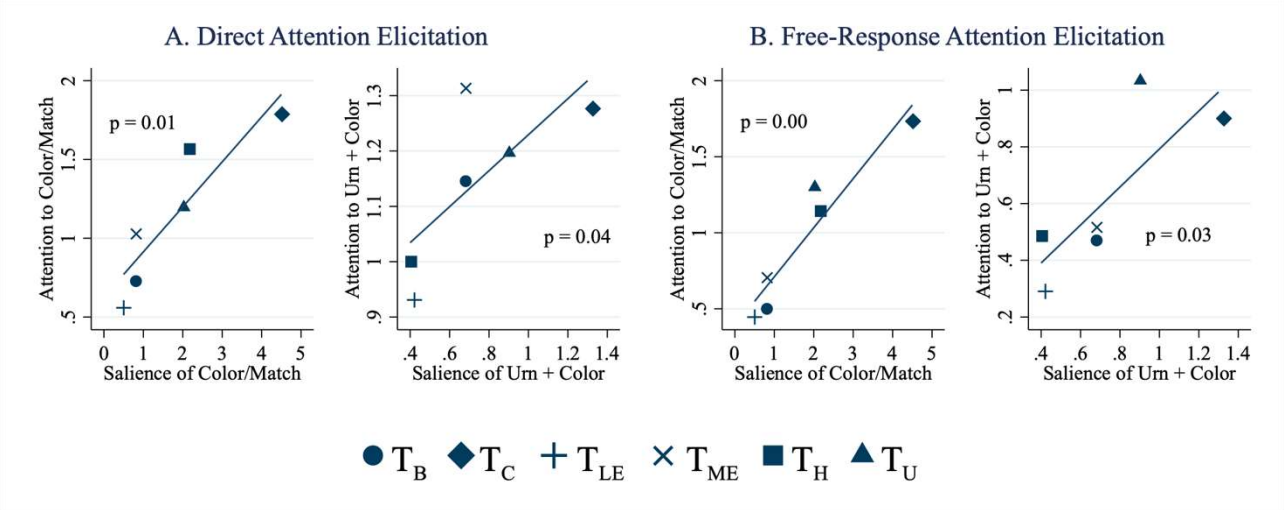

Figure 10. Measured vs revealed attention to features. The x-axis is the estimated salience of each attention profile (where we sum together the color and match salience estimate) relative to the estimated salience of urn. The y-axis is the share of participants who attend to the corresponding profile, as measured by our direct elicitation (Panel A) or free-response measure (Panel B).

Our attention measures are positively correlated with model-implied salience. When beliefs move in a way consistent with an increase in the salience of the signal, match, or the Bayes profile, measured attention on these profiles also increases. Moreover, as our model predicts, the prominence of the ancillary feature, "match", as estimated from beliefs data, is strongly correlated at the treatment level with the independently measured attention to "match", but not to the measured attention to the Bayes profile (participants that report attending to both the color and the urn). For example,

comparing T_B to T_H, attention to (only) the match feature increases from 7.1% to 17.8% ($p < 0.01$), while attention to the Bayesian profile (urn + color) *decreases* from 22.1% to 4.0% ($p<0.01$).[15] Second, we confirm that contrast plays a role in attracting attention to the signal/match feature: the estimated coefficient for contrast, β, is 1.20, with a 95% bootstrap confidence interval of [0.55, 1.80].

6. Additional Implications of Bottom-up Attention

We next derive additional implications of our approach. Section 6.1 shows that bottom-up attention may cause the DM to neglect non-salient hypotheses. Section 6.2 studies the role of the attention limit K, showing that in complex problems it interacts with salience in producing the well-known insensitivity of judgments to sample size (Kahneman Tversky 1972) and to the weight of evidence (Griffin and Tversky 1992). We then present several tests of the model's new predictions.

6.1 Non-Salient Hypotheses: Confirmation Bias and the Gigerenzer-Hoffrage Critique

Nickerson (1998) argues that the confirmation bias, the tendency to interpret data as overly supporting a hypothesis, is often due to the neglect of the alternative hypothesis. A hypochondriac may overreact to mild symptoms because he fails to imagine that such symptoms could also arise with good health. Bottom-up attention accounts for this phenomenon: one hypothesis is salient in the DM's mind, and so is more easily simulated than its alternative. In statistical problems, neglect of a hypothesis can be shaped by its prominence. In our balls and urns experiments the DM is asked "what is the probability that the green ball comes from urn A vs. urn B?" The same question could be phrased as: "what is the probability that the green ball comes from urn A?" The questions are identical but the second phrasing may influence the DM: by leaving urn B implicit, it allows her to neglect B while focusing on A. The DM simulates only the salient hypothesis and does not normalize (Task 3).

[15] In addition, attention only to color also decreases (12.3% vs 6.9%, p=0.02).

To see how this works, denote by $\alpha_B \in \{0,1\}$ the attention to hypothesis H_B. The attention profile is $\alpha = (\alpha_1, .., \alpha_O, \alpha_B)$.[16] When $\alpha_B = 1$ both hypotheses are attended to, which is the case studied so far. When $\alpha_B = 0$, the DM fails to simulate H_B and solves the problem as:

$$\Pr(H_A; \alpha) = \Pr\big(R_\alpha(H_A)\big), \tag{9}$$

setting $\Pr(H_B; \alpha) = 1 - \Pr(H_A; \alpha)$. Equation (9) yields Nickerson's intuition: the DM who neglects H_B forms beliefs by imagining only the focal hypothesis H_A. Bottom-up attention is still determined by Equation (6). The only modification is that $P(\alpha)$ now depends also on the prominence P_B of H_B, and contrast $C(\alpha)$ is computed using (10) whenever H_B is not attended to. The "standard" balls and urns format in which both hypotheses are mentioned has high P_B, whereas the "focal H_A" format in which hypothesis H_B is implicit has low P_B. We then obtain:

Proposition 6 *Moving from a "standard" to a "focal H_A" balls and urns format reduces the Bayes mode and raises the mode at the probability of "A and green",* $\Pr\big(H_A; \alpha_{A \cap g}\big) = \pi_A \cdot q$.

Leaving the alternative H_B implicit reduces the share of correct answers because the Bayes' rule calls for full attention, including to hypotheses. This, in turn, produces two effects. First, it causes a reallocation across the modes of Proposition 4. DMs for whom "urn selection" is salient continue to anchor to the base rate of A, with no effect from whether or not they think about B. DMs for whom "drawing a green ball" is salient think that urn A has q green balls and anchor to q. This is exactly the confirmation bias: these DMs appear to confirm their favored hypothesis A based on its high probability of generating the data, neglecting the fact that the same data could also be generated under B. This is a different logic for reaching the likelihood mode than the one used in Proposition 4. The latter only yields the likelihood of A in a symmetric problem. A DM who instead focuses only on A anchors to its likelihood, regardless of what is in urn B.[17]

[16] In a more cumbersome specification, each hypothesis can have its own attention profile. Neglect of a non-focal H_{-i} can then be formalized as H_{-i} being represented by the feature of being the complement of H_i.

[17] In asymmetric problems, in which $\Pr(g|A) \neq \Pr(b|B)$, neglect of H_B can be detected by DMs' anchoring to the likelihood of A rather than to a combination of the two likelihoods. In particular, neglect of the alternative hypothesis increases the estimated probability of the focal hypothesis iff $\Pr(g|A) > \Pr(g|B)$.

Second, and crucially, the "focal H_A" format creates an entirely "new mode", $\alpha_{A\cap g}$ anchored at $\pi_A \cdot q$. At this mode, which sharply identifies neglect of H_B, the DM attends to both statistical features (the selection of A and the drawing of a green ball from it), and replaces the original question with "what is the probability that a ball is green *and* from A"? These DMs simulate A by computing the joint probability $\pi_A \cdot q$ as in Equation (10), failing to account for B. The deliberate simulation of a specific event further confirms that biases are due to erroneous representations, not to epistemic uncertainty. Remarkably, this use of statistics causes the DM to revise the probability of A below the base rate, despite receiving favorable information! The reason is that the DM fails to appreciate that green balls are even rarer in urn B. To our knowledge, we are the first to unveil this bias despite the fact that in many experiments its incidence is large, as we show next.[18]

We test Proposition 4 by running the "focal H_A" version of the experiment in Section 2. As predicted, making urn B implicit leads to a decrease in the Bayesian mode and a concurrent large increase in the new mode at $\pi_A q = (0.25) * (0.8) = 0.2$.

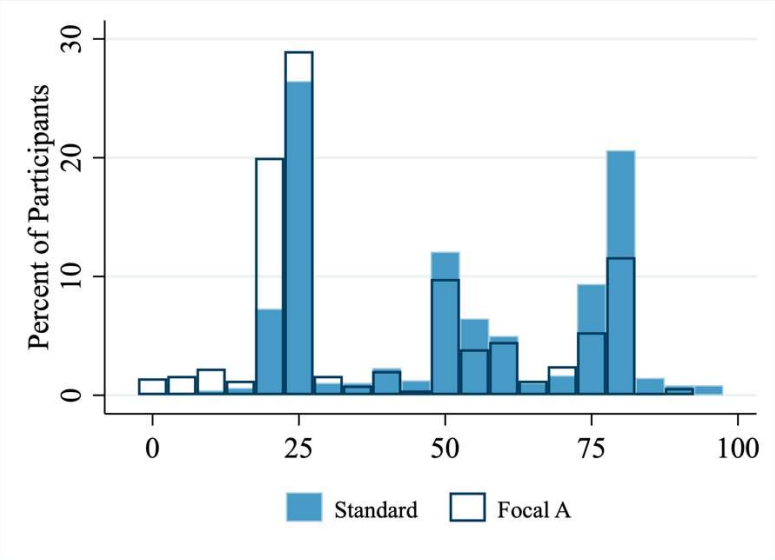

Figure 11. Making one hypothesis implicit reduces Bayesian answers and yields a new mode. Figure shows the distribution of beliefs about $\Pr(A \mid g)$.

[18] Benjamin, Bodoh-Creed, and Rabin (2019) discuss cases in which, after receiving a favorable signal, DM's estimate the probability of the hypothesis below its base rate. In their case the signal and the prior go in the same direction, so anchoring at the likelihood might contribute to this finding.

Keeping the alternative implicit is by all accounts a modest change in description, yet it has a large effect in representations and estimates. The share of subjects anchoring at $\pi_A q = 0.2$ increases from 7.3% to 19.2% ($p < 0.01$). The incidence of this answer is widespread: even in the standard treatment when the alternative hypothesis is explicit. We did not directly elicit attention to hypotheses, but we measure attention to H_B using our free-response attention measure. The share of participants coded as paying attention to the possibility that the drawn marble came from Jar B falls from 49.2% in the standard wording to 39.6% in the Focal A framing (p<0.01).

The new mode is relevant for the debate on base rate neglect. Gigerenzer and Hoffrage (1995, GH) rejected the universality of this and other biases claiming that in "naturalistic contexts" human intuition works well. They showed that better inference can be promoted by describing unconditional frequencies: a share 0.2 of balls are green and in urn A, a share 0.05 are blue and in A, a share 0.15 are green and in urn B, and the remaining share 0.6 are blue balls in B. In this "frequency format" computing the correct answer is easier for it only calls for taking the ratio of 0.2 to 0.15. Our model captures this idea. In this format, in fact, there is a single statistical feature: "drawing a ball from U and of color c", denoted by $f_1 = Uc$ where $c = g, b$, $U = A, B$. The scope for distortions is therefore much reduced: there is no longer anchoring to base rate and likelihoods (which are not mentioned).

GH argue that the efficacy of this format supports the ecological validity of human intuition, since naturalistic contexts expose people to frequencies, not to base rates and likelihoods.[19] This conclusion, however, does not follow from our model. Even in problems with one single statistical feature distortions can arise if people focus on H_A and neglect the alternative hypothesis H_B, or if they focus on ancillary features, phenomena that can both occur in naturalistic settings.

We thus compare the "frequency format", in which both A and B are prominently displayed, to a "focal H_A" frequency format in which H_B is implicit. If exposing people to frequencies is enough

[19] Gigerenzer and Hoffrage's format could also be described as: 25 out of 100 balls are in urn A. Out of those, 20 are blue and 5 are green. The remaining 75 are in urn A. Out of those, 15 are blue and 60 are green. A large body of follow-up work studies how judgments might be improved by designing the communication of statistics and by training people (Visschers et al 2009, Gigerenzer 2014, Operskalski and Barbey 2016).

to promote Bayesian answers, there should be no difference across these versions. If instead bottom-up attention is at play, the new mode should appear in the "focal H_A" frequency format, at the expense of the Bayesian answer. In Appendix B we formally prove this prediction. Figure 12 compares the distribution of answers in balls and urns for the standard and frequency formats in two cases: when the alternative hypothesis is explicit (Panel A) and the "focal H_B" questions where it is not (Panel B).

The results are strongly in line with our prediction. In Panel A, compared to canonical balls and urns, the frequency format sharply increases the mode around the Bayesian answer. This, however, is not due to the fact that the naturalistic frequency format implements Bayesian intuitions.[20] Consider Panel B: as alternative B is made less salient in the "focal H_A" version, the Bayes mode is greatly reduced and the new "$A \cap g$" mode at 20% is strikingly dominant. The benefit of the frequency format over the standard format is no longer clear: many people estimate A to be below its base rate despite the favorable signal.[21]

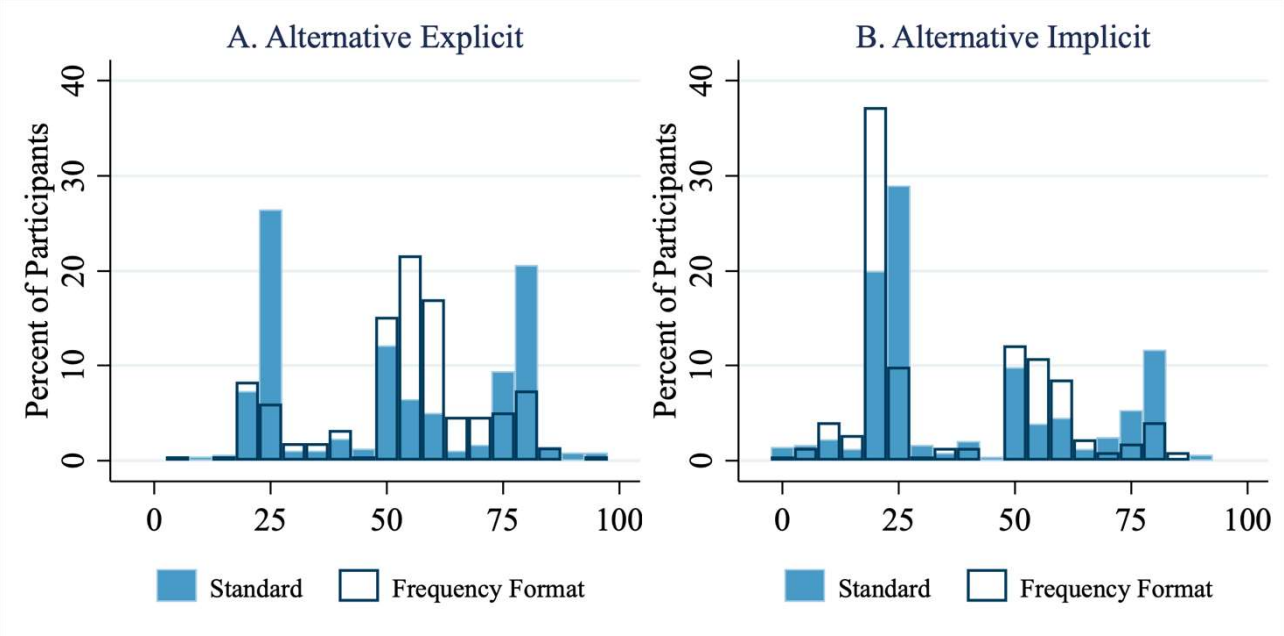

Figure 12. Balls and urns in baseline and frequency formats. Each panel shows the distribution of beliefs about $\Pr(A \mid g)$.

[20] Notably, a number of participants reconstructs the base rate and the likelihood even though they are not saliently presented. Our model can produce this result if DMs may attend to the now ancillary "color" and "urn" features. That is, while the relevant feature is their combination, DMs attention can be spuriously driven to one of them bottom up.

[21] Esponda et al. (2023) show that even the power of experiences is rather weak. They present subjects with standard inference problems where people exhibit base rate neglect (e.g., taxicabs), and then have them repeatedly experience of the joint distribution of signals and true states. Despite the frequentist feedback, many subjects stay anchored to their initial answers. Our model can explain this finding as the byproduct of a stable initial representation of the problem.

Our results suggest that it is too optimistic to expect naturalistic contexts to reduce biases. Bayes rule typically requires attention to many relevant aspects, which may be hard to attain. Psychological work on problem solving is consistent with this view: sometimes naturalistic settings and prior knowledge help, as in solving the Wason task; other times they impair problem solving because people fail to see unusual useful properties of an object, such as in the famous candle problem (Galinsky Moskowitz 2000). Systematically engaging with bottom-up attention, shaped by contrast and prominence, may help design decision architectures conducive to improved judgments.

6.2 Attention limits and Insensitivity in complex problems

Probability estimates are insensitive to the quantity of data. For iid processes, Kahneman and Tversky (1972) and Benjamin, Rabin, and Raymond (2016) document a strong "insensitivity to sample size", whereby estimated sampling distributions fail to converge to the population mean as the sample size grows. Griffin and Tversky (1992) document a strong "insensitivity to the weight of evidence" in inference, where beliefs are insensitive to the number of signals. Our model naturally yields both phenomena: as the sample size/number of signals grows, so do the number of relevant features. Thus, the attention limit K becomes binding, leading to an increasing neglect of data. We show that the interaction of K with salience yields new predictions, which we test.

6.2.1 Insensitivity to Sample Size

Suppose that the DM evaluates the relative likelihood of $H_1 =$ "a sequence of length n has the same number of heads and tails", versus $H_2 =$ "a sequence of length n has only heads". The correct answer is $\Pr(H_1) / \Pr(H_2) = \binom{n}{n/2}$, which increases in n. In experiments, the estimated ratio increases insufficiently (if at all) with n. In our model, the DM's estimate is shaped by the number $r \leq \min(K, n)$ of flips he attends to, captured by attention profile α_r. The latter pins down the

representation $R_{\alpha_r}(H_i)$, which is the union of attended subsequences of length r of the hypothesis' atoms, $\omega \in H_i$.[22] The salience of α_r is additive in the average prominence of its flips $P(\alpha) = P$, contrast $C(\alpha_r)$, and the shock ϵ. In line with our previous assumption, the latter is common to all profiles α in which flips are attended to, so it does not matter here. As we show in Appendix B, contrast increases in r: the more flips the DM attends to, the more she believes that balanced sequences are likelier. This favors rich representations, but the attention limit K may bind. We assume that K is distributed according to a pdf $\pi(K)$ in support $[1, \overline{K}]$. Variations in K across DMs may reflect individual differences in mental faculties, or in situational factors, such as distractions.

Proposition 7 *The average DM underestimates the probability of H_1 vs H_2, the more so when smaller values of K are more likely. As n increases, average beliefs converge to $\bar{\pi}(\overline{K})$.*

Due to attention limits, the DM cannot think about all possible ways of producing balanced sequences for large n. Eventually, beliefs become fully insensitive to n, consistent with KT's finding that people use a "universal distribution" based on a limited number of iid draws. Existing models have wrestled with reconciling the faulty reliance on the law of large numbers in the Gambler's Fallacy with an insufficient reliance on it in large samples (Benjamin, Moore, and Rabin 2017). These phenomena naturally arise in our model: the DM uses a similar representation for the two problems, the class of balanced sequences, whose size however grows insufficiently with n.

As we show in the proof of Proposition 7, our model yields new predictions on the Gambler's Fallacy. First, conditional on committing it, its severity should be higher for DMs who have less severe attention limits, higher K. Second, the average estimated probability of a sequence of n flips and share of heads sh should exhibit insensitivity to the true size of its "share of heads" equivalence class, $\binom{n}{n * sh}$. Intuitively, when the latter becomes larger, it is increasingly difficult – due to attention limits – to simulate its cardinality. Thus, a person focusing on the share of heads will

[22] The ancillary feature shares is relevant in this case but as discussed in Section 3 it does not simplify the estimation process. For simplicity we do not consider it here. Using it is equivalent to hitting the bound $n_\alpha = \min(K, n)$.

estimate the probability of $thth$ to be higher than that of $hhhh$, but less than 6 times, which is the objective ratio of the prevalence of balanced sequences. We can test this prediction using our experiment in Section 4: conditional on a subject committing the Gambler's fallacy, we regress the log of the estimated probability of a sequence on the log of the size of its equivalence class (and on the log of the true probability when we pool different sequence lengths).

Consistent with our prediction, the coefficient on the size of the equivalence class is positive but less than one, showing insensitivity, and is smaller for longer sequences $n = 4,6$ compared to $n = 2$. Thus, bottom-up attention generates three observed behaviors: i) the share of subjects committing the Gambler's Fallacy increases in sequence length n (contrast); furthermore, conditional on committing the fallacy ii) its severity increases with the size of a sequence's equivalence class based on sh (question substitution) but iii) less than proportionally to the latter's size (insensitivity).

	(1) Length 2	(2) Length 4	(3) Length 6	(4) Pooled
Log(Size of Equivalence Class)	0.67***	0.48***	0.43***	0.47***
	(0.04)	(0.02)	(0.02)	(0.05)
Log (Truth)				0.39***
				(0.04)
Constant	-1.26***	-3.48***	-4.89***	-3.51***
	(0.03)	(0.04)	(0.07)	(0.14)
Observations	1128	8528	8016	17672
Individuals	282	533	501	1316
R^2	0.20	0.10	0.06	0.37

Table 5. The dependent variable is the log of the judged probability of each coin-flip sequence of the length indicated in the column heading (pooling all lengths in column 4). Robust standard errors in parentheses. ** and *** indicate significance at the 5% and 1% levels, respectively. Data are restricted to participants for whom judged probabilities and balanced-ness of heads and tails are positively correlated.

6.2.2 Insensitivity to the Weight of Evidence

The same mechanism generates the striking insensitivity to data in inference problems documented by Griffin and Tversky (1992). Consider the inference problem of Section 2 but now allow for multiple draws with replacement from the selected urn. There are $n + 1$ statistical features:

the selected urn, associated with the base rate π_U, and the n draws, each associated with a likelihood. Denote by $D = (n_g, n_b)$ the data, consisting of green and blue balls, $n_g + n_b = n$. The data is favorable to A, $n_g > n_b$, with $\pi_A < 0.5$.

As in Section 3, the DM may neglect drawn balls, focusing only on urn selection, denoted by α_U. He may neglect urn selection and, as in the case of coin flips, attend to $r \leq n$ ball draws, denoted by α_r. Finally, he may attend both to urn selection and to $r \leq n$ draws, denoted by $\alpha_{U,r}$. The salience of each profile is additive in prominence $P(\alpha)$, contrast $C(\alpha)$ and a random shock ϵ_α. As for coin flips, ϵ_α does not depend on the number of draws r. We prove the following result.

Proposition 8 *The average DM is insensitive to the evidence in favor of H_A. Specifically:*

i) He underestimates H_A for sufficiently many green signals $D = (n_g, 0)$, $n_g > n^*$.

ii) The estimate of H_A based on an extra green ball, $D = (N+1, N)$, drops in the number signals N, which also increases attention to urn selection and anchoring to base rates.

Result i) is analogous to insensitivity to sample size: due to capacity constraints, the DM fails to integrate all signals favorable to urn A. The predicted distribution is still multimodal, with some people anchoring at the π_A or the likelihood q (those with $K = 1$) while others integrating more signals and hence yielding more extreme answers, but not to the full extent. Thus, the average estimate is too low compared to what is warranted by the signals. The same mechanism yields, in ii), Griffin and Tversky's insensitivity to the weight of evidence. Relative to a single green signal, adding an equal number of green and blue signals causes the limit K to become binding. This reduces the DM's ability to appreciate that green signals outnumber the blue ones, reducing the contrast associated with the signal. In turn, this increases anchoring to the base rate. This result sharply distinguishes our model from rational inattention. When the DM receives a single green signal, she may anchor to the likelihood, exhibiting a strong overreaction as in Kahneman and Tversky (1972). In contrast, upon receiving many mixed signals he may neglect the data *altogether*, and anchor to the base rate. Instead of being aggregated, different signals *interfere* with one another.

We test these predictions. In the first new treatment, T_{2G}, subjects estimate the probability of A conditional on the draw of two green balls, rather than only one green signal in T_B. Panel A of Figure 12 shows the distribution of beliefs in these two treatments. Consistent with the insensitivity in i), the average response is 52.6% (only 1.4 p.p. higher than in $T_B, p = 0.50$), which exhibits more average under-reaction than when one green ball is drawn. The distribution is also clearly still multimodal, with about 74.1% people anchored at the base rate, the likelihood, and 50:50.

In the second new treatment, T_{5G4B}, we test prediction ii) by harnessing beliefs after 5 green and 4 blue signals, under the same base rate $\pi_A = 0.25$ and the likelihood $q = 0.8$ as T_B. Panel B of Figure 13 compares the resulting distribution of beliefs between T_B and T_{5G4B}. Consistent with prediction ii), the mode at the base rate sharply increases from 26.5% to 39.8%, even though the correct answer is unchanged. In GT's language, increasing the weight and lowering the strength of evidence boosts the share of people who fully neglect the signal in favor of the base rate.[23]

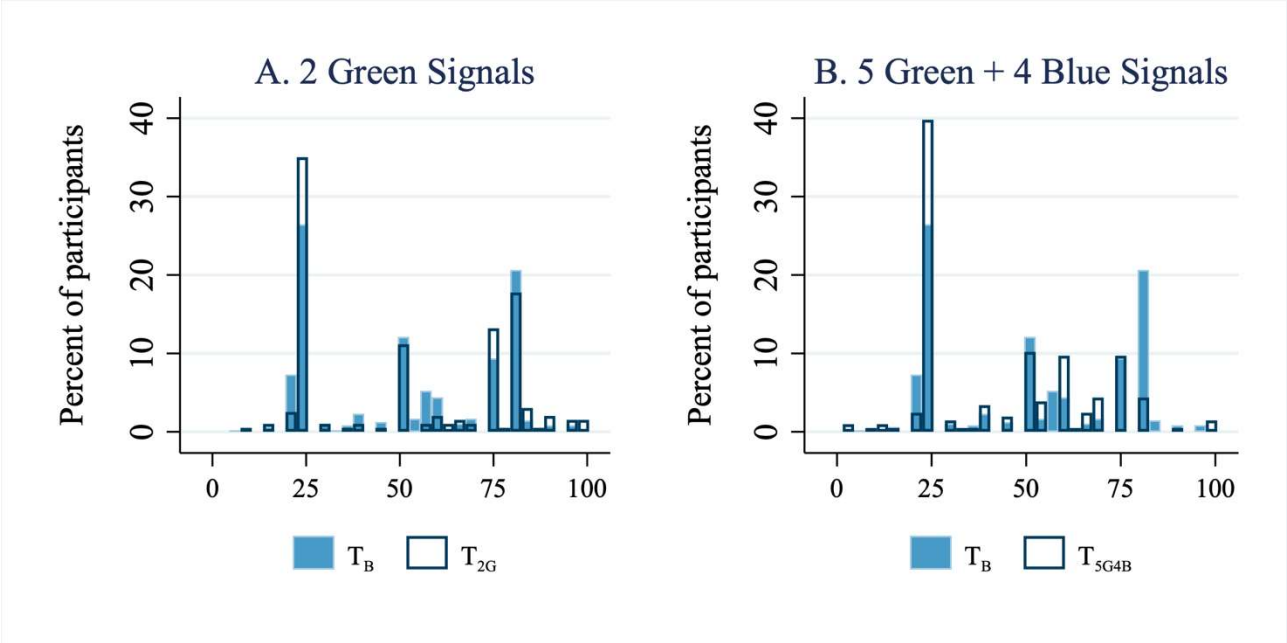

Figure 13. Multiple signals (5 green+ 4 blue and 2 green) in balls-and-urns inference task. Figure shows the distribution of beliefs about the probability of Jar A conditional on the signal(s).

[23] We did not elicit attention to specific numbers and colors of signals, so we cannot test whether treatment effects on measured attention line up with the model. We see, however, that T_{5G4B} increases attention to urn selection, consistent with our mechanism for insensitivity to the weight of evidence.

7. Conclusion

Understanding belief formation is critical to understanding economic behavior. Statistical problems are a very useful laboratory for this enterprise, because they specify a correct answer that can be reached using the information provided. Over the past sixty years, psychologists and behavioral scientists have unveiled many systematic departures of beliefs from the standard Bayesian model (Benjamin 2019), including the Gambler's Fallacy, under-reaction and overreaction in inference, and others. This evidence has led to a proliferation of bias-specific models, reflecting the wide ranging and sometimes contradictory findings. This research has produced important insights but has also opened many doors, leaving a sense that anything goes.

We argued that bias-specific models cannot account for two empirical regularities that we systematically document here: multimodality within a problem and instability across normatively irrelevant variations of the same problem. These phenomena instead reveal a common cognitive structure that helps put many different biases under a common umbrella: bottom-up attention to the features of events. Stylized statistical problems are characterized by multiple features some of which are irrelevant to the problem at hand, but may nevertheless draw attention. Selective attention to features can lead to different distorted representations of the hypothesis, which are in fact different forms of question substitution. This mechanism accounts for many known biases, as well as new ones we document, promising a unified psychological approach to decisions.

Often in the social sciences attention is conceptualized as a scarce resource that is optimally allocated "top down" to further the decision maker's goals. Work on "rational inattention" in economics or the efficient coding approach in psychology follows this approach. Our analysis challenges the ability of this account to capture the cognitive structure of decision biases and their instability. In our experiments all DMs have the same incentives and yet their decisions cluster on different modes and change from one mode to another when goal-irrelevant aspects of the problem are changed. This means that bottom-up attention plays a key role to explaining anomalies, in line with decades of research in psychology showing the importance of bottom up forces for attention. As

we showed in previous work, BGS (2012), bottom-up factors need not operate on wholly irrelevant features. A striking lottery payoff or the striking price of a good (just as a striking statistic in our experiments) may draw attention bottom-up, distracting the decision maker from other equally if not more important goals and relevant features, creating choice instability. An integration of top down and bottom-up attention mechanisms is an important avenue for future work.

In conclusion we describe some important directions for future work. One priority is to integrate the roles of attention and selective memory. In the statistical problems we considered, all relevant data is put in front of subjects. Yet recalled past experiences arguably influence what features they attend to, representations, and estimates. The relevance of a witness statement in court draws attention due to the DM's similar past experiences. Briefly mentioning that a witness is unreliable cues the opposite reaction – we are indeed used to neglecting unreliable data – causing some people to wholly neglect the report's numerical accuracy. Understanding how past experiences in one problem affect which features people recall and attend to in a new problem, is an important ingredient in a theory of prominence and can shed light on why different people represent the same problem in different ways and make different choices. Progress on this front can help understand which narratives or partial models people use in different circumstances, why beliefs can polarize even if there is a great deal of common information, as in the examples of abortion or criminal sentencing, and why learning about a process might be hampered by prominent past experiences (Schwartzstein 2014, Esponda, Vespa, and Yuksel 2022), but also why learning is fast once neglected relevant features are made prominent (Hanna, Mullainathan, and Schwartzstein 2014).

Integrating attention and memory is also important to understand belief formation in naturalistic settings. In these settings, statistics or other numerical information are often unavailable (or anyhow not retrieved or used), and people form beliefs by sampling information from memory. Bordalo, Burro et al. (2022) and Bordalo, Conlon et al (2022) present a model of such sampling based on the psychology of selective recall, and show that it sheds light on several belief anomalies in the field. The approach has also proven fruitful to explain survey data on covid risks, career choices, or

investments (Bordalo, Burro et al. 2022, Conlon and Patel 2023, Jiang et al. 2023). Attention-driven representations can add a crucial ingredient to this theory: which cue in the environment is noticed and acts as an engine for retrieval. For example, the salient losses or failure of an individual bank may draw investors' attention, and cause them to selectively retrieve past episodes of financial meltdown, causing partial neglect of the rarity of cataclysmic events and excessive pessimism.

The combination of memory and bottom-up attention is also relevant for consumer choice. As an example, BGS (2022) offer a theory of consumer choice in which memory and attention interact to shape the perception of the numerical or hedonic magnitude of an attribute, and show that this approach accounts for reference point effects. Our current approach to attention acts at a higher cognitive level, shaping which attributes/features are used to represent choice problems, and which are instead neglected or forgotten. This approach can expand our understanding of the nature, heterogeneity, and instability of decisions made by consumers, investors, voters, etc. Choice options have many features, some relevant/hedonic for a given decision and others ancillary. Some ancillary features can be created artificially or made salient by advertising, and influence decisions by shaping representations. Selective attention to features can create question substitutions of different types. A consumer deciding whether to buy a good may represent the choice as "Is this a fair price?"; an investor considering a firm may represent it as "do I want to invest in a fast growing sector?"; taking a position on a policy can be represented as "am I attached to this party?". The combination of memory and bottom-up attention to features raises the promise of a general theory of intuitive judgments in both naturalistic and abstract settings.

REFERENCES

Behrens, Timothy, Timothy Muller, James Whittington, Shirley Mark, Alon Baram, Kimberly Stachenfeld, and Zeb Kurth-Nelson. "What is a Cognitive Map? Organizing Knowledge for Flexible Behavior," *Neuron* 100 (2018), 490-509.

Benjamin, Daniel, Don Moore, and Matthew Rabin. "Biased Beliefs about Random Samples: Evidence from Two Integrated Experiments." No. w23927. National Bureau of Economic Research, 2017.

Benjamin, Daniel, Matthew Rabin, and Collin Raymond. "A Model of Nonbelief in the Law of Large Numbers," *Journal of the European Economic Association* 14 (2016), 515-544.

Benjamin, Daniel, Aaron Bodoh-Creed, and Matthew Rabin "Base Rate Neglect: Foundations and Implications", mimeo, 2019.

Benjamin, Daniel. "Errors in Probabilistic Reasoning and Judgment Biases," *Handbook of Behavioral Economics: Applications and Foundations 1* (2019), 69-186.

Bordalo, Pedro, Giovani Burro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. "Imagining the Future: Memory, Simulation, and Beliefs about Covid," Working paper (2022).

Bordalo, Pedro, John Conlon, Nicola Gennaioli, Spencer Kwon, and Andrei Shleifer. "Memory and Probability." *Quarterly Journal of Economics* 138 (2023), 265-311.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. "Salience Theory of Choice under Risk," *Quarterly Journal of Economics* 127 (2012), 1243-1285.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. "Salience and Consumer Choice." *Journal of Political Economy* 121 (2013), 803-843.

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

Chetty, Raj, Adam Looney, and Kory Kroft. "Salience and Taxation: Theory and evidence." *American Economic Review* 99 (2009), 1145-1177.

Clancy, Kevin, John Bartolomeo, David Richardson, and Charles Wellford. "Sentence Decisionmaking: The Logic of Sentence Decisions and the Sources of Sentence Disparity." *Journal of Criminal Law and Criminology* 72 (1981), 524-554.

Conlon, John J., and Dev Patel. "What Jobs Come to Mind? Stereotypes about Fields of Study." Working paper.

Dohmen, Thomas, Armin Falk, David Huffman, Felix Marklein, and Uwe Sunde. "The Non-Use of Bayes Rule: Representative Evidence on Bounded Rationality," working paper (2009).

Edwards, Ward. "Conservatism in human information processing." In D. Kahneman, P. Slovic, and A. Tversky (Eds.), *Judgment under Uncertainty: Heuristics and Biases* (pp. 359-369). Cambridge: Cambridge University Press (1968).

Enke, Ben, and Thomas Graeber. "Cognitive Uncertainty," *Quarterly Journal Economics*, forthcoming (2023).

Enke, Benjamin, Thomas Graeber, and Ryan Oprea. "Complexity and Time," working paper (2023).

Esponda, Ignacio, Emanuel Vespa, and Sevgi Yuksel. "Mental Models and Learning: The Case of Base-Rate Neglect," working paper (2022).

Evers, Ellen, Alex Imas, Christy Kang. "On the Role of Similarity in Mental Accounting and Hedonic Editing." *Psychological Review* 129 (2021), 777-789.

Gabaix, Xavier. "A Sparsity-based Model of Bounded Rationality." *Quarterly Journal of Economics* 129 (2014), 1661-1710.

Gabaix, Xavier. "Behavioral Inattention." In *Handbook of Behavioral Economics: Applications and Foundations 1*, vol. 2 (2019), 261-343. North-Holland.

Galinsky, Adam, and Gordon Moskowitz. "Counterfactuals as Behavioral Primes: Priming the Simulation Heuristic and Consideration of Alternatives." *Journal of Experimental Social Psychology* 36 (2000), 384-409.

Gigerenzer, Gerd. "On Narrow Norms and Vague Heuristics: A reply to Kahneman and Tversky." *Psychological Review* 3 (1996): 592-596.
Gigerenzer, Gerd. *Risk Savvy: How to Make Good Decisions*. New York: Viking, 2014.
Gigerenzer, Gerd, and Ulrich Hoffrage. "How to Improve Bayesian Reasoning Without Instruction: Frequency Formats," *Psychological Review* 102 (1995): 684-704.
Grether, David. "Bayes Rule as a Descriptive Model: The Representativeness Heuristic," *Quarterly Journal of Economics* 95 (1980), 537-557.
Griffin, Dale, and Amos Tversky. "The Weighing of Evidence and the Determinants of Confidence," *Cognitive Psychology* 24 (1992), 411-435.
Guyon, Isabelle, and André Elisseeff. "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research* 3 (2003), 1157-1182.
Hanna, Rema, Sendhil Mullainathan, and Joshua Schwartzstein. "Learning through Noticing: Theory and Evidence from a Field Experiment." *Quarterly Journal of Economics* 129 (2014): 1311-1353.
Jiang, Zhengyang, Hongqi Liu, Cameron Peng, and Hongjun Yan. "Investor Memory and Biased Beliefs: Evidence from the Field," Working Paper 2023.
Kahneman, Daniel, and Shane Fredrick. "Representativeness Revisited: Attribute Substitution in Intuitive Judgment," *Heuristics and biases: The psychology of intuitive judgment,* 49 (2002), 81.
Kahneman, Daniel, and Amos Tversky. "Subjective Probability: A Judgment of Representativeness," *Cognitive Psychology* 3 (1972), 430-454.
Khaw, Mel Win, Ziang Li, and Michael Woodford. "Cognitive Imprecision and Small-stakes Risk Aversion." *Review of Economic Studies* 88 (2021), 1979-2013.
Kleinberg, Jon, Annie Liang, and Sendhil Mullainathan. "The Theory is Predictive, but is it Complete? An Application to Human Perception of Randomness." In *Proceedings of the 2017 ACM Conference on Economics and Computation*, 125-126.
Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. "Human Decisions and Machine Predictions." *Quarterly Journal of Economics* 133 (2018), 237-293.
Kruschke, J. K. (2008). "Models of Categorization." In *The Cambridge Handbook of Computational Psychology*.
Li, Xiaomin and Colin F. Camerer. "Predictable Effects of Visual Salience in Experimental Decisions and Games." *Quarterly Journal of Economics* 137 (2022), 1849-1900.
Ludwig, Jens, and Sendhil Mullainathan. "Machine Learning as a Tool for Hypothesis Generation," NBER w31017 (2023).
Nickerson, Raymond. "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises," *Review of General Psychology* 2 (1998), 175-220.
Nosofsky, Robert. "Similarity, Frequency, and Category Representations," *Journal of Experimental Psychology: learning, memory, and cognition,* 14 (1988), 54-65.
Operskalski, Joachim, and Aron Barbey. "Risk Literacy in Medical Decision-Making." *Science* 352 6284 (2016), 413-414.
Rabin, Matthew. "Inference by Believers in the Law of Small Numbers," *Quarterly Journal of Economics* 117 (2002), 775-816.
Reutskaja, Elena, Rosemarie Nagel, Colin Camerer, and Antonio Rangel. "Search Dynamics in Consumer Choice under Time Pressure: An Eye-tracking Study." *American Economic Review* 101 (2011), 900-926.
Schwartzstein, Joshua. "Selective Attention and Learning," *Journal of the European Economic Association*, 12 (2014), 1423-1452.
Selfridge, Oliver. "Pattern Recognition and Modern Computers," in *Proceedings of the March 1-3, 1955, Western Joint Computer Conference*, (1955), 91-93.

Sims, Christopher A. “Implications of Rational Inattention.” *Journal of Monetary Economics*, 50 (2003), 665-690.

Slovic, Paul, Howard Kunreuther, and Gilbert White, “Decision Processes, Rationality, and Adjustment to Natural Hazards: A Review of Some Hypotheses,” in *Natural Hazards, Local, National and Global*, Gilbert White, ed. (Oxford: Oxford University Press, 1974), 39–69.

Thaler, Richard. "Mental Accounting and Consumer Choice." *Marketing Science* 4 (1985), 199-214.

Tversky, Amos. “Features of Similarity,” *Psychological Review* 84 (1977), 327-352.

Tversky, Amos, and Itamar Gati. "Similarity, Separability, and the Triangle Inequality." *Psychological Review* 89 (1982): 123-154.

Woodford, Michael. “Imperfect Common Knowledge and the Effects of Monetary Policy,” in P. Aghion, R. Frydman, J. Stiglitz, and M. Woodford, eds., *Knowledge, Information and Expectations in Modern Macroeconomics*, Princeton: Princeton University Press, 2003.

Woodford, Michael. “Modeling Imprecision in Perception, Valuation, and Choice.” *Annual Review of Economics 12* (2020), 579-601.

Visschers, Vivianne, Ree Meertens, Wim Passchier, and Nanne De Vries. "Probability Information in Risk Communication: a Review of the Research Literature." *Risk Analysis: An International Journal* 29 (2009), 267-287.

APPENDIX

Appendix 1. Proofs

Proof of Proposition 1. To find a rational simplification $\alpha^*(\omega)$, take a coarsest partition of a generic hypothesis H_i in terms of the events to which statistics are assigned. Such partition can be expressed as a collection of unconditional events $K_1(H_i)$, of step 2 events $K_2(H_i)$, and of step n events $K_n(H_i)$, such that the total probability of the hypothesis can be computed as:

$$\Pr(H_i) = \sum_{k_1 \in K_1(H_i)} \pi_{k_1} + \sum_{k_2|k_1 \in K_2(H_i)} \pi_{k_2|k_1} \cdot \pi_{k_1} + \sum_{k_n|k_{n-1},\dots k_1 \in K_n(H_i)} \pi_{k_n|k_{n-1}\dots k_1} \cdot \pi_{k_{n-1}|k_{n-2}\dots k_1} \cdot \dots \cdot \pi_{k_1}. \quad (A.1)$$

At one extreme, the hypothesis can be expressed as a collection of unconditional events only (i.e. only the first, step 1 sum is operative), at the other extreme the finest and coarsest partitions coincide, and only the last, step n sum is operative).

Take a generic event $A \in K_j(H_i)$. Individual elements $\omega \in A$ all share the jth feature. Set this and the preceding features as relevant for them, $\alpha_l^*(\omega) = 1$, $l \leq j$, and all subsequent features as irrelevant $\alpha_l^*(\omega) = 0$, $l > j$, including the ancillary one. Repeat this process for all other events in $K_j(H_i)$, and for all $j = 1, \dots, n$. This yields vector of relevant features $\alpha^*(\omega)$ for each $\omega \in H_i$. The ensuing representation $R_{\alpha^*}(H_i) = \bigcup_{\omega \in H_i} \tilde{f}_{\alpha^*(\omega)}(\omega)$ is simplified: it prunes the ancillary feature, possibly additional ones. Multiplication of the statistics attached to relevant features according to Equation (2) yields the correct chain product probability in Equation (A.1): for each conditional event $A \in K_j(H_i)$ the statistic $\pi_{k_j|k_{j-1}\dots k_1}$ assigned to it is multiplied by the statistics assigned to all conditional events and the conditional event preceding it.

Proof of Proposition 2. Given $\to \infty$, feasible profiles are $A_\infty \equiv \{0,1\}^n \times \{0\} \cup (0,0,\dots,1)$. Consider a DM attending to $r \leq n$ flips (statistical features), denoted by α_r. She simulates the hypothesis-sequence H_i as $\Pr\left(R_{\alpha_r}(H_i)\right) = 2^{-r}$. After normalization, this yields $\Pr(H_i; \alpha_r) = 0.5$, $r \leq n$. By (4) and (5), contrast is zero, $C(\alpha_r) = 0$, prominence is $P(\alpha_r) = r * P_F/r = P_F$. The salience of α_r is $P_F + \epsilon_F$, where ϵ_F is the random prominence of (any number of) flips.

A DM attending to the ancillary feature, sh, profile $\alpha_S = (0,\dots,1)$, simulates the balanced hypothesis-sequence H_1 as $\Pr\left(R_{\alpha_S}(H_1)\right) = \binom{n}{n/2} \cdot 2^{-n}$ and the unbalanced one H_2 as $\Pr\left(R_{\alpha_S}(H_2)\right) = 2^{-n}$. The estimated relative likelihood of the unbalanced in (7) easily follows. By Equation (4), contrast is $C(\alpha_S) = \frac{\binom{n}{n/2}-1}{\binom{n}{n/2}+1}$. Given that $P_S = 0$ the salience of α_S is $C(\alpha_S) + \epsilon_S$. Given the extreme value distribution, the share of DMs attending to sh is:

$$\mu(\alpha_S) = \frac{e^{C(\alpha_S)-P_F}}{e^{C(\alpha_S)-P_F}+1}, \quad (A.2)$$

While other DMs attend to individual flips and issue a 50:50 judgment.

Proof of Corollary 3 Inspection of (A.4) immediately yields that $\mu(\alpha_S)$ is increasing in n and decreasing in P_F, as desired.

Proof of Proposition 4 There are five possible attention profiles $\alpha = (\alpha_U, \alpha_{c|U}, \alpha_m)$, whose implications for representations and estimates were discussed in the text. Consider the salience of these profiles. Under full attention to statistical features $\alpha_\beta = (1,1,0)$ the DM reaches the Bayesian estimate, which yields contrast $C(\alpha_\beta) = |2\beta - 1|$, prominence is $P(\alpha_\beta) = (P_U + P_{c|U})/2$, so the salience of α_β is $C(\alpha_\beta) + P(\alpha_\beta) + \epsilon_\beta$, where ϵ_β is the extreme value shock.

If the DM attends only to urn selection, $\alpha_{BR} = (1,0,0)$, she estimates $\Pr(H_U; \alpha_{BR}) = \pi_U$. Contrast is $C(\alpha_{BR}) = |2\pi_B - 1|$, prominence is $P(\alpha_{BR}) = P_U$, giving salience $C(\alpha_{BR}) + P(\alpha_{BR}) + \epsilon_{BR}$. If the DM attends to the color of the ball only, $\alpha_c = (0,1,0)$, symmetry in urn compositions imply that she estimates $\Pr(H_A; \alpha_c) = q$. Contrast is $C(\alpha_c) = |2q - 1|$, prominence is $P(\alpha_c) = P_{c|U}$, and salience is $C(\alpha_c) + P(\alpha_c) + \epsilon_c$. If the DM attends to the match feature, $\alpha_m = (0,01)$, she also reaches $\Pr(H_A; \alpha_m) = q$. Contrast is again $C(\alpha_m) = |2q - 1|$, prominence is $P(\alpha_m) = P_m$, so the salience of α_m is $C(\alpha_m) + P(\alpha_m) + \epsilon_m$. Finally, under attention to nothing, $\alpha_0 = (0,0,0)$, we have $\Pr(H_A; \alpha_0) = 0.5$. Contrast under this attention profile is $C(\alpha_0) = 0$, prominence is also normalized to zero, so salience is ϵ_0. By the extreme value distribution of shocks, the share of DMs with attention α_j is:

$$\mu(\alpha_j) = \frac{e^{C(\alpha_j)+P(\alpha_j)}}{\sum_{j'} e^{C(\alpha_{j'})+P(\alpha_{j'})}} \qquad j = \beta, BR, c, m, 0. \tag{A.3}$$

Proof of Corollary 5 Using (A.3), the share of DMs at the likelihood or at the Bayes answer relative to the share at the base rate are respectively equal to:

$$\frac{\mu(\alpha_c) + \mu(\alpha_m)}{\mu(\alpha_{BR})} = e^{|2q-1|-|2\pi_B-1|+(P_{c|U}-P_U)+P_{c|U}+P_m} + e^{|2q-1|-|2\pi_B-1|+(P_m-P_U)}, \tag{A.4}$$

$$\frac{\mu(\alpha_\beta)}{\mu(\alpha_{BR})} = e^{\left|2\frac{\pi_A q}{\pi_A q+\pi_B(1-q)}-1\right|-|2\pi_B-1|+\frac{1}{2}(P_{c|U}-P_U)}. \tag{A.5}$$

It is immediate to see that, because $q > 1/2$ and the Bayesian estimate is larger than 0.5, both (A.4) and (A.5) increase in q (and decrease in $\pi_B > 0.5$). The two modes also increase in $P_{c|U}$ while only the likelihood mode increases in P_m.

Proof of Proposition 6 Now the attention profile is $\alpha = (\alpha_U, \alpha_{c|U}, \alpha_m, \alpha_B)$. The contrast of the modes in Proposition 4 is unchanged, only prominence changes. For the likelihood mode, it becomes $\frac{P_{c|U}+\alpha_B P_B}{1+\alpha_B}$ or $\frac{P_m+\alpha_B P_B}{1+\alpha_B}$ depending on whether the DM attends to color or match, for the base rate mode it becomes $\frac{P_U+\alpha_B P_B}{1+\alpha_B}$, and for the Bayesian mode it becomes $\frac{P_{c|U}+P_U+P_B}{3}$. The new mode features contrast $C(\alpha_{A\cap g}) = |2\pi_A q - 1|$ and prominence $\frac{P_{c|U}+P_U}{2}$. The ratio of DMs at the Bayes relative to the new mode is then equal to:

$$\frac{\mu(\alpha_\beta)}{\mu(\alpha_{A\cap g})} = e^{\left|2\frac{\pi_A q}{\pi_A q+\pi_B(1-q)}-1\right|-|2\pi_A q-1|-\frac{P_{c|U}+P_U}{6}+\frac{P_B}{3}}, \tag{A.6}$$

so that, evidently, lower prominence P_B of the alternative hypothesis reduces the incidence of the Bayes mode relative to the new mode. By similar arguments, it is immediate to see that lower P_B reduces also the incidence of α_β and increases the incidence of $\alpha_{A\cap g}$ compared to all other modes.

The prediction for the frequency format experiment easily follows. There is one statistical feature. If the DM attends to it and to H_B, the answer is Bayesian. If she attends only to the statistical feature the answer is 20%. The difference in contrast is as in (A.6), that in prominence is $(P_B - P_S)/2$, where P_S is the statistical feature's prominence. The relative prevalence of the Bayes mode increases in P_B. If the DM attends to the ancillary feature, she anchors to 80% regardless of whether H_B is attended to or not. In line with our previous analysis, then, lower P_B decreases the incidence of α_β and increases the incidence of $\alpha_{A\cap g}$ compared to all other modes.

Proof of Proposition 7 The hypotheses now are H_1 = "share of heads is 0.5" and H_2 ="share of heads is 1". Consider first a DM without attention limits. A DM attending to $r \leq n$ flips (wlog the first r flips), represents H_2 as a unique sequence of r heads, and H_1 as the union of all its subsequences of length r. Denoted attention by α_r, we have that for $r \leq n/2$, $\left|R_{\alpha_r}(H_1)\right| = 2^r$: each

sequence of length r is a subsequence for a suitably chosen balanced sequence. For $r > n/2$, the sequence is in $R_{\alpha_r}(H_1)$ if and only if it has at least $r - n/2$ heads. Thus, we obtain:

$$\left|R_{\alpha_r}(H_1)\right| = \begin{cases} 2^r & if \quad r \le n/2 \\ \sum_{s=r-n/2}^{n/2} \binom{r}{s} & if \quad r > n/2 \end{cases}. \tag{A.7}$$

Given that each sequence in $R_{\alpha_r}(H_1)$ is simulated as 2^{-r} and that $R_{\alpha_r}(H_2)$ is also simulated by 2^{-r}, the probability estimate is under profile α_r is

$$\Pr(H_1; \alpha_r) = \frac{\left|R_{\alpha_r}(H_1)\right|}{1 + \left|R_{\alpha_r}(H_1)\right|}. \tag{A.8}$$

Consider salience now. Attention to the ancillary feature is equivalent to profile α_n, so for simplicity we do not separately consider the ancillary feature here. The contrast of α_r is equal to:

$$C(\alpha_r) = \frac{\left|R_{\alpha_r}(H_1)\right| - 1}{1 + \left|R_{\alpha_r}(H_1)\right|}, \tag{A.9}$$

prominence is again P_F, so that salience is $C(\alpha_r) + P_F + \epsilon_F$. Note that, holding sequence length n fixed, $C(\alpha_r)$ is strictly increasing in r for $r < n - 1$. To see this, note that there is a natural surjection from $R_{\alpha_{r+1}}(H_1)$ to $R_{\alpha_r}(H_1)$ that simply drops the last element of the sequence. The surjection is strict for $r < n - 1$. Thus, when estimating H_1 vs H_2, a DM whose attention is unconstrained will always choose the richest representation, setting $r = n$.

Consider attention limits. By the previous arguments, a DM with limit K will attain the richest representation $r = K$. Given the distribution π_K of K, the average estimate will be equal to:

$$\pi(H_1; n) \equiv \sum_{K \ge 1} \Pr(H_1; \alpha_K)\, \pi_K, \tag{A.10}$$

Let us now characterize $\Pr(H_1; \alpha_K)$. From the fact that $C(\alpha_r)$ is strictly increasing in r, we immediately obtain $\Pr(H_1; \alpha_K) < \Pr(H_1; \alpha_n) = \frac{\binom{n}{n/2}}{\binom{n}{n/2}+1}$. As $n > 2\overline{K}$ all $\Pr(H_1; \alpha_K)$ converge to $\frac{2^K}{2^K+1}$, with mean beliefs converging to $\bar{\pi}(\overline{K}) = \sum_{K \le \overline{K}} \pi(K) \cdot \frac{2^K}{2^K+1}$, which is fully insensitive to n.

Consider next the implication of the attention limit for the Gambler's Fallacy when hypothesis H_1 is a specific balanced sequence. If the DM attends to individual flips, then attention limits do not matter and she correctly estimates $\Pr(H_1; \alpha_r) = 0.5$. If she attends to the share of heads, she instead computes it as $\Pr(H_1 = share\ of\ heads\ is\ 0.5; \alpha_K)$, using the Equation in (A.8). Thus, there is now a distribution of people, some of which exhibit stronger gambler's fallacy than others (those with larger K), and on average the fallacy is insensitive to n, following (A.10).

Proof of Proposition 8. Again, consider first the case $K \to \infty$. The possible attention allocations are: i) urn selection, α_U, ii) urn selection and $r \le n$ draws, $\alpha_{U,r}$, and iii) $r \le n$ draws, α_r. Again, for simplicity we abstract from the ancillary feature. At profile α_U, the DM behaves as in Proposition 3, and salience is equal to $C(\alpha_{BR}) + P_U + \epsilon_U$. At attention profile α_r, using the same logic of Proposition 7, one obtains that simulation is equal to:

$$\Pr\left(R_{\alpha_r}(H_U)\right) = \begin{cases} 2^r & if \quad r \le n_b \\ \sum_{s=r-n_b}^{r} \binom{r}{s} q_U^s (1-q_U)^{r-s} & if \quad n_b < r \le n_g \\ \sum_{s=r-n_b}^{n_g} \binom{r}{s} q_U^s (1-q_U)^{n_\alpha - s} & if \quad r > n_g \end{cases},$$

where q_U is the share of green balls in urn U, so $q_A = q = 1 - q_B$. The estimate at attention α_r is $\Pr(H_A;\alpha_r) = \Pr\left(R_{\alpha_r}(H_A)\right) / \left[\Pr\left(R_{\alpha_r}(H_A)\right) + \Pr\left(R_{\alpha_r}(H_B)\right)\right]$. Given $q > 0.5$, contrast is:

$$C(\alpha_r) = 2 \cdot \Pr(H_A;\alpha_r) - 1, \qquad (A.11)$$

prominence is $P(\alpha_r) = P_{c|U}$, and salience is $C(\alpha_r) + P(\alpha_r) + \epsilon_{c|U}$, where $\epsilon_{c|U}$ is the shock to the drawing of colored balls features. Finally, a DM at attention profile $\alpha_{U,r}$ simulates hypotheses by:

$$\Pr\left(R_{\alpha_{U,r}}(H_U)\right) = \begin{cases} \pi_U 2^r & if \quad r \le n_b \\ \pi_U \cdot \sum_{s=r-n_b}^{r} \binom{r}{s} q_U^s (1-q_U)^{r-s} & if \quad n_b < r \le n_g \\ \pi_U \cdot \sum_{s=r-n_b}^{n_g} \binom{r}{s} q_U^s (1-q_U)^{n_\alpha - s} & if \quad r > n_g \end{cases},$$

where q_U is the share of green balls in urn U, so $q_A = q = 1 - q_B$. The estimate at profile $\alpha_{U,r}$ is $\Pr(H_A;\alpha_{U,r}) = \Pr\left(R_{\alpha_{U,r}}(H_A)\right) / \left[\Pr\left(R_{\alpha_{U,r}}(H_A)\right) + \Pr\left(R_{\alpha_{U,r}}(H_B)\right)\right]$. Given $q > 0.5$, contrast is:

$$C(\alpha_{U,r}) = 2 \cdot \Pr(H_A;\alpha_{U,r}) - 1, \qquad (A.12)$$

prominence is $P(\alpha_{U,r}) = (P_U + rP_{c|U})/(r+1)$, so salience is $C(\alpha_r) + P(\alpha_{U,r}) + \epsilon_{U,c|U}$, where $\epsilon_{U,c|U}$ is the extreme value shock to the salience of both kinds of statistical features.

There is a large set of possible attention profiles, but the following result, which we prove in lemma B, is useful: for a DM who considers only ball draws, $r = n$ maximizes contrast and hence salience. For a DM who considers urn selection and ball draws, salience is maximized at $1 \le r^* \le n$ which may be $r^* < n$ if $P_{c|U} < P_U$. The attention allocation for an unconstrained DM is the most salient one between α_U, α_{U,r^*} and α_n. For a DM with constraint K, the final attention allocation is the most salient one between α_U, $\alpha_{U,\hat{r}}$ and α_K, where $\hat{r} = \min[K-1, r^*]$.

Let us characterize these beliefs (and the relative mass at the beliefs) for the two cases described in Proposition 8. Case 1: $(n_g, n_b) = (n_g, 0)$. Estimates entail the likelihood ratios:

$$\frac{\pi_A}{\pi_B}, \qquad \left(\frac{q}{1-q}\right)^K, \qquad \frac{\pi_A}{\pi_B} \cdot \left(\frac{q}{1-q}\right)^{\hat{r}}, \qquad 1$$

Which yield the following salience (minus the stochastic factor):

$$P_U + (2\pi_B - 1), \qquad P_{c|U} + \frac{\left(\frac{q}{1-q}\right)^K - 1}{\left(\frac{q}{1-q}\right)^K + 1}, \qquad \frac{\hat{r} \cdot P_{c|U} + P_U}{\hat{r}+1} + \frac{\frac{\pi_A}{\pi_B} \cdot \left(\frac{q}{1-q}\right)^{\hat{r}} - 1}{\frac{\pi_A}{\pi_B} \cdot \left(\frac{q}{1-q}\right)^{\hat{r}} + 1}, \qquad 0$$

Note that, regardless of the profile chosen, estimates undershoot the Bayesian answer if

$$\max\left\{\frac{\pi_A}{\pi_B}, \quad \left(\frac{q}{1-q}\right)^K, \quad \frac{\pi_A}{\pi_B} \cdot \left(\frac{q}{1-q}\right)^{\hat{r}}\right\} < \frac{\pi_A}{\pi_B} \cdot \left(\frac{q}{1-q}\right)^n,$$

Given that $\pi_A < \pi_B$ this occurs if $\left(\frac{q}{1-q}\right)^{n-K} > \frac{\pi_B}{\pi_A}$, which immediately yields the first statement: by assumption, with a single signal the Bayesian answer is above 0.5, $\pi_B(1-q) < \pi_A q$. Thus, as long as some agents are constrained, there is under-reaction on average.

Case 2: $(n_g, n_b) = (N+1, N)$. We show that if attention limits are sufficiently strong, there is insensitivity to data. Consider a DM for which $K < n_b = N$. This DM simulates $\Pr\left(R_{\alpha_r}(H_U)\right) = 2^r$, under profile α_r and $\Pr\left(R_{\alpha_{U,r}}(H_U)\right) = \pi_U \cdot 2^r$ under profile $\alpha_{U,r}$. Critically, this DM is fully insensitivity to the color of signals, so her modes are entirely given by π_A/π_B and by 1. For DMs with such attention limit these modes respectively occur with (un-normalized) probability

$\exp\big(P_U + (2\pi_B - 1)\big) + \exp\Big(P_{U,c|U} + (2\pi_B - 1)\Big), \exp\left(P_{c|U}\right) + 1$. In this expression we have $P_{U,c|U} = \big(r \cdot P_{c|U} + P_U\big)/(r+1)$.

We studied before the single green signal $\big(n_g, n_b\big) = (1,0)$. With attention limits, if $K = 1$, the DM chooses between the base rate and the likelihood, if $K > 1$ she behaves as in Proposition 3. This implies that, for any K, the mass at the base rate for $\big(n_g, n_b\big) = (N+1, N)$ is given by:

$$\frac{\exp\big(P_U + (2\pi_B - 1)\big) + \exp\Big(P_{U,c|U} + (2\pi_B - 1)\Big)}{\exp\big(P_U + (2\pi_B - 1)\big) + \exp\Big((P_{c|U} + P_U)/2 \; + (2\pi_B - 1)\Big) + \exp\left(P_{c|U}\right) + 1}$$

And the mass at the base rate for $\big(n_g, n_b\big) = (1,0)$ is given by:

$$\frac{\exp\big(P_U + (2\pi_B - 1)\big)}{\exp\big(P_U + (2\pi_B - 1)\big) + \exp\left((P_{c|U} + P_U)/2 + \dfrac{\frac{\pi_A}{\pi_B} \cdot \left(\frac{q}{1-q}\right) - 1}{\frac{\pi_A}{\pi_B} \cdot \left(\frac{q}{1-q}\right) + 1}\right) + \exp\Big(P_{c|U} + (2q - 1)\Big) + 1}$$

Thus, the conclusion that there is a greater mass at the base rate follows from the fact that the numerator of the first fraction is bigger than that of the second fraction, and the denominator is bigger for the second fraction (if the Bayesian answer has less contrast than the base rate). Regarding mean beliefs, the same argument implies that there is a greater mass also at 50-50 for $\big(n_g, n_b\big) = (N+1, N)$ than $\big(n_g, n_b\big) = (1,0)$. Lastly, only $\big(n_g, n_b\big) = (1,0)$ has a positive mass at the Bayesian answer $\beta > 0.5$ (by assumption) and the likelihood q, which proves that mean beliefs are strictly greater under $\big(n_g, n_b\big) = (1,0)$ as desired.

Appendix B: Data

This appendix describes the experiments in greater detail.

Recruitment and logistics. Participants were recruited through Prolific and had to be at least 18 years old, reside in the US or UK, and have previously submitted at least 50 other studies on prolific with at least a 95% approval rate. The sample we recruited was balanced on gender (as Prolific skews female). The study was described to potential participants simply as a "Short survey for laptop or desktop computer" with the longer description reading "This quick survey should take around 15 minutes and is part of a research study. Note: you must use a laptop or desktop computer to take the survey (mobile devices will not work)." Participants received a $2.50 bonus for completing the survey, plus any bonus they earned. A total of 4,799 participants completed the survey, one fewer than our preregistered sample size (because we drop the second submission of one participant who took the survey twice).

Preregistration. The study was preregistered on the AEA RCT registry (ID AEARCTR-0011166).

Links with more information. This online document contains screenshots of the questions for each of the inference and gamblers'-fallacy treatments, as well as of the attention self-report questions. The survey itself can be accessed at this link.

In the main text, we analyze a measure of attention derived from free-response questions where participants describe in their own words how they solved the inference and gambler's fallacy problems. We code these responses by querying GPT 3.5, prompting it with yes-no questions about whether the response appears to indicate that the participant was paying attention to various features. This online document lists these prompts.

Intended and actual sample sizes. The tables below lists each treatment, including some not mentioned in the main text (but described in greater detail below), along with the intended and actual sample size for each. Differences between intended and actual sample size are due to chance.

Treatment	Intended sample size	Actual sample size
Standard balls and urns	500	480
Standard taxicabs	200	199
T_1: Undermining witness	200	196
T_2: Highlighting match	200	202
T_{LE}: Less extreme likelihood	500	497
T_{ME}: More extreme likelihood	500	487
5 Green 4 Blue Signals	200	206
2 Green Signals	200	197
1 Green 1 Irrelevant Signals	500	480
1 Green 0 Irrelevant Signals	500	525
Focal A	500	490
T_S	200	193
T_L	200	207
Frequency format	200	217
Frequency format (Focal A)	200	223

Table A1. Treatment groups and sample sizes for inference questions.

Treatment	Intended sample size	Actual sample size
th vs *hh*	400	434
ththht vs *hhhhhh*	400	405
T_{full}	1000	1038
T_{last}	1000	978
$T_{control}$	1000	971
T_{share}	1000	973

Table A2. Treatment groups and sample sizes for gambler's fallacy questions.

Statistics as frequencies rather than conditional probabilities.

We now describe a test in which the contrast of the ball's color, in a balls and urns inference question, increases but the correct answer stays the same. To this end, we describe urns using absolute rather than relative frequencies. Specifically, urns A and B are selected with probability 0.5. Urn A contains 5 green and 5 blue balls. Urn B contains only green balls, but their absolute number varies across treatments. In treatment T_S urn B is "small", holding 5 green balls. In T_L urn B is "large", containing 15 green balls. In both treatments, the correct answer is $\Pr(A|g) = 1/3$.

In this format, the event space is $\Omega = \{(A, g), (A, b), (B, g)\}$. As we hinted in Section 3, though, now the generic event $\omega = (U, c)$ has three statistical features, not two. The first is urn selection, $U = A, B$, linked to the 0.5 base rate. The second is the "number of ways in which color c can be drawn in U", denoted as $\#c|U$. This is associated with the number of balls of color c in U. For

instance, urn A contains 5 green balls, so $\#g|A$ is associated with 5. The third feature is "number of balls in U", denoted by $\#U$. This feature is associated to the (inverse of) total number of balls in the urn. For instance, urn A contains 10 balls, so $\#A$ is associated to 1/10. This allows for a proper simulation in Task 2 by a rational agent who pays full attention, $R(H_U) = (U, \#g|U, \#U)$:

$$\Pr\big(R(H_U)\big) = (0.5) * \frac{\#\ of\ g\ in\ U}{\#\ of\ balls\ in\ U}. \quad (9)$$

The rational DM imagines selecting U with probability 0.5, picking one of its several green balls, but dividing by the total urn size. This is the logical process behind the Bayes' rule. Conversion of absolute into relative frequencies implies that the rational DM gives the same answer in T_S and T_L.

Selective bottom up attention, driven by contrast, leads this process astray. The DM can pay attention to any statistical feature in isolation or to any combination of them.[24] Critically, though, color contrast sharply changes across treatments. In the small urn B treatment T_S, the color of the ball is not salient, because both urns have 5 green balls. In the large urn B treatment T_L, the color is very salient because B has many more green balls than A: 15 vs 5. This change creates instability.

Prediction 5 *Relative to T_S, the T_L treatment increases attention to the color of the drawn ball, reduces the mode at $Pr(H_A; \alpha) = 0.5$ and generates a mode at $Pr(H_A; \alpha) = 0.25$.*

When urn B is small, there should be a large mode at 0.5 for two reasons. First, color is not salient, so DMs focus on urn selection, anchoring to base rates. Second, even the DMs focusing on color but neglecting the size of urns, $\#U$, issue a 50:50 judgment because urns A and B have the same absolute number of green balls. When B is large, both aspects change. First, color becomes more salient, reducing attention to the base rate. Second, attention to color causes people to simulate urn A with 5 green balls and urn B with 15 green balls, yielding a mode at 0.25.

[24] For simplicity we do not consider the ancillary feature "match" here, which is hard to parse in this format because color shares are not transparently given to subjects. This feature takes value 1 for the unique state of urn B and 0.5 for both urn A states, so attending to it is equivalent to neglecting the color of the ball.

Figure A1 reports the result of this experiment, which is consistent with Prediction 5. Moving from T_S (Panel A) to T_L (Panel B) leads to a large and significant drop in the mode at 0.5 (from 52.3% to 33.3%, $p = 0.00$) and a sizable increase in the mode at 0.25 (from 28.5% to 37.2%, $p = 0.06$).

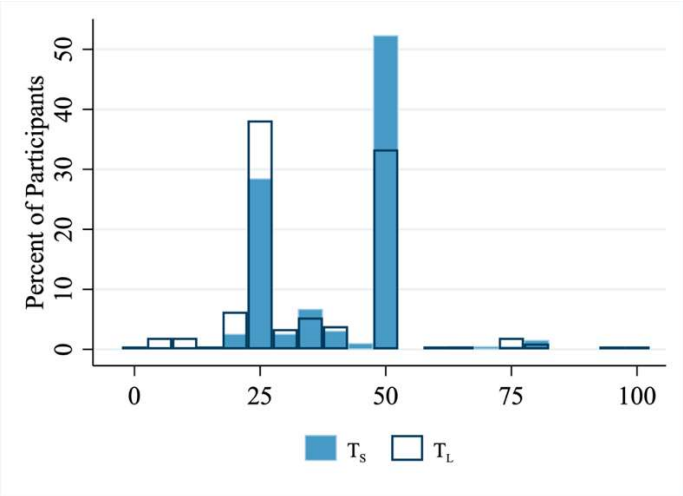

Figure A1. Increasing urn size increases the contrast of, and anchoring to, the ratio of green balls.

The increased large mode at 0.25 is not consistent with any specific heuristic or anchoring to any specific given number. In our model, it reflects deliberate mental simulation under a representation that neglects some relevant features, as explained by Prediction 5 in the case of treatment T_S. Note that the sizeable mode at 0.25 in treatment T_S likely appears because in this problem it happens to coincide with the "new mode" described in Section 5.1 (i.e., the unconditional probability drawing a green marble from urn A, failing to renormalize by the corresponding probability with urn B).

Irrelevant Signals

We also included treatments that added to the balls-and-urns paradigm an irrelevant dimension of possible signals. In particular, in these treatments, each urn now contained five black-and-white marbles in addition to the green-or-blue marbles. In *both* urns, three of these marbles were striped, while two were solid. Thus, any black-and-white marble drawn from the randomly selected urn is uninformative about whether it was from Urn A or Urn B. Figure A2 compares a treatment where the only signal is a single green marble to a treatment where the signal is one green marble and one of these irrelevant striped marbles. Despite these two problems being normatively identical, we see a shift away from the likelihood mode (23% vs 10%, $p < 0.01$) and toward the base rate (29% vs 35%, $p = 0.01$).

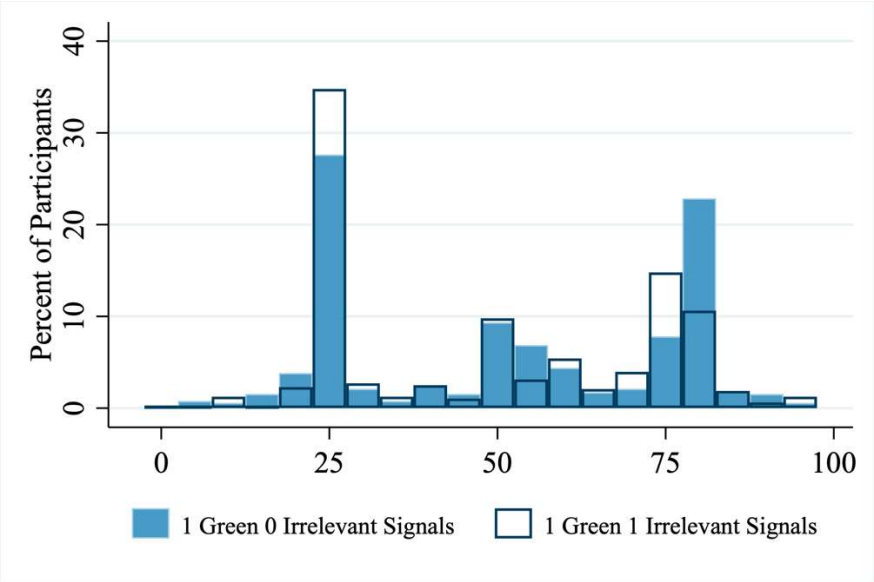

Figure A2. Histograms show the distribution of posterior beliefs that the computer chose Urn A.

Priming share heads through earlier survey questions.

In the main text, we describe treatments T_{full} and T_{last}, which manipulate the prominence of share heads in gambler's fallacy question by changing the question wording (while keeping the underlying problem identical). We also included two treatments meant to test whether more subtle manipulations could achieve similar effects. In particular, in T_{share} and $T_{control}$ we had participants rate 15 pairs of (randomly generated) length-six sequences of coin flips according to how similar they were. This occurred *before* answering the main gambler's fallacy question, which asked about the relative likelihood of $ththht$ and $hhhhth$. In T_{share}, we told participants that by "similar" we meant how much each pair of sequences differed in terms of the fraction of their flips that were heads. In $T_{control}$, we instead defined it as how many individual flips differed between them (i.e., do they disagree on heads vs tails for the first flip, the second flip, etc.). Participants had to answer these questions correctly before they could proceed. Note that these participants did not later rate the similarity between pairs of coin flips (as other participants did after the main inference and gambler's fallacy questions) to avoid confusion.

Figure A3 compares the distribution of answers to the gambler's fallacy question (which was identical across these treatments). We see at most a small effect, with the mean answer in T_{share} decreasing in the expected direction from 42.9 to 41.4 ($p = 0.05$) compared to $T_{control}$. The share committing the gambler's fallacy (i.e., answering with less than 50) does increase from 43.0% to 45.9%, although this difference is not significant ($p = 0.20$). Directly elicited attention to share does increase by 9.2 percentage points (from 56.4% to 65.6%, $p < 0.01$), though the effect on free-response attention to shares is smaller (4.1 p.p., 39.5% to 43.6%, $p = 0.07$). Regressing a dummy of whether

participants in $T_{control}$ commit the gambler's fallacy on dummies for each of these attention measures (separately) yields coefficients of 0.34 and 0.15 (for direct and free-response measures, respectively). Naively multiplying these with the corresponding treatment effects on attention, we might then expect an increase in the incidence of the gambler's fallacy of about 3.1 percentage points (9.2*0.34) or 0.6 p.p. (4.1*0.15) from $T_{control}$ to T_{share}. The actual difference of 2.9 percentage points is not statistically different from either these numbers, so we take these results to be somewhat inconclusive.

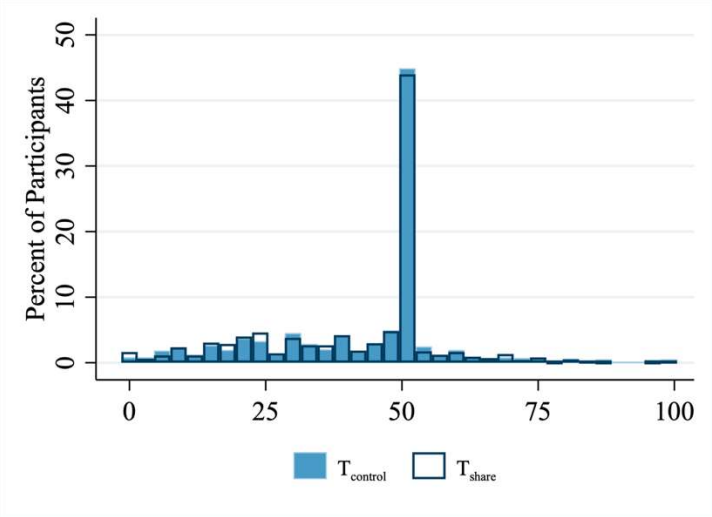

Figure A3. Manipulating prominence of share heads by varying earlier questions in the survey. Lower answers correspond to believing that the more mixed sequence is more likely.

Similarity ratings.

All participants not in $T_{control}$ to T_{sha} , as described in the main text, rated pairs of coin flips according to how similar they found them to be. They also judged the frequency of individual sequences. Each participant was randomly assigned to rate sequences of length 2, 4, or 6. This length was the same for similarity and frequency judgments. For length-2 sequences, frequency judgments were made out of 100 (i.e., how many sequences out of 100 are expected to be X). For length-4 and length-6 sequences, they were out of 500 and 1000, respectively.

Figure A4 shows how frequency and similarity ratings correlate with each other for length-4 sequences, which were omitted from the main text for brevity. We see a similar pattern to length-2 and length-6 sequences: completely unbalanced sequences (the darkest dots) are deemed less similar to other sequences and also less frequent.

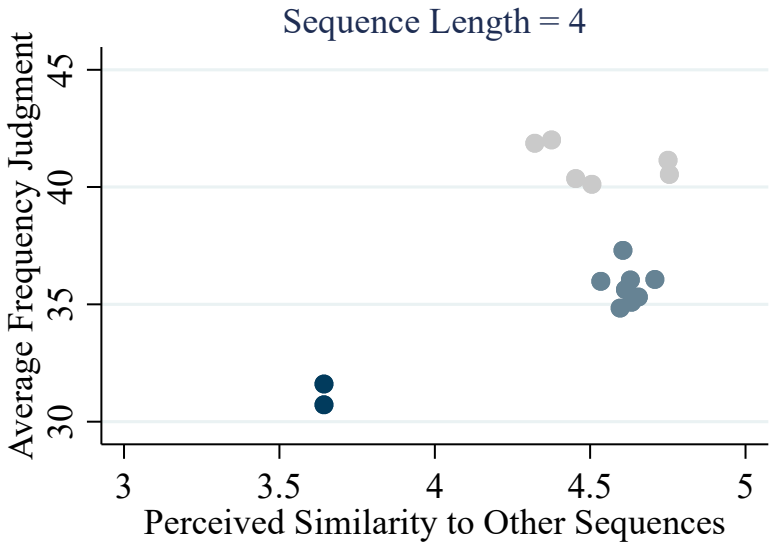

Figure A4. Similarity to average sequence predicts estimated frequency. Each dot corresponds to a coin-flip sequence of length four. The x-axis shows the average judged similarity between that sequence and other sequences of the same length. The y-axis shows the average belief about the likelihood of that sequence (per 500 four-flip sequences).

We mentioned in the main text that, after residualizing on share of heads, similarity and frequency judgments were no longer correlated. Table A2 shows regressions of average frequency judgments and average similarity judgments with and without fixed effects for the number of heads in the sequence for length-4 and length-6 sequences (for length-2 sequences, the fixed effects regression contains the same number of regressors as observations). We see that, absent these fixed effects, similarity and frequency judgments are highly correlated (as Figures 3 and A4 suggest). With these fixed effects, the coefficient on similarity ratings is not significant.

	(1) Length 4	(2) Length 4	(3) Length 6	(4) Length 6
Similarity Rating	5.227***	-1.656	8.392***	1.464
	(1.622)	(1.447)	(0.667)	(1.192)
FEs for # Heads	No	Yes	No	Yes
N	16	16	64	64

Table A2. Table shows OLS regressions. The dependent variable is the average judged frequency of each coin-flip sequence of the length indicated in the column heading. Robust standard errors in parentheses. *** indicates significance at the 1% level.

Appendix C: Model Estimation

In this section, we describe in detail the estimation procedure for the full likelihood model for inference problems in Section 5. First, we exclude participants who are not at any mode, as well as participants at the 50-50 mode.[25] Given the difficulty of exactly computing the Bayesian answer, we assign a participant to the Bayesian mode if her answer is within 5% of the Bayesian answer.

For the remaining participants, we estimate the likelihood function, which is given by:

$$P(i \in e|treatment\ t) = \exp\left(Prom_{e,t} + \beta\left[C\left(\alpha_{e,t}\right)\right]\right)/S_t,$$

Where $S_t = \sum_e \exp\left(Prom_{e,t} + \beta\left[C\left(\alpha_{e,t}\right)\right]\right)$ is the normalizing constant, or the sum of all of the salience terms for $e \in \{match, color, BR, Bayes\}$ holding fixed a treatment t.

Following the theory, we impose that the prominence of the Bayes profile for a given treatment is the arithmetic mean of the prominence of the base rate (urns) and the signal (color):

$$Prom(Bayes|\ t) = \frac{1}{2}\left(Prom(color|t) + Prom(BR|t)\right).$$

Otherwise, we allow the prominence of each feature to be unrestricted, across all treatments. Importantly, we impose a constant β, the loading on contrast, across all treatments.

We construct standard errors by bootstrapping with replacement, and for each bootstrapped sample using gradient descent to maximize the log-likelihood.

Table C1 shows the point estimates and confidence intervals of each parameter.

[25] The reason we do the latter is because reporting 50-50 may not be purely driven by a complete lack of attention to any features.

	Estimate	95% confidence interval	
β	1.20	0.55	1.80
Prominence of Urn: T_B	0.69	0.57	0.95
Prominence of Color: T_B	-0.32	-0.81	0.48
Prominence of Match: T_B	-0.35	-1.42	0.08
Prominence of Company: T_C	-0.41	-0.79	0.63
Prominence of Report: T_C	-0.08	-1.22	2.22
Prominence of Accuracy: T_C	0.56	-2.78	1.26
Prominence of Company: T_U	0.14	-0.25	1.79
Prominence of Report: T_U	-0.30	-1.40	2.65
Prominence of Accuracy: T_U	0.28	-4.35	0.99
Prominence of Urn: T_H	0.49	0.04	1.13
Prominence of Color: T_H	-1.56	-2.86	-0.57
Prominence of Match: T_H	1.08	0.43	1.78

Table C1. Parameter estimates and confidence intervals.