How People Use Statistics

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

September 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 biases in judgments about i.i.d. draws, such as the Gambler's Fallacy and insensitivity to sample size, with biases in inference such as under- and overreaction and insensitivity to the weight of evidence. The model makes predictions about how changes in the salience of specific features should jointly shape the prevalence of these biases and measured attention to features, but also create entirely new biases. We test and confirm these predictions experimentally. 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, Stanford University, Bocconi University and IGIER, Brown University, and Harvard University, respectively. We are grateful to Nicholas Barberis, Ben Enke, Thomas Graeber, Alex Imas, Daniel Kahneman, Giacomo Lanzani, Steven Ma, Dominic Russel, Kunal Sangani, Jesse Shapiro, Claire Shi, Josh Schwartzstein, Cassidy Shubatt, Jeffrey Yang, and Florian Zimmermann for helpful comments. Gennaioli thanks the European Research Council (GA 101097578) for financial support.

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 unifying answer to these questions. A large body of work formalizes specific biases such as the Gambler's Fallacy (GF, Rabin 2002) and sample size neglect in i.i.d. draws (Benjamin, Rabin, Raymond 2016), and 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 offer a new approach in which people pay attention to the salient features of a problem and neglect non-salient ones, even if relevant. Biases arise because, in this process, people represent hypotheses erroneously, exhibiting a form of question substitution (Kahneman and Frederick 2002). The model accounts for and reconciles well-known biases in judgments about i.i.d. draws and in inference. It also explains multimodality and instability of responses in both domains, making new predictions, which we test, on how changes in the salience of specific features shape the prevalence of different biases and measured attention. Last, we predict and find previously undocumented errors.

To see the basic idea, consider the famous duck-rabbit illusion, in which a drawing 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. Nobody sees both animals at once, and nobody says it is 50% chance a duck, and 50% chance a rabbit. Attention selects one representation. When 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 how much he took, leading to different sentences for the same crime under the same law. In bail decisions, some judges may even focus on irrelevant aspects, such as whether a defendant is well groomed (Ludwig and Mullainathan 2023). In these examples, decision makers attend to some but not all features of the problem they face, leading to different representations and judgments.

We argue that the same logic is at play when people solve statistical problems, except here there is an objectively correct answer. These problems also have many features, which people can selectively attend to. When judging two sequences of a fair coin such as $hthtth$ vs. $hhhhhh$ people may focus on the individual flips of each sequence, or on the sequences' share of heads (0.5 vs. 1). When judging the probability that a green ball comes from urn A (vs B), people may focus on the ex-ante probability of selecting urn A, or on the draw of a green ball from it. Depending on which feature is attended to and which ones are neglected, the same hypotheses are represented differently.

We model a decision maker (DM) who, conditional on her representation, correctly uses the statistics given in the problem. Mistakes only arise because her attention to features is shaped by salience. 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. 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 attention, prominence, depends 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. In a statistical problem, a feature is prominent due to the language used to describe the sampling process or the hypotheses, or due to the naturalistic context in which the problem is cast, which cues the relevance of certain features from past experiences. We do not measure prominence directly, but the model tightly disciplines the joint movement of responses and attention to specific features when these are made more prominent in the description.

The model makes two broad predictions, which we test experimentally. First, salience causes neglect of relevant features, leading to bias. And because different features are salient to different DMs, due to either random variation or different experiences, this bias entails multi-modality in the distribution of answers to a problem. When assessing coin flip sequences $hthtth$ vs. $hhhhhh$, some DMs attend to the share of heads, neglecting that each flip is 50:50. They replace the original question with the relative likelihood of obtaining a balanced vs an unbalanced sequence and overestimate $hthtth$. Other DMs attend to individual flips, and do not commit GF. In inference, some DMs attend to the prevalence of hypotheses and anchor to base rates (under-reaction), others attend to the signal and anchor to the likelihood (over-reaction). In either case, one piece of data is used, the other is neglected. DMs attending to both features combine the base rate and the likelihood, sometimes achieving the Bayesian answer. Consistent with this prediction, we document multi-modality, which in inference takes the form of large groups of subjects anchoring exactly at the base rate or at likelihood (see also Dohmen et al. 2009). Critically, in both iid draws and inference measured attention to the features dictated by the model predicts which answers people give, consistent with our mechanism. This occurs even when these features are not associated with a statistic given in the problem, as in the case of iid draws.

Our second and key prediction is that changes in the salience of a feature cause joint shifts in attention and in the distribution of estimates. We test this prediction by manipulating the salience of some features. In i.i.d. draws we make individual flips prominent by describing the same hypothesis in terms of these flips, and show that this reduces both measured attention to the share of heads and the incidence of the GF. In inference, we increase the contrast of the signal by raising the likelihood, and show that doing so jointly boosts attention *and* anchoring to the likelihood, also increasing the share of people who neglect the base rate. In inference, our model also predicts that describing the likelihood in terms of the similarity between the signal and different hypotheses should increase measured attention to such feature and anchoring to the likelihood. We show that this mechanism accounts 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.

The model also explains why the so called "frequency format" (Gigerenzer and Hoffrage 1995) promotes Bayesian answers: it curbs neglect of either the base rate or the likelihood. We however show that the frequency format is not by itself the panacea against distortions caused by bottom up attention. To this end, we manipulate the salience of a hypothesis by not mentioning its alternative in the question. Consistent with the model, this treatment unveils a new bias: many subjects now estimate the prominent hypothesis as the product of its base rate and likelihood, fully neglecting the features of the other hypothesis. This result casts doubt on the notion that human intuition is generally ecologically optimal and sheds new light on the confirmation bias.

Our model explains the coexistence of biases typically attributed to different heuristics such as availability, representativeness, or anchoring (Kahneman and Tversky 1972, Gigerenzer 1996), and why one bias or the correct answer becomes more prevalent when a specific feature becomes salient. Our findings challenge existing models of biases both in i.i.d. draws, which rely on a fixed mis-specified sampling model (e.g., Rabin 2002), and in inference, which rely on a stable distortion of the Bayes' rule. Such distortions may be due to fixed heuristics (Grether 1980), to perceptual noise/complexity (Enke and Graeber 2023, Khaw, Li, and Woodford 2021), or to a combination of the two (Ba, Bohren, Imas 2023). These models are not consistent with the key patterns in the data: i) multimodality of answers to a problem, ii) instability, including across normatively equivalent problems, and iii) a systematic correlation between different biases and attention to different features.

We formalize the key concept of salience-driven, associative simplification of features using insights from psychology and machine learning (Tversky 1977, Kruschke 2008, Selfridge 1955, Guyon and Elisseeff 2003) to model our key findings of multimodality and instability of beliefs. Compared to models of goal-optimal attention (Sims 2003, Woodford 2003, 2020, Gabaix 2019), we explain why highly goal-relevant information can be neglected and why goal irrelevant changes dramatically shape attention and biases. Our paper relates to a growing body of work showing that biases can persist even in the presence of feedback and incentives due to selective attention, which can arise from incorrect models (Schwartzstein 2014, Gagnon-Bartsch, Rabin, and Schwartzstein 2023, Esponda et al, 2022) or computational complexity (Simon 1957, Enke and Zimmermann 2019, Enke 2020, Graeber 2023). In our model, selective attention is driven by bottom-up salience of features, and bias arises even in computationally simple problems. In coin flips, it is trivial to avoid GF by recognizing that each flip is 50:50. Bias is caused by the salience of a feature, the share of heads, that is relevant for a different problem. Moreover, shifts in bottom-up salience lead to instability of choices, whereas much of earlier work focuses on stickiness in biases.

Statistical problems provide a great setting to study our mechanism because the given statistics offer anchors for detecting shifting attention to and the use of information, yielding sharp multi-modality. Our mechanism is also relevant for belief formation and choice more broadly, including in domains with no clear anchors. In such domains we should not expect to see sharp multimodality, but selective attention to the features of events will still produce sharp disagreement and instability in the nature and magnitude of the average bias in the population. That bias may be under-reaction in some domains (e.g., climate change) and overreaction in others (e.g., financial bubbles), but may change rapidly when events change which features are salient for a significant group of people.

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 "iid draws" 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.

For iid draws, we told participants that we created a large number of sequences from tosses of a fair coin. In the first treatment, 100 of these sequences were either $H_1 = th$ or $H_2 = hh$. In the second treatment, they were either $H_1 = ththht$ or $H_2 = hhhhhh$. We asked participants for their best guess of how many of these sequences were from H_1 or H_2. Panels A and B of Figure 1 show the distribution of beliefs about the relative share of the balanced sequence for each treatment.

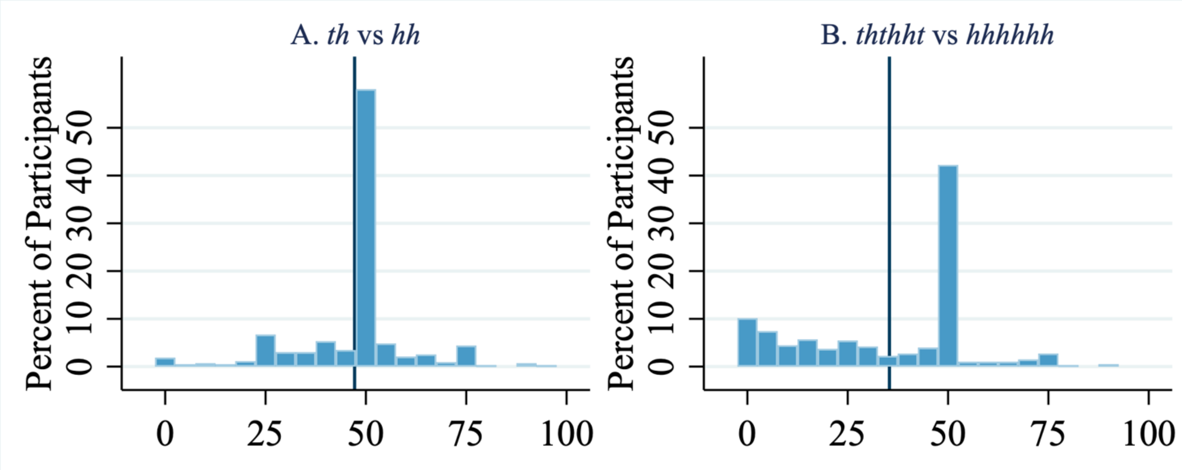

Figure 1. Each panel reports the distribution of estimated $\Pr(H_2|H_1 \cup H_2)$. Answers closer to 0 indicate higher probability of the balanced sequence H_1. The blue bar marks the mean answer.

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, GF 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, $p = 0.00$). Instability in the share of people committing GF is inconsistent with a mechanical, possibly heterogeneous, tendency to use a mis-specified sampling model (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 inference next. We presented a problem in two different yet normatively equivalent formats. In the "balls and urns" treatment (Edwards 1968), participants were told that an urn A contains 80% green and 20% blue balls, while 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. They are then asked the probability that it was drawn from A vs. B. In the more naturalistic "cabs" treatment (Kahneman and Tversky 1972), participants were told 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, while 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%. They are the asked the probability that the errant cab was indeed Green vs. Blue. We run the two formats with identical statistical parameters with two sets of participants, which to our knowledge has not been done before. Using Bayes' rule, the correct answer is $\Pr(A|g) = \Pr(Green|g) = 0.57$ in both problems. The distribution of answers is reported in Figure 2.

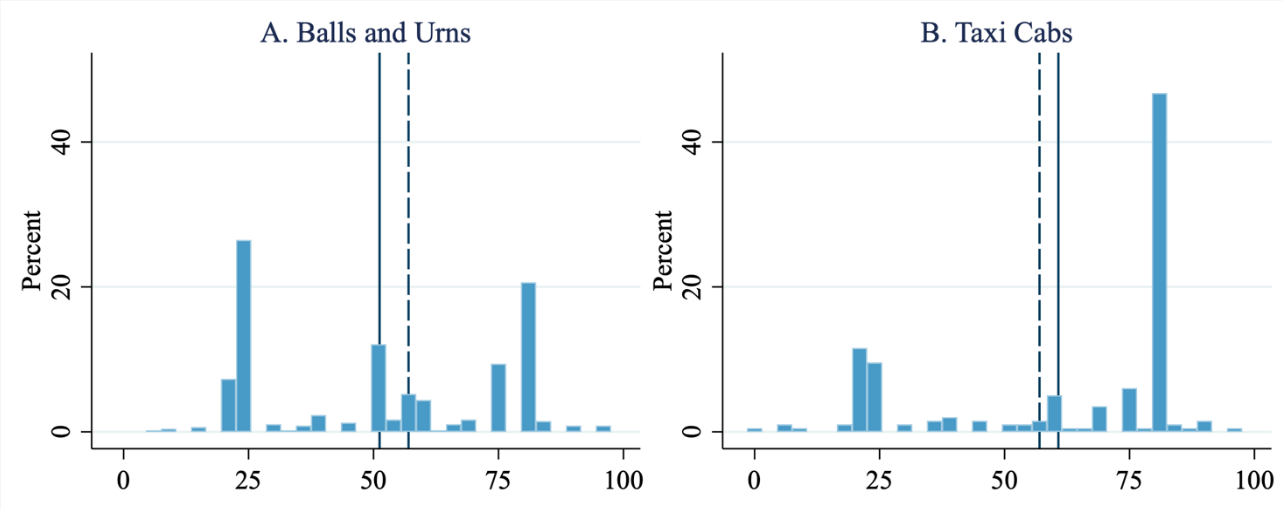

Figure 2. The left panel reports the distribution of $\Pr(A|g)$, the right panel of $\Pr(Green|g)$. The solid line indicates the mean answer, while the dashed line indicates the Bayesian answer of 0.57.

Consistent with previous work (Benjamin 2019), 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 also 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), with a shrinkage of beliefs to the prior due to noise (Woodford 2020, Enke and Graeber 2023), or with any fixed heuristics. Even answers typically attributed to epistemic uncertainty (De Bruin et al 2000) 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 combine 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.

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 features associated with the base rates (underreaction) or the 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 hypotheses. 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.

Formally, a statistical problem has three components: i) the sampling process, ii) the statistics, e.g. the probabilities of specific events, 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, but the model also covers the case in which they correspond to absolute frequencies (see Appendix B). 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 the probabilities in b) do not add up 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. The DM reaches the correct answer if she attends to the relevant features, but commits errors if not. Section 3.3 formalizes two key drivers of DM's bottom-up attention to features: contrast and prominence. Section 3.4 describes how to apply the model and test its predictions 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 features $j \leq n$ "statistical", because each of them is associated with a statistic $\Pr(f_j)$: the *true probability* of each such event. 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 balls 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.

Features $f_{n+1}, \ldots, f_F$ 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 notable feature because it determines the similarity of a sequence to its data generating process: (h, t) is similar to the fair coin that produced it because its 0.5 share of heads is what a fair coin tends to produce.[2] In inference, there is also an ancillary feature that captures the similarity of realized data to the data generating process: whether the realized signal is the most likely outcome of the hypothesis or not. In the example 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, $m = 0$.

By capturing similarity to the data generating process, the share of heads in coin flips and match in inference are connected to KT's "representativeness" heuristic: an event is representative of a statistical process if it resembles salient features of that process. In our model, though, there are no stable heuristics. There are instead many features. Some, the statistical ones, are tied to sampling steps. Others, like the similarity of a sequence/signal to the statistical process, capture different properties. These features "compete" for the 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

The DM solves the problem by executing three tasks: 1) construct a simplified feature-based representation of the hypotheses based on selective attention, 2) compute the probability of these representations using the statistics, and 3) normalize 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 $\alpha = (\alpha_1, .., \alpha_{n+1})$. 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. For simplicity, she attends either to statistical or ancillary features, not to the mixtures of the two (this restriction can be relaxed). Denote the set of feasible attention profile by A_K. Selective attention the 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 $\tilde{f}_\alpha(\omega) = (\tilde{f}_{\alpha,1}, \dots, \tilde{f}_{\alpha,n+1})$, 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 $R_\alpha(H_i) = \cup_{\omega \in H_i} \tilde{f}_\alpha(\omega)$.

[2] Longer sequences have more ancillary features, e.g. (h, t, h, t, h, t) is "alternating", and (t, t, t, h, h, h) is "sorted".

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:*

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

The DM computes the joint probability of the features-events she attends to. If she attends to more than one statistical feature, for each vector $\tilde{f}(\omega) \in R(H_i)$ she computes $\Pr\big(\tilde{f}_r \cap \ldots \cap \tilde{f}_s\big)$ 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 simulating $R_\alpha(H_1) = (\varphi, \varphi, 1)$, computing the probability of obtaining only heads $\Pr(sh = 1) = (0.5)^2$, and by simulating $R_\alpha(H_2) = (\varphi, \varphi, 0.5)$, computing 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 final estimate is reached by normalizing simulated probabilities.

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

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

Normalization only matters if 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 GF. 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 GF. 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). Given a 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 using attention* $\alpha^*(\omega)$ *in Task 1 and then following Tasks 2 and 3, 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 all features that are relevant to the problem while neglecting others. With the correct simplification strategy in Equation (1), Tasks 1, 2 and 3 guarantees a correct solution. As we show in the proof, the minimum number of relevant features of hypotheses can be found using a coarsest partition of them in terms of events whose probability can be computed. In our example, there is a unique partition of 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 correctly solves this problem when she attends to the first and second flip while she commits the Gambler's Fallacy when she attends to the share of heads.[3] But what shapes attention? We address this question next.

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 stimulus to the goals of the DM. Rational inattention models formalize this idea (Sims 2003; Gabaix 2019, Woodford 2003, 2020; Khaw et al. 2021). Bottom-up attention reflects instead an involuntary focus on salient stimuli which causes neglect of non-salient ones, even if relevant (BGS 2012, 2013, 2022, Li and Camerer 2022, Evers, Imas, and Kang 2023). Sometimes the attention-drawing stimulus is relevant to the task but still distorts the decision. While driving, a surprising police radar may cause us to neglect the car behind us, so we break too heavily. But a stimulus may draw attention even if it is entirely irrelevant, such as when a black stain on the wall distracts us from a conversation.

Section 2 highlighted the role 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). More broadly, in standard models of bias people know Bayes' rule but distort true probabilities, because they use a misspecified sampling process (Rabin 2002, Rabin and Vayanos 2010) or because they overweight the prior or the signal (Grether 1980), e.g. due to rational inattention or perceptual noise (Khaw, Li, and Woodford 2021, Enke and Graeber 2023). Hypotheses are properly represented, statistics are

[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.

combined, at least to some extent, and biases are stable: the share of biased people and the type of bias they commit should not change as they do in Figures 1 and 2.

In contrast, in our model salience driven shifts in attention can account for instability by changing the representation of hypotheses and the use of statistics. The 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 black stain is on a white 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. the stain is in front of us). Thus, 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, 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:

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

The numerator 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 $|\Pr(sh = 1) - \Pr(sh = 0.5)|/(\Pr(sh = 1) + \Pr(sh = 0.5)) = 1/3$. The contrast induced by attention to individual flips, $\alpha = (1,1,0)$, is instead zero, $|\Pr(h,h) - \Pr(h,t)|/(\Pr(h,h) + \Pr(h,t)) = 0$. Here contrast encourages attention to sh. More generally, contrast is shaped by the objective parameters of the problem. In coin flips, it is shaped by the probability of a head and the sequence length n. In inference, it is shaped 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 prominent if the description of the problem brings it to the top of mind. There are two possible mechanisms for this. First, some formal ingredients of the problem, such as the sampling process producing Ω and the hypotheses H_1 vs. H_2, can be described in a way that makes a specific feature salient. In balls and urns, the composition of the urns could be described as “The color of a drawn ball matches 80% of the time the color of the urn (Green vs. Blue) it comes from”. This description of the sampling process is logically equivalent to that in Section 2, but it makes the "match" feature more prominent. Similarly, describing the hypothesis as "Urn-A" vs "Urn-B" as in Section 2 makes the urn selection feature more prominent than describing them as: does the green ball "match" vs. "not" the color of the urn it comes from? Again, the two ways of describing hypotheses are logically identical, but the latter raises prominence of the match feature.

A second source of prominence is the context in which the statistical problem is cast, which causes—due to past experiences—certain features to be more salient than others and hence top of mind. In consumer choice, the role of past experiences is well established. For instance, demand for insurance increases after floods because the recent experience brings this risk top of mind (Slovic, Kunreuther, and White 1974). In a statistical problem, describing the same inference problem in a courtroom context, as in taxicabs, can cause the witness statement to be salient due to many direct or fictional experiences a participant remembers with high relevance of witness reports in court.

In our experiments we manipulate prominence by changing the description of the problem in ways that intuitively make certain features prominent, as we just discussed. We do not measure prominence externally, which may be possible to do using text analysis. To validate our prominence manipulations, we measure attention to features and correlate it with biases. Our model makes strong predictions for that correlation. To derive these predictions, we introduce prominence as a latent variable that affects attention α. The prominence of feature j is a scalar P_j, and the prominence of profile α, denoted $P(\alpha)$, is formalized as the average prominence of its features:

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

Equation (5) captures, in the simplest way, two important aspects of attention. First, making a feature more prominent, increasing P_j, increases the salience of all representations using this feature, also in conjunction with others, i.e. of all profiles having $\alpha_j = 1$. Second, there is interference: if a DM attends to feature j', increasing the prominence of feature j is less impactful, because the DM's attention is divided. Interference creates sparsity. 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 in an individual specific extreme value term ϵ_α. This term captures stable individual differences in prominence due to different past experiences, as well as 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:*

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

The term $\epsilon_{\tilde{\alpha}}$ captures an individual level component of salience, yielding a multinomial distribution of attention and, using Tasks 1-3, a distribution of judgments. Within a treatment, attention and biases should be correlated at the individual level, due to variation of $\epsilon_{\tilde{\alpha}}$ across people. Second, and critically, attention and biases should be correlated across treatments: an increase in the salience of a feature should increase the share of people attending to it and making the associated judgment. In our experiments we test both predictions. 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

To apply our model, the analyst must specify and measure two objects: features and attention. Some features are 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 need not be explicitly mentioned. They capture broader properties of events, in our case the similarity of an event to its data generating process, motivated by representativeness (Kahneman and Tversky 1972). In more complex problems, many ancillary features may shape beliefs, just like many non-hedonic yet salient features, such as advertising and broader context, shape consumer choice. These features can be empirically discovered by asking people for a rationale for their choices, by using text analysis or algorithms.[4] Specifying/discovering features is the key first step.

Once some explanatory features are identified, the model can be tested by studying how beliefs, 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, 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

[4] 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 specified by the analyst ex ante.

1982, Nosofsky 1988): people judge two objects to be more similar when they attend to features the two objects share.[5] We then assess whether different measures yield comparable results.

In sum, 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 $\Pr(H_i; \alpha)$. When this is done, the predictions of Equation (6) can be tested by examining the individual level correlation between attention and behaviour (multimodality), and the joint aggregate shifts in these measures (instability). In Sections 4 and 5, we showcase this method in the domains of 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 the GF.

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

Attention and Representation. A DM attending to all statistical features, individual flips, while ignoring the share of heads, $\alpha_n = (1,1,..,0)$, 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 $\Pr\left(R_{\alpha_n}(H_i)\right) = (0.5)^n$, which is identical across hypotheses, yielding 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 both hypotheses as $\Pr\left(R_{\alpha_r}(H_i)\right) = (0.5)^r$. By contrast, a DM attending only to the share of heads, $\alpha_{S,n} = (0, \dots 0,1)$, represents hypotheses as $R_{\alpha_{S,n}}(H_i) = (\varphi, \dots, \varphi, sh)$. By (2) she simulates them by the probability of its share of heads, $\Pr(sh)$, which causes her to underestimate H_1 and commit the GF.

Endogenous Attention and Estimates. To determine the distribution of attention and estimates in an experiment, we must describe the attention profile of different DMs. Denote by P the scalar

[5] 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.

prominence of each individual flip relative to sh. Denote by $C(\alpha_{S,n})$ the contrast of $\alpha_{S,n}$, which depends on length n. Proposition 2 characterizes multimodality, Corollary 3 instability.

Proposition 2 *A share $\mu(\alpha_{S,n})$ of DMs attends to the share of heads and for $n > 1$ commits the Gambler's Fallacy, estimating the relative probability of the unbalanced sequence as:*

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

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%.[6] The key new prediction is their connection to measured attention: a DM committing the GF should also be more likely to attend to the share of heads. The model also predicts that bias *and* attention should change when the salience of the same feature changes.

Corollary 3 *The share $\mu(\alpha_{S,n})$ of DMs who attend to the share of heads and commit the GF increases in sequence length* n *and decreases in the prominence of individual flips P.*

As n increases, more people commit the GF because the contrast-based salience of sh, $C(\alpha_{S,n}) = \left[\binom{n}{n/2} - 1\right] / \left[\binom{n}{n/2} + 1\right]$, rises with 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 salient representation, and for many DMs replaces the original question. Thus, our model explains the fall in the 50:50 mode when moving from Panel A to Panel B in Figure 1: it is caused by the higher contrast of the share of heads when $n = 6$ compared to $n = 2$.[7] Corollary 3 also predicts a prominence effect: increasing the salience of individual flips in the problem's description causes them to be top of mind, draws attention away from sh, in turn reducing the incidence of the GF.

These predictions distinguish our model from existing accounts of biases in i.i.d. draws. In these models, bias is due to the use of incorrect sampling models, such as draws without replacement (Rabin 2002, Rabin and Vayanos 2010). These models do not predict a link between bias and attention to an irrelevant feature of hypotheses: hypotheses are correctly represented and estimated according to a stable but incorrect model. A fortiori, these models do not predict the instability in the share of people who attend to an irrelevant feature and commit the GF. We next test these predictions.

---

[6] 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 GF than others.

[7] In our model the severity of the GF increases with n also because, conditional on attending to the share of heads, the faulty equivalence class of balanced sequences gets larger, so bias in (7) increases.

Coin Flip Experiments. Table 1 provides a summary of the treatments. In all treatments, individuals are asked to judge the relative likelihood of a given unbalanced and balanced sequence and also report what features of the data they attended to. In treatments T_2 and T_6, which we showed in Section 2, the two sequences are given by $H_1 = hh$ vs. $H_2 = th$ and $H_1 = hhhhhh$ vs. $H_2 = ththht$ respectively. We also introduce two new treatments to study the role of prominence. In T_{full}, subjects are asked to estimate $H_1 = hhhhhh$ vs. $H_2 = hhhhht$, where the hypotheses are described by full sequences, as in T_2 and T_6. In T_{last}, we instead tell subjects, "the first five flips were $hhhhh$. What is the probability that the final flip was heads or tails?" T_{last} is logically equivalent to T_{full}, but the description of hypotheses makes the last flip more prominent.

After eliciting participants' estimates, we independently measure free-response and direct-elicitation proxies for attention to features. The 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. Similarity judgments should then reflect attention: if the DM attends to the share of heads rather than to individual flips, the same two sequences should be less similar because, while they have several flips in common they sharply differ along sh.[8] We thus interpret low similarity as a proxy for attention to the share of heads.

Across the four treatments, we test two sets of predictions. First, as predicted by Proposition 2, there should be an individual level association between beliefs and attention within each treatment: a participant's attention to the ancillary feature sh should be positively correlated with her tendency to commit GF. Second, across treatments, there should be instability in biases driven by contrast and prominence, as predicted by Corollary 3. The share of participants committing GF and those attending to the share of heads should be greater for longer sequences (T_2 vs T_6) and smaller when individual flips become more prominent (T_{full} vs T_{last}).

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}

[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.

T_{last}	978	Ask about final flip. in 6-flip sequences i.e.. P(h vs t | $hhhhh$)	Increase prominence of final flip compared to T_{full} (and thereby reduce attention to share heads

Table 1. Treatments manipulating salience in the gambler's fallacy problem.

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 she commits the GF (i.e., reports 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 (Table 2, 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 2. 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.

Consistent with our model, a subject attending to the share of heads is more likely to commit GF (Column 1), and a subject perceiving the same two sequences as more similar, which indicates less attention to sh, is less likely to commit GF (Column 2). Each measure of attention has predictive power conditional on the others (Column 3). These findings support the notion that bias arises due to an erroneous representation of hypotheses caused by a salient yet irrelevant feature.

Instability in Beliefs and Attention. Consider instability next. In Figure 1, increasing sequence length from $n = 2$ to $n = 6$ increases the incidence of GF. Figure 3 compares beliefs for T_{last} and T_{full}: we find that the mean estimate of H_1 is significantly higher (49.3 vs 44.4 out of 100, $p < 0.01$) for T_{last} than T_{full}, driven also by an increase in the mode at 50: 50 (68% vs 54% of participants, $p < 0.01$). Consistent with Corollary 3, changing the description of hypotheses in a way that renders individual flips salient reduces the share of people committing the GF. This is consistent with the idea that instability in bias is generated by instability in the "bottom up" representation of hypotheses.

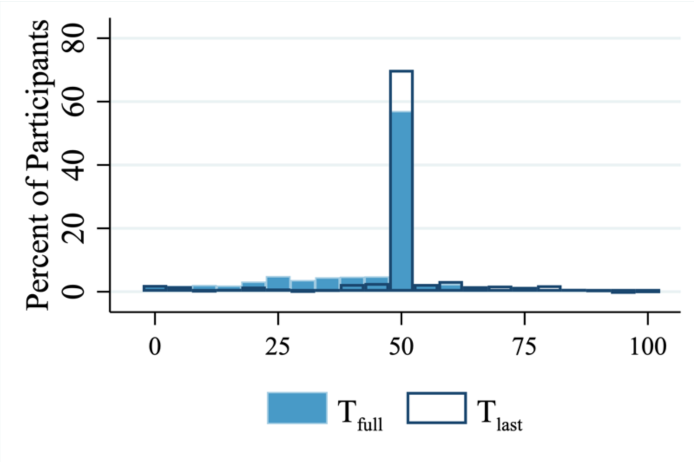

Figure 3. Making the last flip more prominent reduces the Gambler's Fallacy. This figure reports the distribution of estimated Pr(*hhhhhh* | *hhhhht or hhhhhh*). Answers closer to 0 indicate higher probability of the balanced sequence.

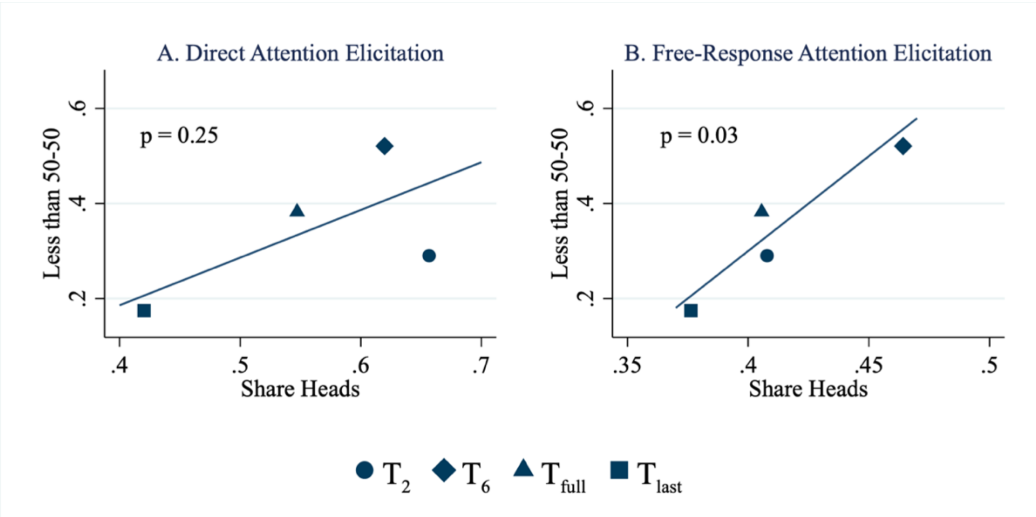

Figure 4. Treatment effects in Gambler's Fallacy and attention. The x-axis is the fraction of participants in each treatment that attend to share heads according to our direct-elicitation (Panel A) and free-response (Panel B) attention measures. The y-axis is the fraction of participants across treatments who judge the balanced sequence to be more likely than the unbalanced sequence.

We next test whether treatment effects in beliefs correspond to changes in attention, which proxies for the changing salience of different features. Figure 4 plots the fraction of subjects in each treatment who commit the GF along with that of attending to sh according to the direct-elicitation (Panel A) and the free-response (Panel B) proxies. 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, based on subjects' reasoning, detects model-consistent instability in attention across all treatments. As predicted by our model, instability the GF is closely associated with shifting bottom-up attention to an irrelevant feature, the share of heads.

We conclude by exploring the connection between attention to the share of heads and similarity judgments. At the end of the survey, all participants answered two additional modules. In $Probability_n$, participants rated the unconditional probability of multiple randomly generated n-flip sequences. In $Similarity_n$, 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) sequences and two (eight) non-overlapping pairs. For $n = 6$, they rated 16 randomly selected sequences and non-overlapping pairs (we correct for the fact that some sequences were more likely to be selected). The similarity measure in Table 1 came from answers in $Similarity_n$.

Figure 5 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 target sequences. In both panels, more balanced targets are perceived to be more similar to the average sequence than unbalanced ones. That is, a target with 0.5 share of heads is perceived as similar to the many other balanced sequences, despite the differences in individual flips. This pattern closely tracks the GF: there is a clear positive correlation between judged frequency of a sequence and its average similarity to other sequences ($p < 0.05$ for both panels). Our mechanism predicts this relationship: the hypothesis of a balanced sequence is misrepresented, it gets confused with many other balanced sequences to which is similar, boosting its estimated frequency. Furthermore, the share of heads appears to be the feature that drives this pattern: controlling for the share of heads removes any significant correlation between similarity and frequency (see Appendix B).

[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 ($p=0.10$).

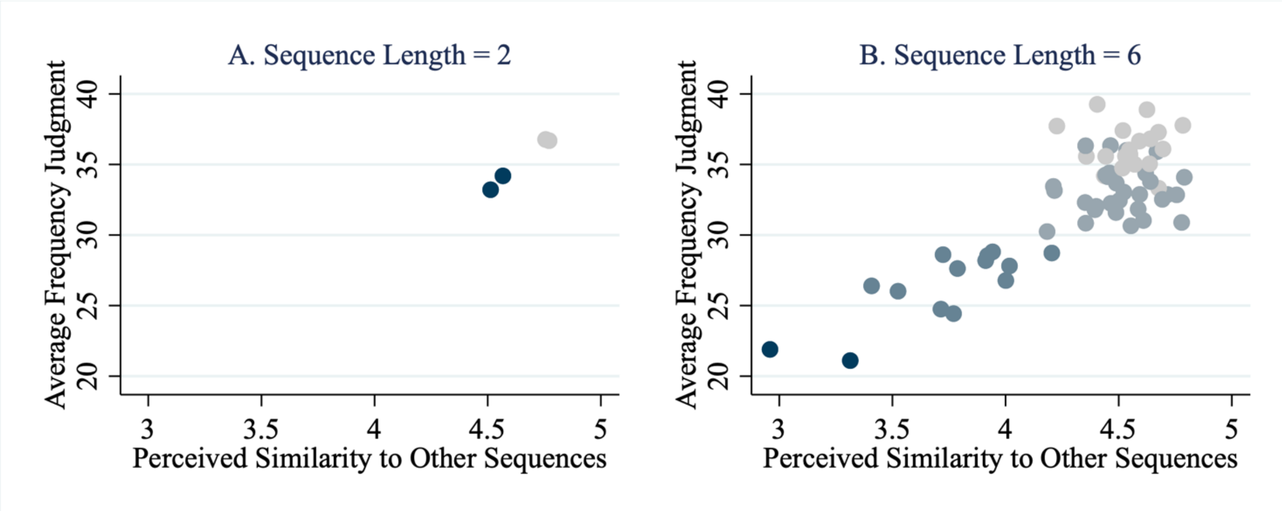

Figure 5. Average judged similarity to other sequences predicts frequency judgments. Lighter dots indicate more balanced sequences, indicating that share heads drives both measures. Frequency judgments are expected number of sequences out of 100 (Panel A) or 1000 (Panel B).

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 that, when thinking about Tom, people attend to his described features – "a meek and tidy soul" – and simulate a librarian, neglecting many non-salient features that may cause Tom to land in a different job. Similarity and probability judgments are driven by partial attention to features.

5. Salience, Multimodality and Instability in Inference

We show that selective bottom-up attention to certain relevant or irrelevant features accounts for the coexistence of under and over-reaction in inference and for their instability documented in Figure 2, creating 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, 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 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 as first selecting the urn and next drawing a green ball from it, $R_{\alpha_\beta}(H_U) = (U, g|U, \varphi)$. This DM simulates the hypothesis as $\pi_{g|U}\pi_U$ and obtains, after normalization, the Bayesian answer, $\Pr(H_A; \alpha_\beta) = \beta$. 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* is drawn 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 that a ball drawn from A is green compared to one drawn from B?", formally $R_{\alpha_c}(H_U) = (\varphi, c|U, \varphi)$. This DM simulates H_U using its likelihood $\pi_{g|U}$, yielding 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 a ball matches the urn's color?", $R_{\alpha_m}(H_U) = (\varphi, \varphi, m)$. This DM simulates H_A as $\Pr(m = 1) = \pi_{g|A}\pi_A + \pi_{b|B}\pi_B$, which also yields $\Pr(H_A; \alpha_m) = q$.

In the last two 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 think "a green ball could come from either urn" and report $50:50$.[10] This bias does not reflect a sophisticated reaction to epistemic uncertainty, but rather the fact that no feature is salient to the DM. When a feature becomes salient, anchoring to 50:50 should drop, as we find in Figure 2.

Endogenous Attention and Estimates. Proposition 4 collects the results above by allowing for individual level variation in attention in Equation (6).

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* $\Pr(H_A; \alpha_0) = 0.5$.

Due to individual-level differences in attention, the model predicts, within an experimental treatment, a systematic relationship between measured attention to features and the probability estimate which accounts for the multi-modality observed in Figure 2. As in coin flips, the model then also predicts instability. Denote by P_l the scalar prominence of feature $l = U, c|U, m$.

[10] Here, no attention to features can also capture the possibility that the DM's attention jumps between "urn selection" and "color of ball", which favor different hypotheses, without settling on 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*.*

Due to contrast, making the signal more informative boosts attention it gets and the share of people anchoring to the likelihood (the opposite occurs if the base rate becomes more extreme). Due to prominence, purely contextual changes do the same, jointly increasing attention to a feature and anchoring to its associated statistic (the likelihood) at the expense of other features.

Corollary 5 offers an explanation for the instability in Figure 2: features of the likelihood are more prominent in taxicabs than in balls and urns, relative to the base rate. Consider the description of the sampling process. In balls and urns, the likelihoods are described separately as the composition of urns A and B, making the urns prominent. In taxicabs, the likelihood is described in terms of the probability the signal matches the hypothesis: "a test reveals that the witness can correctly identify each cab color with probability 80%". This raises the prominence of the "match" feature. Hypotheses are also described differently: in balls-and-urns the hypotheses are framed as "A" vs "B", making urn selection prominent, in taxicabs they are framed as whether "the errant cab is indeed Green vs Blue (as the witness claimed)", raising the prominence of the match. Lastly, the courtroom context of taxicabs may also increase the prominence of the signal, due to personal or fictional past experiences of relevant witness reports in court. All of these irrelevant changes may shape bottom-up attention and explain the instability of biases. Critically, our model specifies how these description changes should be reflected in changes in attention to specific features, which we test in our experiments.

The connection between bias and attention in Proposition 4 and Corollary 5, leading to the instability of biases in statistically identical problems as in moving from balls and urns to taxicabs, does not arise in standard models of biased inference. In these models, people apply the Bayes' rule in a distorted way but: i) they use *both* the base rate and the likelihood, and ii) distortions are due to stable weights (Grether 1980, Enke and Graeber 2023).[11] Due to i), people should pay attention to both statistics, at least to some extent, but not to the irrelevant "match" feature capturing the similarity between the signal and its generator. Due to ii), attention and bias should not change when the problem is reframed. Consider instability in detail. With respect to contrast, Corollary 5 predicts that making one relevant piece of information more extreme (the likelihood) interferes with attention to,

[11] In Enke and Graeber (2023) people perceive likelihoods imprecisely, which causes: i) a dispersion of estimates, and ii) a shrinkage of posteriors toward the prior which gives an average under-reaction bias. In our data, we see some estimates that are not anchored to the base rate or likelihood or to 50:50, but we do not see the concentration around the middle that is the hallmark of under-reaction in that model.

and hence the use of, another relevant piece of information (the base rate). In standard models, making one statistic more extreme does not inhibit the use of the other. With respect to prominence, Corollary 5 predicts that normatively irrelevant changes in description should shape attention to specific features and judgments, which does not happen in standard models, which postulate that attention is focused on the (unchanged) relevant features. We now test Proposition 4 and Corollary 5.

Inference Experiments. Table 3 provides a summary of the treatments. T_B and T_C are our baseline balls-and-urns and cabs treatment, which we described in Section 2. To test the role of contrast and prominence in beliefs and attention, we add 4 new treatments. T_{LE} and T_{ME} test the role of contrast: in the "less extreme" likelihood treatment, T_{LE}, the base rate is 0.15 and the likelihood is 0.70, while in the "more extreme" treatment, T_{ME}, the base rate is again 0.15 but the likelihood is increased to 0.90. The wording of T_{LE} and T_{ME} are otherwise identical to that of T_B.

T_H and T_U test the role of prominence, which we hypothesized to play a role in the instability across T_B and T_C: while the underlying statistical problem remains the same as that of T_B and T_C, the treatments differ in how the hypotheses and sampling processes are described. In treatment T_H, we modify T_U to increase the prominence of match. We label the urns by their modal color, "Green-urn" vs. "Blue-urn," and describe the likelihood (80%) as the probability a drawn ball "matches" the color of the urn.[12] The rewording thus increases the prominence of the "match" and the "color of ball" features, which we also verify experimentally. In treatment T_U, we conversely change T_C to make features of the underlying signal less prominent. We do so by modifying: i) the description of the witness to "the court found that the witness was very unreliable: he was able to identify each color correctly only about 80% of the time…", and ii) the description of the base rate to "the large majority of cabs in the city—75% to be exact—are Blue, while the remaining 25% are Green." These changes decrease the perceived relevance of the report and increase that of the base rate by relying on past experience (i.e., very unreliable witnesses are irrelevant in court), which affect attention and biases even though the statistical informativeness of the signal is unchanged.

To measure attention, in each treatment we ask participants to justify their probability estimates in free form and then ask them to choose which features they attended to from a list that in balls and urns 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

[12] 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."

witness report.[13] The free response measure of attention is again based on asking chat GPT to choose which answer in 1)-4) the subject likely chose.

Across the six treatments, we again test two sets of predictions. First, as predicted by Proposition 4, there should be correlated multimodality in beliefs and attention within each treatment: reported attention to urns, color, and match should align with which mode the DM anchors to. Second, comparing across treatments, there should be correlated instability in biases and attention driven by contrast and prominence, as predicted by Corollary 5. A rise the contrast of the likelihood (T_{LE} vs T_{ME}) or the prominence of match (T_H vs T_B) should boost both attention to the signal and anchoring to the likelihood. Conversely, lowering the prominence of the signal (T_U vs T_C) should shift attention away from the signal and increase anchoring to base rates. Finally, we test whether when moving from balls and urns (T_B) to taxicabs (T_C), there is greater attention to match and color.

Treatment	Base Rate	Likelihood	N	Summary	Purpose
T_B	0.25	0.80	480	Balls and urns: baseline	Compare to T_H
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 3. Treatments manipulating salience in inference problems.

[13] 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. In practice, this choice does not affect our main results, as by far the most common such attention profile (28% of participants) involves paying attention to both the signal and the match feature (recall that attending to either feature in our model would yield the same answer to the inference problem).

Multimodality in Attention and Estimates. We test Proposition 4 by connecting within each treatment multimodality in attention and judgments. The large majority of answers are anchored to one of the modes in Proposition 4 (ranging from 68.2% to 78.2% of answers depending on treatment). Pooling all inference treatments in Table 4, 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.

	(1)	(2)	(3)	(4)
	Base Rate	Likelihood	Bayes	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 4. 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 4 shows that measured attention profiles strongly predict estimates in a way consistent with Proposition 4. For example, participants who report attending to only the urn feature are 41.8

percentage points more likely to anchor to the base rate. Free-response attention to urn further increases that probability by 16.9 percentage points. Similar results hold for other modes. Furthermore, many people report paying attention to only one feature, which is either a statistic or the irrelevant match feature, which is then reflected in which statistics they use or neglect. Participants who pay attention to both features are more likely to make a correct judgment.

One potential concern is that the link between reported attention and estimates comes from participants mechanically reporting features associated with their estimates. This, however, does not explain why attention to ancillary features that are not associated with statistics, such as the share of heads or match in balls and urns, also predicts beliefs. Furthermore, we also elicit attention using free responses, which provide a more semantic and less mechanical description of how respondents thought about the problem. Table 4 shows that free responses have additional explanatory power beyond directly-elicited attention, suggesting that the correlation between attention and choice genuinely reflects the heterogeneity of how participants represent and solve the problem.

Attention and Instability in Estimates We next show the effect of controlled manipulations of contrast and prominence. We first look at estimates, and then document shifts in attention as predicted by Corollary 5. Consider contrast first. The left graphs of Figure 6 compare the T_{LE} vs. T_{ME} likelihood treatments. In Panel A, consistent with the model, increasing the likelihood from 0.7 in T_{LE} to 0.9 in T_{ME}, increases the share anchored to the likelihood (from 15.5% to 22.8%, $p = 0.00$), and decreases the share anchored to the base rate (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] Consequently, in Panel B the relative share of answers at the likelihood or Bayes vs. the base rate increases, 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.

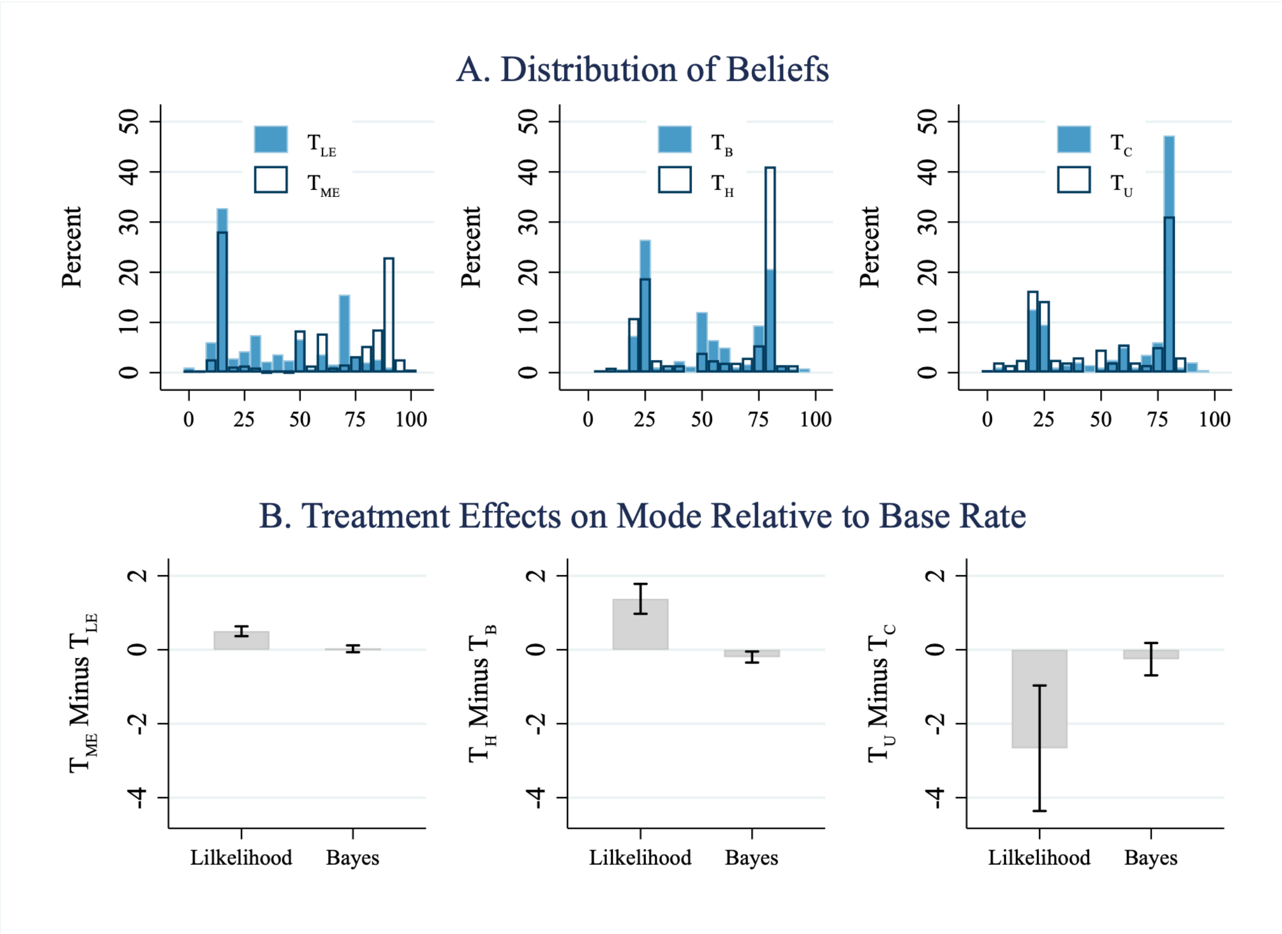

Figure 6. A shows the distribution of beliefs about $\Pr(A \mid g)$ across inference treatments. Panel B shows treatment effects on the fraction of participants who anchor to the likelihood or Bayesian mode divided by the fraction who anchor to the base rate. Whiskers show +/- one standard error.

In a broad class of Bayesian or quasi-Bayesian models people integrate the prior and the likelihood, with a greater revision in beliefs if the likelihood is higher. This is inconsistent with the role of contrast, which shows that – rather than integrating the base rate and the likelihood – people select one piece of information out of many. Consistent with our model, a higher likelihood causes a sharply bimodal adjustment of beliefs: a fraction of people shifts to anchoring to the likelihood, increasing neglect of the base rate, while a fraction of people continues to neglect the signal.[15]

We next show that prominence reconciles the balls and urns and taxicabs formats. The middle graphs of Figure 6 compare balls and urns when the match feature is made salient, T_H, versus T_B when it is not. Panel A shows that by describing the problem in terms of the match feature, T_H dramatically increases the share of participants who anchor to the likelihood compared to standard balls and urns T_B, in absolute terms (22.8% vs 15.5%, p<0.01) and relative to the base rate (2.2 vs

[15] Augenblick, Lazarus, and Thaler (2021) find that average beliefs underreact more for higher likelihoods. Their format is different from ours in several respects, but their finding about average beliefs is consistent with our model: it arises when the fraction of people anchoring to the likelihood increases slowly with the likelihood itself. This condition holds in our data: in terms of odds ratio, mean beliefs for A are twice as high for T_{ME} than for T_{LE}, compared to the Bayesian benchmark in which it should be three times higher.

0.8, p<0.01), in line with Corollary 5. There is also a modest reduction in the relative prevalence of the Bayesian answer. Similarly, the right graphs of Figure 6 show that the "undermining the witness" treatment T_U, designed to reduce the salience of the signal relative to the base rate, increases anchoring to the base rate and decreases anchoring to the likelihood: one feature crowds out another, despite the fact that statistics are unchanged.

If these changes in bias are due to the changing salience of specific features, attention to these features should change accordingly, as in Corollary 5. To see if this is the case, Figure 7 plots on the x axis the share of subjects 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 at the base rate. Panel A reports the results using the direct elicitation measure, Panel B using the free response measure. Both measures of attention are consistent with Corollary 5. 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.

These results underscore the centrality of shifting bottom up attention for understanding bias. The evidence does not support a stable mapping between objective probabilities and judgments, nor the primacy of a specific statistic (the base rate in under-reaction models, the likelihood in base rate neglect ones). The evidence supports a mapping between attention and estimates, so that changes in salience can reconcile various biases and their instability. While "balls and urns problems" are worded in a way that makes the individual urns A and B more prominent, the statistically equivalent base rate neglect problems, e.g. cabs, are worded to highlight how the signal is similar to the underlying hypothesis. To understand which biases are dominant in a given setting, one needs to go beyond objective probabilities and independently measure attention and feature salience.[16]

[16] A distinction has also been drawn between balls and urns and "forecasting", in which overreaction also prevails (Fan Liang, and Peng 2021). One explanation is that forecasting tasks (in which people must guess a future signal rather than the urn the current signal comes from) also make signals more salient compared to inference, fostering overreaction.

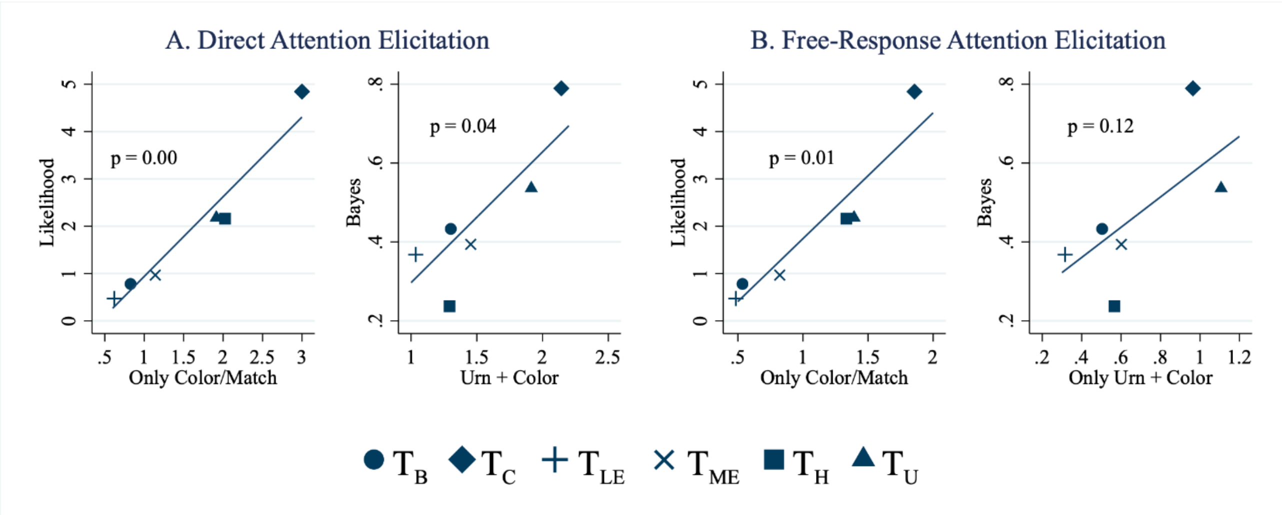

Figure 7. Treatment effects on beliefs and attention. The x-axis is the fraction of participants in each treatment attending to color and/or match (left figure within each panel) and to urn + color (right with each panel) divided by the fraction attending only to urn according to our direct-elicitation (Panel A) and free-response (Panel B) measures. The y-axis is the fraction of participants who anchor to the likelihood (left within each panel) or close to the Bayesian answer (right within each panel) divided by the fraction who anchor at the base rate.

Model Estimation. We provide a structured test of our model by estimating it via maximum likelihood (details are in Appendix C). This allows us to infer the latent cognitive primitives of contrast and prominence from observed probability estimates and assess whether the pattern of attention predicted by the model matches measured attention out-of-sample. We test two additional restrictions. First, the treatment-level prominence of the ancillary feature ("match") should be associated only with increases in measured attention to "match" itself, not to Bayes. Second, the estimates tell 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:

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

where $(P_e - P_U)$ is the prominence of attention profile α_e, while the second term is its contrast, all relative to urn selection. $C(\alpha)$ is pinned down by the statistics of the problem, but here we test whether $\beta > 0$. The constant in (8) captures the relative prominence of e. Figure 8 plots on the x axis model-implied salience and on the y axis measured attention to the same feature profile.

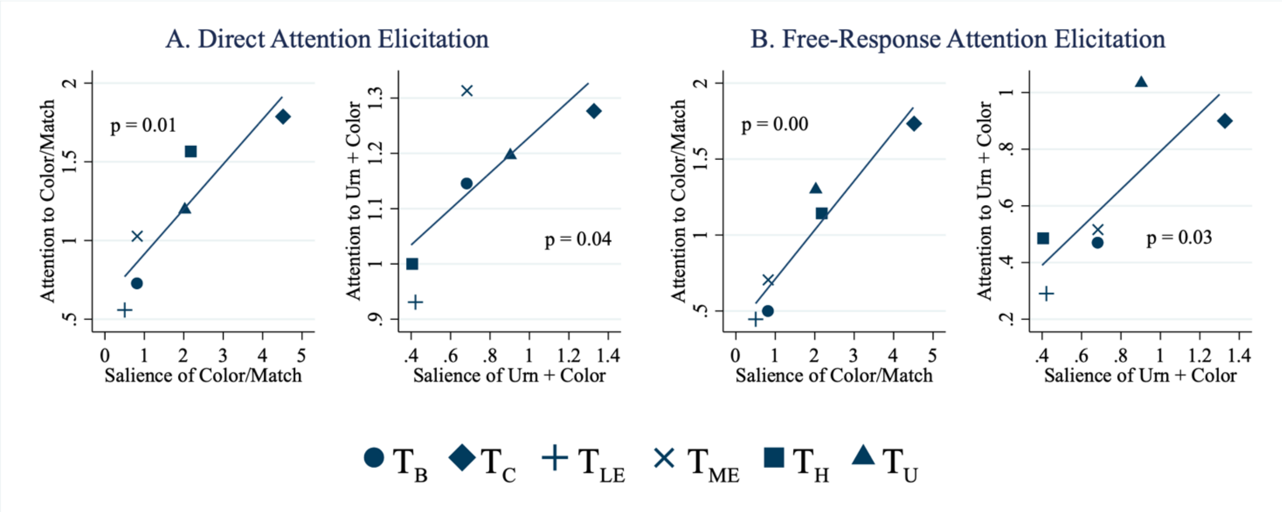

Figure 8. Measured vs revealed attention to features. The x-axis is the estimated salience of each attention profile (where we sum together the color and match salience estimate) relative to the estimated salience of urn. The y-axis is the share of participants who attend to the corresponding profile, as measured by our direct elicitation (Panel A) or free-response measure (Panel B).

First, measured attention is 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. Second, contrast matters: the coefficient on contrast is estimated as $\beta = 1.2$, with a 95% bootstrap confidence interval of [0.55, 1.80]. Third, consistent with our model, the prominence of the "match" feature, 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% ($p<0.01$). Consistent with interference, the salience of match also reduces attention to "only color" (12.3% vs 6.9%, p=0.02).

6. Additional Implications of Bottom-up Attention

We now derive and test additional implications of our approach. Section 6.1 shows that salience may cause the DM to neglect certain hypotheses. Section 6.2 shows that in complex problems, where the attention limit K is binding, partial attention generates the insensitivity of judgments to sample size (Kahneman Tversky 1972) and to the weight of evidence (Griffin and Tversky 1992).

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 by failing to imagine that the latter 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, the salience of a hypothesis can be shaped by its prominence. In balls and urns we described hypotheses as "what is the probability that the ball is drawn from A vs. B?" The same question could be phrased as: "what is the probability that ball is drawn from A?" The questions are identical but the second phrasing, leaving urn B implicit, may allow the DM to neglect B. Thus, she simulates only A and fails to normalize (Task 3).

To see how this works, denote by $\alpha_B \in \{0,1\}$ the attention to hypothesis H_B. The attention profile is $\alpha = (\alpha_1, .., \alpha_O, \alpha_B)$.[17] 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 (9) 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",* $\Pr\big(H_A; \alpha_{A\cap g}\big) = \pi_A \cdot q$.

Neglect of H_B reduces the share of correct answers because the Bayes' rule calls for full attention, including to hypotheses. It also increases the base rate and likelihood modes, which remain feasible because these statistics are already normalized, so they do not need Task 3. Interestingly, DMs who neglect H_B and attend only to "drawing a green ball" exhibit a kind of confirmation bias. They think only about urn A, appreciate that it has q green balls, and thus estimate its probability as q. They seem to confirm their favoured hypothesis A based on its high probability of generating the data, neglecting that green balls are also in B. This logic causes anchoring to A's likelihood q regardless of the color composition of B, which is not the case for the mechanism in Proposition 4.[18]

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

[17] 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.

[18] 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.

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 (9). The deliberate simulation of a specific event further confirms that biases are due to erroneous representations. Remarkably, at this mode the DM sets the probability of A below its 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.

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

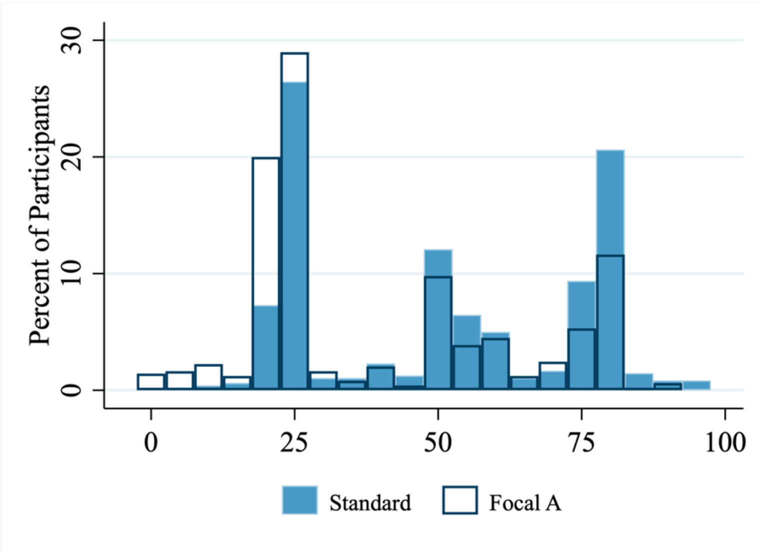

Figure 9. The Figure shows the distribution of beliefs about $\Pr(A \mid g)$.

Keeping the alternative implicit is by all accounts a modest change in description, yet it has a large effect. 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 mode is widespread, even in treatments when H_B is explicit. We did not directly elicit attention to hypotheses, but we can use 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 format to 39.6% in the Focal A one (p<0.01).

The new mode is relevant for the debate on base rate neglect. Gigerenzer and Hoffrage (GH, 1995) showed that more accurate 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 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.

To test whether displaying frequencies is sufficient to promote Bayesian answers, we compare two versions of balls-and-urns where probabilities are described in frequency format. In the standard frequency format, both hypotheses A and B are prominently displayed. In the "focal H_A" frequency format, H_B is implicit. If exposing people to frequencies is enough to promote Bayesian answers, there should be no difference across these versions. If it is also necessary to draw bottom-up attention to the alternative hypothesis, the new mode $\pi_A \cdot q$ should appear in the "focal H_A" frequency format, at the expense of the Bayesian answer. Figure 10 compares the distribution of answers in the standard frequency format (Panel A) and the "focal H_A" format (Panel B).

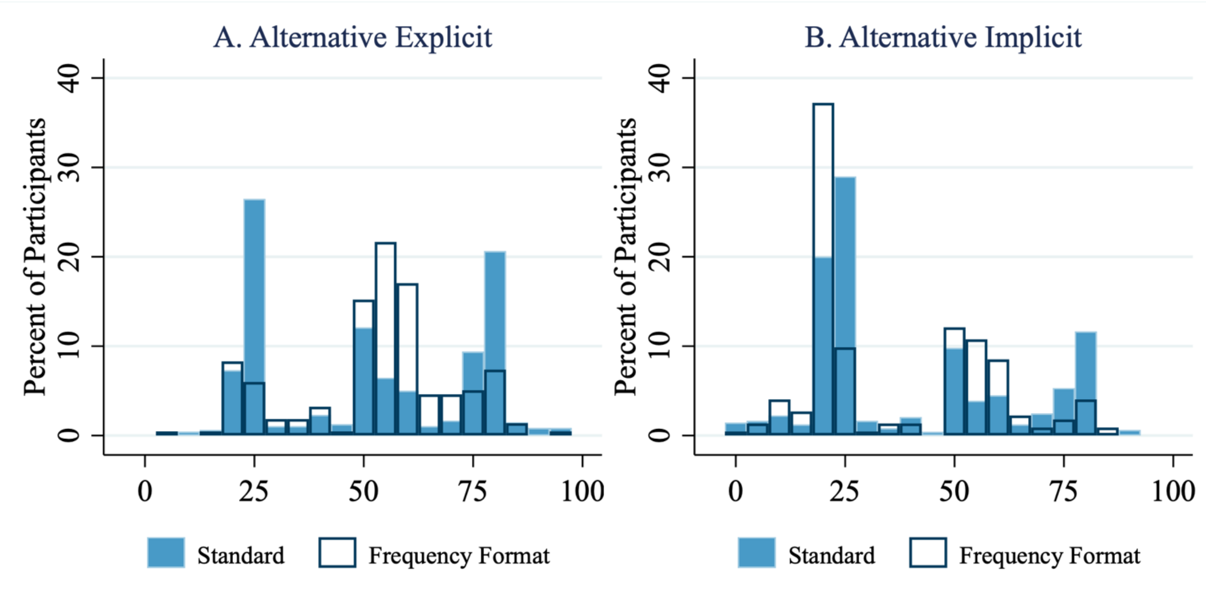

Figure 10. Balls and urns in baseline and frequency formats. Each panel shows the distribution of $\Pr(A \mid g)$.

The results are strongly in line with our model. 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. Consider Panel B: as alternative B is made less salient in the "focal H_A" version, the new "$A \cap g$" mode at

[19] The frequency 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 work studies the effect of training and communication of statistics (Visschers et al 2009, Gigerenzer 2014, Operskalski and Barbey 2016).

20% is strikingly dominant. The benefit of the frequency format over the standard format is no longer clear: in the former many people estimate A to be below its base rate despite the favorable signal.[20]

As this example illustrates, it is too optimistic to expect naturalistic contexts to reduce biases. Bayes rule typically requires attention to many relevant features, 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, 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.[21]

6.2 Attention limits and Insensitivity in complex problems

In complex problems, in which the attention limit K is binding, our model yields well known forms of insensitivity of probability estimates to the quantity of data. Intuitively, as the sample size/number of signals grows, so does the number of relevant features, bolstering the role of salience in selecting which ones to attend to, up to the maximum of K, and which ones to neglect.

6.2.1 Insensitivity to Sample Size

For iid processes, Kahneman and Tversky (1972) and Benjamin, Rabin, and Raymond (2016) document a strong "insensitivity to sample size": estimated sampling distributions fail to converge to the population mean as the sample size grows. Specifically, 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 true answer is $\Pr(H_1) / \Pr(H_2) = \binom{n}{n/2}$, which increases in n. In experiments, the estimated ratio increases too little, if at all, with n.

Consider how this phenomenon arises 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'

[20] Notably, even in the frequency format a number of participants anchors to the base rate and the likelihood. Our model can produce this result if DMs attend to the now ancillary "color" and "urn" features. Esponda et al. (2022) show that even the power of experienced frequencies is rather weak. Their subjects solve standard "base rate neglect problems" (e.g., taxicabs), and then receive feedback on the joint distribution of signals and states. Despite the feedback, many subjects stay anchored to their initial answers. Stable representations can help explain this fact.

[21] We only considered prominence as a source of hypothesis neglect, but contrast may also play a role. Ba, Bohren and Imas (2023) show that overreaction to data increases when a neutral urn C with a 50-50 color compositions and a large prior probability is added to urns A and B. One explanation of this finding is that, upon observing a green ball, neglect of urn C maximizes contrast. As the DM edits out this urn and its high prior, she strongly reacts to data.

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. As before, ϵ 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 than unbalanced ones. While contrast favours rich representations, 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 $\overline{\pi}(\overline{K})$.*

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 estimated size grows insufficiently with n.

As we show in the proof of Proposition 7, this mechanism 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. Heterogeneity in K therefore yields the heterogeneity in the severity of GF observed in Figure 1. 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, $\binom{n}{n * sh}$. As 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 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

[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 $n_\alpha = \min(K, n)$.

committing the GF 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). Property iii) follows from our model but to our knowledge has not been documented before.

	(1)	(2)	(3)	(4)
	Length 2	Length 4	Length 6	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

Griffin and Tversky (1992) document a strong "insensitivity to the weight of evidence" in inference, where beliefs are insensitive to the number of signals. To see how this can arise in our model, consider the inference problem of Section 2, but allow for multiple draws with replacement from the 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. Or she may neglect urn selection and, as in the case of coin flips, attend to $r \leq n$ ball draws, denoted by α_r. Finally, she 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) *She 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. 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, in turn reducing the contrast associated with the signal itself, which boosts 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). Upon instead receiving the same favourable evidence for A in terms of mixed signals, she may neglect *all signals* 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 11 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]

[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.

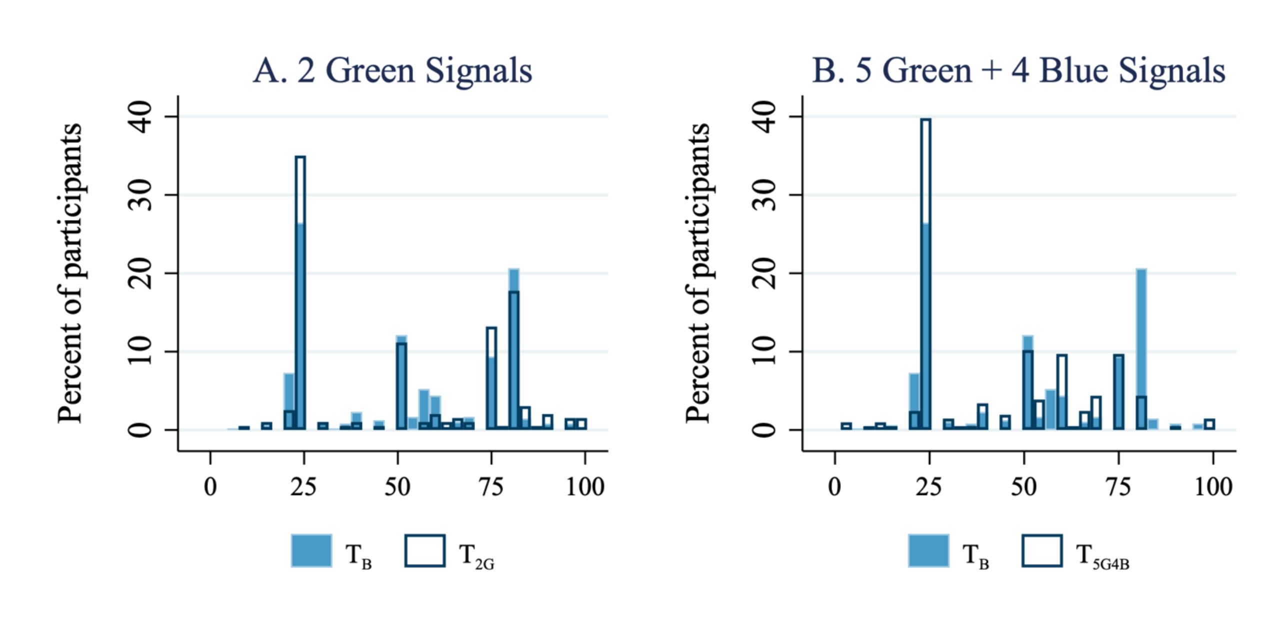

Figure 11. Multiple signals (5 green+ 4 blue and 2 green) in balls-and-urns inference task. Figure shows the distribution of beliefs about the probability of Jar A conditional on the signal(s).

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 statistics 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 point to a 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 to further the decision maker's goals. Work on "rational inattention" in economics or the efficient coding approach in psychology follows this approach. While scarcity of attention is uncontroversial, our analysis challenges the assumption of goal-optimality. 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 suggests 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 allocation. As we showed in previous work, BGS (2012), bottom-up factors can also shape attention towards goal-relevant features. A striking lottery payoff or the surprising 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 goal-driven and bottom-up attention mechanisms is an important avenue for future work.

We conclude by describing other 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 to itself 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. Such a theory of prominence would deepen the account of multimodality and individual fixed effects in solving statistical problems. More broadly, it can shed light on which narratives or partial models people use in different cases, why beliefs diverge despite a great deal of common information, why learning about a process might be hampered by prominent past experiences (Schwartzstein 2014, Esponda, Vespa, and Yuksel 2022), but also why learning can be sped up once neglected relevant features are made prominent (Hanna, Mullainathan, and Schwartzstein 2014, Graeber 2023).

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, characterizing the sources of both disagreement and of average bias in the distribution of estimates. 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 triggers retrieval. This mechanism may be relevant for other well know puzzles such as the hot hand fallacy, but also in the field. For example, the salient losses or failure of an individual bank may draw investors' attention, causing them to selectively retrieve past episodes of financial meltdown, and to neglect the rarity of cataclysmic events and strong pessimism.

The combination of memory and bottom-up attention is also relevant for consumer choice. 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. Selective attention to features, driven by contrast, prominence but also surprise, 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. This process 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

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

Ba, Cuimin, Aislinn Bohren, and Alex Imas. "Over- and Underreaction to Information." Working paper (2022).

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," 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.

De Bruin, Wändi Bruine, Baruch Fischhoff, Susan G. Millstein, and Bonnie L. Halpern-Felsher. "Verbal and Numerical Expressions of Probability: "It's a Fifty–fifty Chance"." *Organizational Behavior and Human Decision Processes* 81 (2000), 115-131.

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, Benjamin. "What You See is All There is," *Quarterly Journal of Economics* 135 (2020), 1363-1398.

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

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

Enke, Benjamin, and Florian Zimmermann. "Correlation Neglect in Belief Formation," *Review of Economic Studies* 86 (2019), 313-332.

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 (2022), 777-789.
Fan, Tony, Yucheng Liang, and Cameron Peng. "The Inference-Forecast Gap in Belief Updating." SSRN 3889069 (2021).
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.
Graeber, Thomas. "Inattentive inference." *Journal of the European Economic Association* 21, no. 2 (2023): 560-592
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 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," *Quarterly Journal of Economics,* forthcoming (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.

Rabin, Matthew, and Dimitri Vayanos. "The Gambler's and Hot-Hand Fallacies: Theory and Applications." *The Review of Economic Studies* 77 (2010), 730-778.

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.

Simon, Herbert. "Models of Man". New York: Wiley, 1957.

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.