Review of Economic Studies (2026) **93**, 250–285 https://doi.org/10.1093/restud/rdaf022

Advance access publication 22 May 2025

How People Use Statistics

Pedro Bordalo
Saïd Business School, University of Oxford, UK
John Conlon
Carnegie Mellon University, USA
Nicola Gennaioli
Bocconi University and IGIER, Italy
Spencer Kwon
Brown University, USA
and
Andrei Shleifer
Harvard University, USA

First version received September 2023*; Editorial decision October* 2024*; Accepted April* 2025 *(Eds.)*

For standard statistical problems, we provide new evidence documenting (1) multimodality and (2) instability in probability estimates, including from irrelevant changes in problem description. The evidence motivates a model in which, when solving a problem, people represent each hypothesis by attending to its salient features while neglecting other, potentially more relevant, ones. Only the statistics associated with salient features are used. 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 for how changes in the salience of specific features jointly shapes known biases and measured attention to features, but also create entirely new biases. We test and confirm these predictions experimentally. Salience-driven attention to features emerges as a unifying framework for biases conventionally explained using a variety of stable heuristics or distortions of Bayes' rule.

Key words: Selective attention, Inference, Overreaction, Underreaction, Disagreement, Representation

JEL Codes: D01, D81, D9

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 (GF), 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

The editor in charge of this paper was Elias Papaioannou.

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 Bayes' rule.

Why do people make these systematic mistakes? And why are these mistakes unstable, changing from one problem to the next and across 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 GF (Rabin, 2002), sample size neglect in i.i.d. draws (Benjamin *et al.*, 2016), base-rate neglect (Grether, 1980), and underreaction in inference (Enke and Graeber, 2023), but does not connect biases across problems or even within different versions of the same problem.[1]

We build a new model based on selective attention to address these questions. When assessing a hypothesis, the decision maker (DM) focuses on its salient features and neglects other features, even if relevant. The hypotheses are thus incorrectly represented, a form of question substitution (Kahneman and Frederick, 2002). Incorrect representation leads to incorrect beliefs. This mechanism unifies different biases and their instability based on well-known regularities in salience-driven attention.

As motivation, Section 2 documents that the distribution of estimates in famous statistical problems is multimodal and unstable. When judging the relative likelihood of two sequences of a fair coin, some people commit the GF, while others give the correct 50:50 answer. The share of correct answers falls as sequences get longer, even though the correct answer remains the same. Similarly, when inferring the probability of a hypothesis based on a noisy signal, some people anchor to the prior, others to the likelihood, and very few integrate the two (see also Dohmen *et al.*, 2009). Again, changes to the formulation that leave the correct answer the same create instability: holding fixed the numerical statistics, describing the signal as the report of a witness in court rather than abstractly sharply raises the neglect of the prior. Multimodality and instability are inconsistent with existing models, which rely on a fixed mis-specified model (Rabin, 2002), stable distortions of Bayes' rule (Grether, 1980), or perceptual noise (Khaw *et al.*, 2021; Enke and Graeber, 2023).

To see how this evidence connects to selective attention, consider the famous duck–rabbit illusion, in which a drawing can be interpreted as either a duck or a rabbit. Some people focus on the beak and see a duck, while others focus on 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, nor do people represent the picture as a mixture of duck and rabbit. Selective attention leads to only one representation at a time.[2] When sentencing a confessed bank robber (Clancy *et al.*, 1981), some judges focus on the defendant's age, others on whether he was armed, and still others on how much money he took, leading to different sentences for the same confessed 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, 2024). In these examples, selective attention to features yields sharply 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* versus *hhhhhh*,

1. The vast majority of this research is concerned with either inference problems or i.i.d. random sequences and sampling distribution problems. In Benjamin's (2019) review, out of the 123 cited papers that experimentally elicit beliefs in statistical problems, 91 are on inference, 14 cover random sequences, and 23 cover sampling distributions (some cover both). No paper to our knowledge jointly covers inference and random sequence problems.

2. In line with this intuition, a literature in neuroscience documents competition between different representations as a key neuronal mechanism underlying visual attention (McClelland and Rumelhart 1981; Desimone and Duncan 1995), which can also lead to instability in visual representations over time.

people may focus on the individual flips of each sequence, or on the sequences' share of heads (0.5 versus 1). When judging the probability that a green ball comes from urn *A* (versus *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.

Following Bordalo *et al*. (2022), we formalize two drivers of attention: contrast and prominence. First, a feature has high contrast and draws attention if it sharply discriminates between different options. In consumer choice, price has high contrast if one good is strikingly cheaper than another. In statistical problems, a feature analogously has high contrast if it sharply favours one hypothesis over another. Second, a feature can be prominent based on the description of the problem, drawing attention because it is highly visibly displayed, as in Chetty *et al*.'s (2009) study of sales taxes. Similarly, different wording or framing of equivalent statistical problems may highlight their different features, bringing some but not others prominently to mind even though statistics are unchanged. In our approach, changes in the contrast or prominence of specific features may cause sharp changes in attention and thus representations, delivering the instability in measured biases.

The model accounts for and reconciles well-known biases in judgments about i.i.d. draws and inference based on the same mechanism, delivering multimodality and instability in both domains. It also makes two new predictions, which we test experimentally. First, our framework predicts a connection between bias and attention to specific features. In i.i.d. draws, DMs committing the GF should be more likely to attend to the sequences' share of heads. In inference, overreacting DMs should be more likely to attend to the signal relative to the prior, whereas people arriving at the Bayesian answer should attend to both features. We measure attention by: (1) analysing free-response reports on how people solve problems, (2) having participants select features they attended to from a list, and (3) eliciting similarity judgments. Across all methods, the model's prediction is confirmed.

Our second and key prediction is that exogenous changes in the salience of a feature should cause joint shifts in attention and the distribution of estimates. To test this prediction, in i.i.d. draws we make individual flips prominent by describing the same hypotheses in terms of the flips that differentiate them, and show that doing so 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, and increases the share of people who neglect the base rate. We also increase the prominence of the match between the signal and different hypotheses by describing the likelihood in terms of the similarity between the two, and show this increases measured attention to this feature and anchoring to the likelihood. This mechanism accounts almost fully for the large shift in estimates from the balls and urns format (Edwards, 1968), in which many people anchor to the base rate, to the "taxicabs" format (Kahneman and Tversky, 1972), in which they anchor to the likelihood.

Our model explains why presenting inference problems in a "frequency format" (Tversky and Kahneman, 1983; Gigerenzer and Hoffrage, 1995) promotes Bayesian answers: it curbs the neglect of either the base rate or the likelihood. This format is no panacea, though. To show this, 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 predicted by our model: many people estimate the described hypothesis as the product of its base rate and likelihood, fully neglecting the alternative. Salience-driven attention thus casts doubt on the general ecological rationality of human intuition and highlights the sensitivity of judgment to irrelevant features of context.

We explain biases attributed to heuristics such as availability, representativeness, or anchoring (Kahneman and Tversky, 1972; Gigerenzer, 1996) as by-products of selective attention to

features. We formalize such attention using insights from psychology and machine learning (Selfridge, 1955; Tversky, 1977; Guyon and Elisseeff, 2003; Kruschke, 2008). Compared with models of goal-optimal attention (Sims, 2003; Woodford, 2003, 2020; Gabaix, 2019), we explain why highly relevant information can be neglected while irrelevant changes shape attention and biases. In our set-up, prominence is treated as a latent variable which is not theoretically micro-founded but is disciplined through measurement. Bordalo *et al*. (2025b) generalize the current model to include a theory of prominence based on experienced categories and visually salient cues.

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 *et al.*, 2023, Esponda *et al*., 2024) or computational complexity (Simon, 1957; Enke and Zimmermann, 2019; Enke, 2020; Graeber, 2023). In our model, bias arises even in computationally simple problems because of selective attention and incorrect representation (see also Ba *et al*., 2024). In coin flips, it is trivial to avoid the GF by recognizing that each flip is 50:50. Bias arises because a particular feature, the share of heads, is salient but irrelevant for the problem at hand. Moreover, shifts in salience lead to instability of choices, whereas much of earlier work focuses on stickiness in biases.

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 "i.i.d. 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% points of the correct ones. Supplementary Appendix B describes the experimental protocol and pre-registration.

For i.i.d. 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. Figure 1A and B shows the distribution of beliefs about the relative share of the unbalanced sequence for each treatment.

As in previous studies (Benjamin, 2019), the mean response is <0.5, confirming the GF, the belief that a specific balanced sequence is more likely than an unbalanced one. There are, however, two new findings. First, the 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 partly occurs because the share of people answering *exactly* 50% drops by about 14% points (54.8% in panel A versus 40.7% in panel B, $p = 0.00$), with an increase in answers around 5%.

Instability in the share of people committing the GF is inconsistent with a heterogeneous yet stable tendency to use a mis-specified sampling model (Rabin, 2002). In Rabin and Vayanos (2010), the GF can become more severe when sequences get longer only on the intensive margin, *i.e.* in the extent of error conditional on committing the GF. Comparing Panel A to Panel B, in contrast, there is also a sharp increase in the *extensive* margin of people committing the GF. The difference in the extensive margin suggests 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?

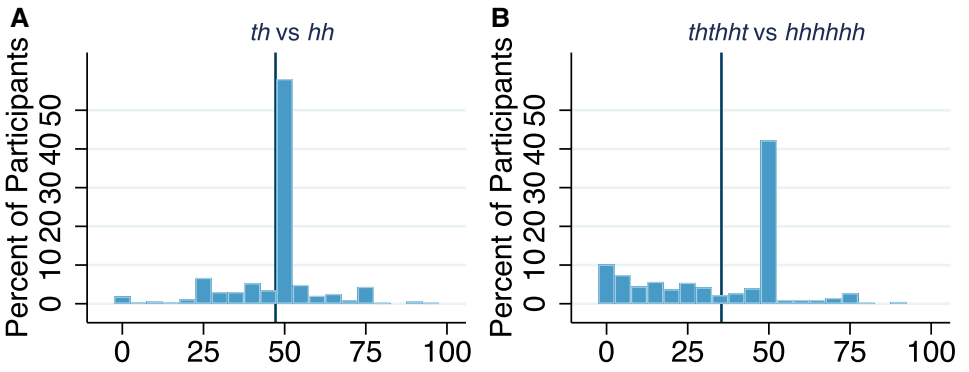

FIGURE 1

Panels A and B report the distribution of estimated $\Pr(H_2|H_1 \cup H_2)$ for the two treatments described in the text. Answers closer to 0 indicate higher probability of the balanced sequence H_1. The blue bar marks the mean answer

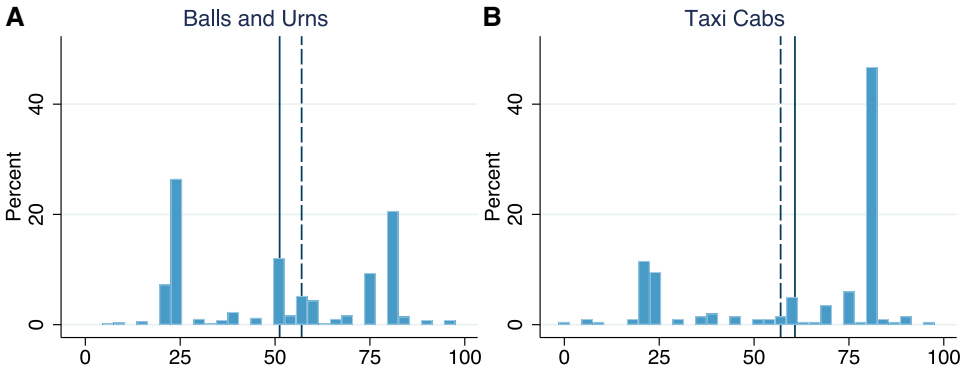

FIGURE 2

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

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 versus B. In the more naturalistic "cabs" treatment (Kahneman and Tversky, 1972), participants were instead told there are two taxicab companies, the Blue and the Green, according to the colour 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 colour cab with probability 80%. They are the asked the probability that the errant cab was indeed green versus 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 $\Pr(A|g) = \Pr(\text{Green}|g) = 0.57$ in both problems. The distribution of answers is reported in Figure 2.

Consistent with previous work (Benjamin, 2019), in balls and urns (Panel A) underreaction to the data prevails on average: the mean answer (solid line) is 52%, lower than the correct

answer (dashed line). There is however pronounced multimodality: many answers cluster on the base rate 25%, the likelihood 80%, and 50%. Where do these different modes come from?

Crucially, there is also instability: in the taxicab frame (Panel B), many more people anchor at or around 80%, so on average they overreact. 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 heuristic. 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 colour of the drawn ball, answering with the base rate. In taxicabs, they instead neglect the baseline frequency of blue cabs, answering with the likelihood. Why are different features neglected in different frames?

We also ran the treatments of Figures 1 and 2 within subjects (see Supplementary Appendix B for details). The results confirm both multimodality and instability. On one hand, at the individual level, there is a systematic heterogeneity in how subjects represent a given problem. In i.i.d. draws, subjects who commit the GF in one problem are more likely to do the same in a similar problem. In inference, subjects anchoring to the base rate/likelihood in one problem are more likely to do the same in a subsequent similar problem. On the other hand, comparing across different settings, there is instability in how a person represents a problem: the shift in beliefs in Figures 1 and 2 is almost fully explained by subjects who move between the dominant modes. For i.i.d. draws, a large fraction of subjects judge *ht* as equally likely as *hh*, and yet commits the GF when judging six flip sequences. In inference, the instability in Figure 2 is in good part explained by subjects who switch from anchoring to the base rate in balls and urns to anchoring to the likelihood in taxicabs.

These findings raise two challenges. First, summarizing beliefs in an experiment by the mean or modal response is highly misleading in the presence of multimodality. In Figures 1 and 2, hardly anyone is near the mean. This is especially clear 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 fail to elicit what people do naturally: grasp at straws in a complex situation. Answers to standard statistical look like the duck–rabbit problem.

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: underreaction 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 rate (underreaction) or the likelihood (overreaction). To account for these findings, we need a new framework.

3. THE MODEL

We present a model in which the patterns described in Section 2 arise from selective 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 attention.

Formally, a statistical problem has three components: (1) the sampling process, (2) the statistics, *e.g.* the probabilities of specific events, and (3) 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 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 Supplementary 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: the elementary events $\omega \in \Omega$ constituting the 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 $k_1 = h, t$. A generic 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 colour 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.

A rational solution consists of: (1) expressing each hypothesis as a partition of the events about which statistics are provided, (2) computing the probability of each hypothesis using these statistics, and (3) normalizing the estimate if the probabilities in (2) do not add up to one, *i.e.* if $H_i \cup H_{-i} \subset \Omega$. Sometimes different partitions of hypotheses exist, but they all lead to a correct answer.

We describe 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 attention to features: contrast and prominence. Section 3.4 describes how to apply the model and test its predictions in the laboratory, offering some guidance for 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, $k_2|k_1$, ... 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 the *true probability* $\Pr(f_j)$ 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 colour 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, they capture observable properties of the event, *i.e.* equivalence classes to which it belongs. In coin flips, one such feature is a sequence's "share of heads", denoted by $sh \in [0, 1]$, which identifies the class of sequences with the same share of heads as ω. This is

a notable feature for it is connected to similarity: $(h,\ t)$ is similar to the fair coin that produced it because its 0.5 share of heads is what a fair coin tends to produce.[3] 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 colour of the ball is similar to the urn, and $m = 0$ otherwise. This feature defines two equivalence classes: $(A,\ g)$ and $(B,\ b)$ are events in which the signal and the hypothesis are similar, $m = 1$, while in $(A,\ b)$ and $(B,\ g)$ they are dissimilar, $m = 0$.[4]

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 $j = 1,\ \ldots,\ n + 1$, 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 well-established attention limits. 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 then distorts representations as follows.

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

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

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

3. 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".

4. Inference problems with many draws may have more ancillary features, such as the match between the distribution of draws and the urn composition. Attention to this feature provides a foundation for the notion of exact representativeness (Camerer 1987, 1990; Grether 1980), whereby subjects overestimate the probability a set draws came from an urn with the corresponding composition.

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 $H_1 = (h,\ h)$ versus $H_2 = (h,\ t)$. 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:*

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

The DM computes the 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 $\Pr(\tilde{f}_r \cap \cdots \cap \tilde{f}_s)$ by multiplying their probabilities. She then sums the products across all vectors. A DM attending to individual flips simulates $H_1 = (h,\ h)$ and $H_2 = (h,\ t)$ by multiplying the 0.5 statistic attached to these features, $\Pr(R_\alpha(H_1)) = \pi_{h_1} \cdot \pi_{h_2} = (0.5)^2$ and $\Pr(R_\alpha(H_2)) = \pi_{h_1} \cdot \pi_{t_2} = (0.5)^2$. If instead the DM attends to the share of heads, she simulates representations $R_\alpha(H_1) = (\varphi,\ \varphi,\ 1)$ and $R_\alpha(H_2) = (\varphi,\ \varphi,\ 0.5)$, computing the probability of the same hypotheses as $\Pr(sh = 1) = (0.5)^2$ and $\Pr(sh = 0.5) = 2 * (0.5)^2$, respectively. Different representations focus the DM on different features, leading to different simulations. The final step normalizes the simulated probabilities.

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

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

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 $H_1 = (h,\ h)$ versus $H_2 = (h,\ t)$ 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* $\alpha^*(\omega) = (\alpha_1^*,\ \ldots,\ \alpha_{n+1}^*)$, $\omega \in H_i \cup^{H_{-i}}$, *containing at least one zero such that a DM using attention* $\alpha^*(\omega)$ *in Task 1 and then following Tasks 2 and 3, implements 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 1–3 guarantees a correct solution.[5] But what shapes attention? We address this question next.

3.3. *Salience-driven attention to features*

Selective attention comes in several forms. It can be goal-optimal, as for example in rational inattention models (Sims, 2003; Woodford, 2003, 2020; Gabaix, 2019; Khaw *et al.*, 2021). Attention can also reflect a focus on salient stimuli which causes neglect of other less salient, even if relevant, ones (Bordalo *et al.*, 2012, 2013, 2022; Evers *et al.*, 2022; Li and Camerer, 2022; Conlon, 2025). Sometimes the salient feature is relevant to solving the problem but by drawing attention away from other relevant features still distorts the decision. While driving, a surprising police radar may cause us to neglect the car behind us, and brake too heavily. But a feature may draw attention even if it is entirely irrelevant to the current task, such as when a stain on the wall distracts us from a conversation.

Section 2 showed that different people use different statistics, both within and across problems, despite having the same incentives for accuracy: they do not choose the "most accurate" statistics for a given attention limit K, as for instance in models of sparsity (Gabaix, 2014) or other approaches to optimal selective attention. Instead, the evidence suggests that systematic variation in attention is driven in part by salience, which generates instability in representations and the use of statistics. We focus on two drivers of such salience (Bordalo *et al.*, 2022): contrast and prominence. In visual attention, a feature is contrasting if it sharply differs from the background (*e.g.* a dark stain on a white wall, a very low price). A feature is prominent if the description of the problem brings it top of mind, either through prominent display or through its proximate relevance for solving the problem (home runs are top of mind, but walks are not, leading to errors in the evaluation of baseball players, see Lewis, 2003).

We formalize these forces using salience theory (Bordalo *et al.*, 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 favours one of the goods. In a statistical problem, we likewise say that attending to a feature induces contrast if it sharply favours one hypothesis over the other. Formally, the contrast of α is

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

The numerator captures the extent to which the representation favours one hypothesis over the other, while the denominator captures diminishing sensitivity, as in Bordalo *et al.* (2012, 2013). To illustrate, when assessing $(h,\ h)$ versus $(h,\ t)$, 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

5. Another attention limit implicitly imposed in Task 1 compared with 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. 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.

shaped by the probability of a head and the sequence length n. In inference, it depends on the base rate and the likelihood. In our experiments, we manipulate contrast by changing statistics.

Next, consider prominence. In Bordalo *et al*. (2022), as in Chetty *et al*. (2009), an attribute such as the price or sales tax is more prominent 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 top of mind. Some formal ingredients of the problem, such as the data-generating process and the hypotheses H_1 versus H_2, can be described in a way that makes a specific feature prominent. In balls and urns, the composition of the urns could be described as "80% of the time the colour of a drawn ball matches the colour of the urn (green versus blue) it comes from". This data-generating process is equivalent to that in Section 2 (where the contents of each urn are described directly), but the description makes the "match" feature more prominent. Likewise, describing the hypotheses as "Urn A" versus "Urn B" as in Section 2 makes the urn selection feature more prominent than describing the hypotheses as whether the ball "matches" the colour of the urn, though the two ways of describing hypotheses are logically identical.[6]

In our experiments, we manipulate prominence by changing the description of the problem in ways that intuitively bring certain features top of mind. We validate our manipulation by separately measuring prominence using similarity judgments and attention measures. Denote the prominence of feature j as a latent scalar P_j. We let the prominence of attention profile α, denoted $P(\alpha)$, be the average prominence of its features:

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

which is a latent variable that can be recovered from measured attention α. 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, *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, in the sense that people tend to attend to one feature or another, but rarely both. 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 $P(\alpha)$, but also in an individual-specific extreme-value term ϵ_α. The latter captures transient fluctuations in attention, but also stable individual differences in the prominence of specific features (for example, due to past experiences of attending to them). To simplify, we assume salience is additive in these terms.

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

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

6. As discussed in Bordalo *et al*. (2022), much early work on prominence comes from the field of visual attention. This field has identified reliable predictors of visual attention such as a stimulus' visual contrast and its centrality in the visual field. These correspond to our notions of contrast and prominence in description. This literature also shows that attention is spontaneously drawn to stimuli that have recently been attended to, even if these stimuli are not relevant in the current problem (Remington *et al*., 1992). This links prominence of a feature to its use in past problems, a channel we abstract from here.

The term $\epsilon_{\tilde{\alpha}}$ yields 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. Statistical features are explicitly given by the problem. Ancillary features need not be explicitly mentioned, but can be discovered by intuition (*e.g.* our ancillary features were motivated by representativeness, Kahneman and Tversky, 1972), directly asking people for a rationale for their choices, or using text analyses or algorithms.[7] Specifying/discovering features is the key first step.

Once a set of 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. Several approaches to measuring attention are available. Eye tracking (Reutskaja *et al.,* 2011) is often used to capture visual attention, but for our purposes a measure of semantic attention, related to how a problem is solved, is more useful. We pre-registered three approaches to such measurement. 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 large language model to code these responses according to whether the participant appeared to pay attention to specific features (see Supplementary Appendix B for details). Second, after the free-response, a multiple-choice question asks participants to select from a list the features they felt they attended to. Third, we ask respondents to rate the *similarity* between events and infer attention from these ratings. The connection between similarity and attention to features is well established (*e.g.* Tversky and Gati, 1982; Nosofsky, 1988): people judge two objects to be more similar when they agree on salient features. Conversely, this insight implies that similarity measurements can be used to assess the prominence of a feature independently of the original probability judgment task.[8] We check whether different measures yield comparable results.

7. Kleinberg *et al.* (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 2024). This feature was discovered via machine learning, rather than specified by the analyst ex ante.

8. 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 $j = 1,\ \ldots,\ F$ 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. Empirically, the psychology literature has developed the technique of multidimensional scaling (Torgerson 1952), which uses similarity assessments to embed objects in Euclidean space. The embedding function is then informative of which features weigh prominently in judgment. See Nosofsky *et al.* (2018) for a recent application.

With a set of features and measured attention to them in hand, 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 across settings (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 THE GF

We first apply our model to coin flips. It delivers the patterns in Figure 1 and yields new predictions, which we test, on how changes in the description of the problem affects measured attention to features and the GF.

4.1. *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 versus 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 $f(\omega) = (f_1,\ \ldots,\ f_n,\ sh)$.

4.2. *Attention and representation*

A DM attending to $r \leq n$ statistical features, individual flips, while ignoring the share of heads, $\alpha_r = (1,\ 1,\ \ldots,\ 0)$, represents the generic hypothesis by $R_{\alpha_r}(H_i) = (f_1,\ \ldots,\ f_r,\ \varphi)$. This DM behaves rationally: by Equation (2) she simulates $\Pr(R_{\alpha_r}(H_i)) = (0.5)^r$, which is identical across hypotheses, yielding after normalization the correct estimate $\Pr(H_1|\alpha_r) = 0.5$. In contrast, a DM attending only to the share of heads, $\alpha_{S,n} = (0,\ \ldots 0,\ 1)$, represents hypotheses as $R_{\alpha_{S,n}}(H_i) = (\varphi,\ \ldots,\ \varphi,\ sh)$. 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.

4.3. *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 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 GF, estimating the relative probability of the unbalanced sequence as:*

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

The remaining DMs attend to a subset of flips and answer 50:50.

There are two modes for beliefs *and* attention: one at 50% with attention to individual flips, another below 50%, as in Equation (7), with attention to the share of heads.[9] The key new

9. 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.

TABLE 1
Treatments manipulating salience in the GF problem

Treatment	N	Summary	Purpose
T_2	434	Balanced versus unbalanced 2-flip sequences	Compare with T_6
T_6	405	Balanced versus unbalanced 6-flip sequences	Increase contrast of share compared with T_2
T_{full}	1,038	Ask about full 6-flip sequences $H_1 = hhhhht$ versus $H_2 = hhhhhh$	Compare with T_{last}
T_{last}	978	Ask about final flip. in 6-flip sequences	Increase prominence of final flip compared with T_{full} and thereby reduce attention to share heads *i.e.* P(h versus t | $hhhhh$)

prediction besides multimodality is the connection between judgments modes and measured attention. 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 with a 50:50 outcome. The share of heads sticks out as a salient representation, and for many DMs it replaces the original question. Thus, our model explains the fall in the 50:50 mode when moving from Figure 1A to B: it is caused by the higher contrast of the share of heads when $n = 6$ compared with $n = 2$.[10] 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.

4.4. *Coin-flip experiments*

Table 1 provides a summary of the treatments. In all treatments, individuals are asked to judge the relative likelihood of two given sequences and 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 $H_1 = hh$ versus $H_2 = th$ and $H_1 = hhhhhh$ versus $H_2 = ththht$, respectively. We also introduce two new treatments to study the role of prominence. In T_{full}, subjects are asked to estimate

10. In our model, the severity of the GF, conditional on committing it, 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.

TABLE 2
Correlating measures of attention with the GF

	Dependent Variable: Commit the GF		
	(1)	(2)	(3)
Directly elicited attention to share	0.169***		0.174***
	(0.017)		(0.032)
Free-response attention to share	0.083***		0.092***
	(0.017)		(0.032)
Similarity between judged sequences		−0.078***	−0.076***
		(0.021)	(0.020)
Treatment FEs	Yes	Yes	Yes
N	2,855	846	846
R^2	0.110	0.093	0.137

Table shows ordinary least squares 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.

$H_1 = hhhhhh$ versus $H_2 = hhhhht$, 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.

We validate our treatment by separately measuring attention to features—(1) the share of heads, (2) whether the final flip is heads or tails, and (3) anything else—in three complementary ways, as outlined in Section 3.4. First, after eliciting participants' probability assessments, we independently measure free-response and direct-elicitation proxies for attention to features. Next, for a subset of participants, later in the survey we also elicit perceived similarity between sequences, independently of the original probability assessment. In line with the discussion in Section 3.4, this assessment should reflect the prominence of sequences' share of heads: if the share of heads of a sequence is prominent to the DM, then two sequences H_1 and H_2 should be less similar if the sequences differ sharply along *sh*.[11] We interpret perceived similarity as a more direct proxy for prominence of the share of heads, whereas the first two measures reflect the total salience of features, which may also be driven by contrast.

Across the four treatments, we test two predictions. First, by Proposition 2, a participant's attention to the ancillary feature *sh* should be positively correlated with her tendency to commit the GF. Second, by Corollary 3, there should be instability across treatments: the share of participants committing the GF and those attending to the share of heads should be greater for longer sequences (T_2 versus T_6), due to contrast, and smaller when individual flips are more prominent (T_{full} versus T_{last}).

4.5. *Multimodality in attention and estimates*

We first document multimodality in attention and probability estimates within each treatment. Pooling across all treatments and adding treatment fixed effects, we run regressions of a respondent-level indicator for whether she commits the GF (*i.e.* reports a belief of <50 out of 100 for the unbalanced sequence) on indicators for directly elicited and free-response attention

11. Using the similarity function in footnote 6, if the DM attends to all individual flips the similarity between a balanced and an unbalanced sequence is 0.5, if she attends to the share of heads it is zero.

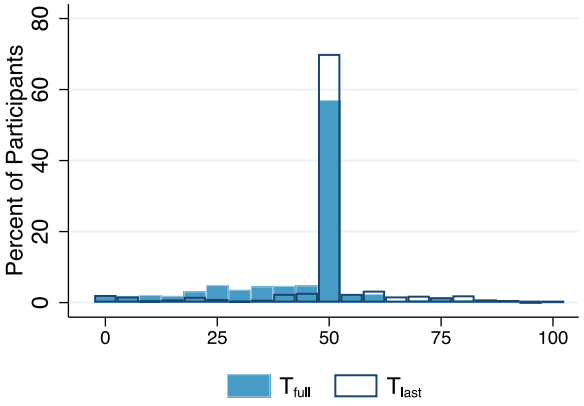

FIGURE 3

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

to share of heads (Table 2, Column 1), on the perceived similarity between sequences (Column 2), and on all three attention proxies (Column 3).

Consistent with our model, a subject attending to the share of heads is more likely to commit the 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 the 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.

4.6. *Instability in beliefs and attention*

Consider instability next. In Figure 1, as we saw in Section 2, increasing sequence length from $n = 2$ to $n = 6$ increases the incidence of the 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 versus 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% versus 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 representation of hypotheses.[12]

We next test whether treatment effects in beliefs correspond to changes in attention, which proxy 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

12. Previous literature finds some prevalence of the GF in settings where, similar to T_{last}, subjects are asked about the next coin flip in a sequence (Benjamin 2019). Our prediction here concerns not the level of the GF (a positive prevalence in such problems is fully consistent with our model) but the comparative static that arises from making the last flip more prominent. Note also that while we find a significant prevalence of the GF in T_{last}, it is weaker than in our baseline treatment T_6, where the more balanced sequence is *hthtth*. This is in line with our model, since the share of heads has higher contrast-driven salience in T_6, as well as with the literature.

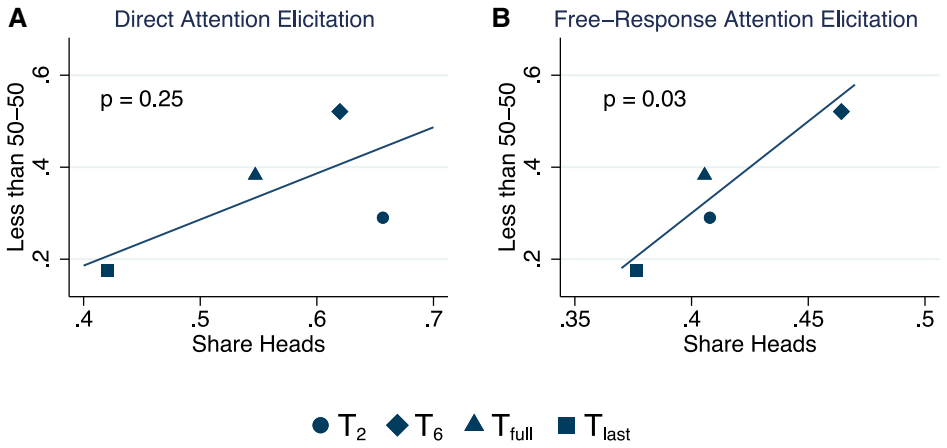

FIGURE 4

Treatment effects on the incidence of the GF 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

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}).[13] Reassuringly, the free-response measure, based on subjects' report of their reasoning, detects model-consistent instability in attention across all treatments. As predicted by our model, instability in the GF is associated with shifting attention to an irrelevant feature, the share of heads. The correlation coefficient between treatment-level attention to the share of heads and the fraction of participants committing the GF is 0.74 for the direct-attention elicitation and 0.95 for the free-response measure.

We conclude by further connecting attention to the share of heads, similarity, and probability 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 Supplementary appendix 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. Intuitively, a balanced sequence is similar to the many other unbalanced sequences despite

13. 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% versus 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 ($p = 0.10$).

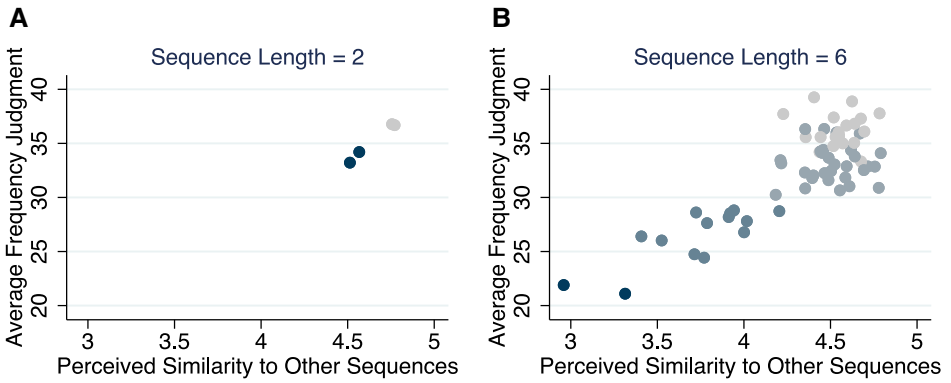

FIGURE 5

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)

the differences in individual flips. Consistent with our model's account of 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). When the DM attends to *sh*, the balanced sequence is 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 Supplementary Appendix 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 next show that salience-driven attention to features accounts for the patterns of beliefs in inference problems shown in Figure 2, and assess new predictions regarding the link between measured attention and beliefs and the instability in inference.

5.1. *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 colour 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 versus 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 $\beta > 0.5$.

5.2. *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 colour, $\alpha_{BR} = (1,\ 0,\ 0)$, represents the problem as "what is the probability that *a ball* is drawn from A versus 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 with 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 colour?", $R_{\alpha_m}(H_U) = (\varphi,\ \varphi,\ m)$. 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 versus another is true?". These DMs think "a green ball could come from either urn" and report 50:50.[14] 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.

5.3. *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 colour of the ball or to "match", α_c and α_m, respectively, and anchor to the likelihood* $\Pr(H_A;\ \alpha) = q$. *The remaining DMs neglect all features and answer* $\Pr(H_A;\ \alpha_0) = 0.5$.

The model predicts, within an experimental treatment, a systematic relationship between measured attention to features—which varies across people due to the shock ϵ_α—and the probability estimate which accounts for the multimodality observed in Figure 2. As in coin flips, the model then also predicts instability. Denote by P_l the scalar prominence of feature $l = U,\ c|U,\ m$.

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

14. Here, no attention to features can also capture the possibility that the DM's attention jumps between "urn selection" and "colour of ball", which favour different hypotheses, without settling on either.

Due to contrast, making the signal more informative boosts the 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 colour 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" versus "B", making urn selection prominent, in taxicabs they are framed as whether "the errant cab is indeed green versus blue (as the witness claimed)", raising the prominence of the match. Lastly, the courtroom context of taxicabs may also increase the prominence of the witness' accuracy for some participants, due to personal or fictional past experiences of witness reports in court. All of these irrelevant changes may shape attention and explain instability. A key prediction of our model, which we now test, is that such description changes should be reflected in changes in attention to specific features.

Proposition 4 and Corollary 5 capture a key, distinctive implication of our model, namely a tight connection between judgment, attention, and instability. Approaches that assume stable distortions of Bayes' rule do not capture the neglect (in both judgment and measured attention) of relevant features, the use of irrelevant features, and the instability in attention and bias when the problem's statistical contrast or framing change.[15]

With respect to contrast, Corollary 5 predicts that making one relevant piece of information more extreme (the likelihood) interferes with attention to, 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. We next test Proposition 4 and Corollary 5.

5.4. *Inference experiments*

Table 3 summarizes our inference treatments. T_B and T_C are our baseline balls-and-urns and cabs treatment of Section 2. To test the new predictions, we add four 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 stays at 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 the problem's description. In treatment T_H, we modify T_U by labelling the urns by their modal colour, "Green-urn" versus "Blue-urn," and by describing the likelihood

15. In Enke and Graeber (2023), people perceive likelihoods imprecisely, which causes: (1) a dispersion of estimates, and (2) a shrinkage of posteriors toward the prior which gives an average underreaction 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 underreaction in that model.

TABLE 3
Treatments manipulating salience in inference problems

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

(80%) as the probability a drawn ball "matches" the colour of the urn.[16] The rewording thus intuitively increases the prominence of the "match" and the "colour of ball" features, which we also verify experimentally.

In treatment T_U, we conversely change T_C to make the signal less prominent. We modify: (1) the description of the witness to "the court found that the witness was very unreliable: he was able to identify each colour correctly only about 80% of the time…", and (2) that 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 and prominence of the report, which affect attention and biases even though the statistical informativeness of the signal is unchanged.

To measure attention, we ask participants to justify their probability estimates in a free-response elicitation and then 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 versus B, (2) whether the drawn ball was green or blue, (3) whether the drawn ball matched many balls in the jar it came from, and (4) none of the above. For taxicabs, analogous options appeared about the cab companies and the witness report.[17] Free-response attention is measured by querying OpenAI's GPT 3.5 model with separate yes–no questions about whether the free form response appears to indicate attention to each of the features (1)–(4).

Across the six treatments, we again test two predictions. First, by Proposition 4, reported attention to urns, colour, and match should align with which mode the DM anchors to. Second, by Corollary 5, an increase in the contrast of the likelihood (T_{LE} versus T_{ME}) or the prominence of match (T_H versus T_B) should boost both attention to the signal and anchoring to the likelihood. Conversely, lowering the prominence of the signal (T_U versus T_C) should shift attention

16. 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 colour 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 colour of the jar it came from about 80% of the time. About 20% of the time, it was the opposite colour."

17. 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).

TABLE 4
Multimodality in attention and in estimates

	(1) Base rate	(2) Likelihood	(3) Bayes	(4) 50%
Directly elicited attention				
Only urn	0.418***			
	(0.022)			
Only colour/match		0.408***		
		(0.023)		
Only urn and colour			0.128***	
			(0.026)	
Nothing				0.166***
				(0.041)
Free-response attention				
Only urn	0.169***			
	(0.022)			
Only colour/match		0.121***		
		(0.027)		
Only urn and colour			0.110***	
			(0.026)	
Nothing				0.054***
				(0.011)
Treatment FEs	Yes	Yes	Yes	Yes
N	2,061	2,061	2,061	2,061
R^2	0.296	0.256	0.069	0.052

The dependent variable is whether participants' answers were the base rate (column 1), the likelihood (column 2), within 5% 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.

away from the signal and increase anchoring to base rates. Finally, we test in the experiment of Section 2 whether moving from balls and urns (T_B) to taxicabs (T_C) leads to greater attention to match and colour.

5.5. *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 ordinary least squares 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.

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% points more likely to anchor to the base rate. Free-response attention to urn further increases that probability by 16.9% 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, when selecting attention from the list, participants may mechanically ex post select 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, this is much less of

a concern for the free-response measure, which is based on how respondents themselves describe how they thought about the problem, and that in Table 4 exhibits additional explanatory power beyond directly elicited attention.

5.6. *Attention and instability in estimates*

We next show the effect of controlled manipulations of contrast and prominence. We first look at participants' 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} versus 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% points of) the Bayesian answer (from 12.1% to 9.2%, $p = 0.15$).[18] Consequently, in Panel B the relative share of answers at the likelihood or Bayes versus the base rate increases, consistent with Corollary 5.

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 evidence, which instead points to a failure of integration, as in our mechanism. 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 continues to neglect the signal.[19]

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 describing the problem in terms of the match feature, T_H greatly increases the share of participants who anchor to the likelihood compared with standard balls and urns T_B, both in absolute terms (22.8% versus 15.5%, $p < 0.01$) and relative to the base rate (2.2 versus 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.[20]

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 colour, match, or both, relative to

18. Changing the likelihood also changes the correct answer. In the Supplementary Appendix, we describe a sharper test in which the contrast of the ball's colour 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 versus 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.

19. Augenblick *et al.* (2025) 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 with the Bayesian benchmark in which it should be three times higher.

20. Tversky (1977) discusses a variation of the taxicab problem in which the base rate refers to the frequency of accidents instead of the frequency of cabs (the problem states "although the two companies are roughly equal in size, 85% of cab accidents in the city involve Green cabs, and 15% involve Blue cabs.") This description makes the base rate of the hypotheses "green" and "blue" more explicit, thereby increasing its prominence. Consistent with this interpretation, there is substantially less overreaction in this version of the problem.

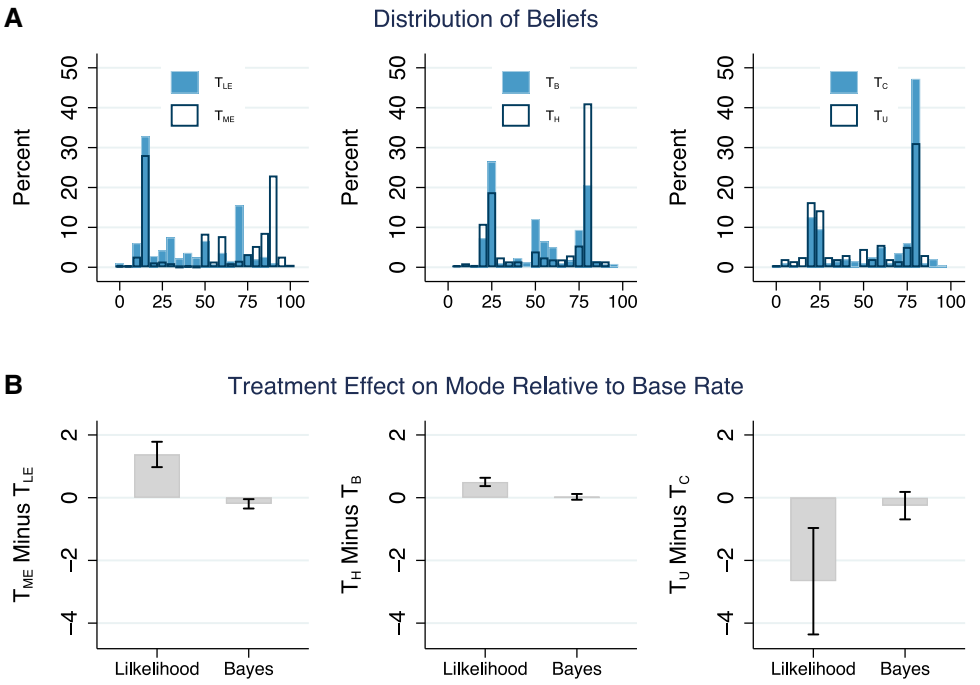

FIGURE 6

Panel 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

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 colour 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 with 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 salience-driven attention for understanding bias: there is 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 hypothesis. To understand biases, one needs to go beyond objective probabilities, and instead measure attention and feature salience.[21]

21. A distinction has also been drawn between inference problems and "forecasting", in which overreaction also prevails (Fan *et al.*, 2024). 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 with inference, fostering overreaction.

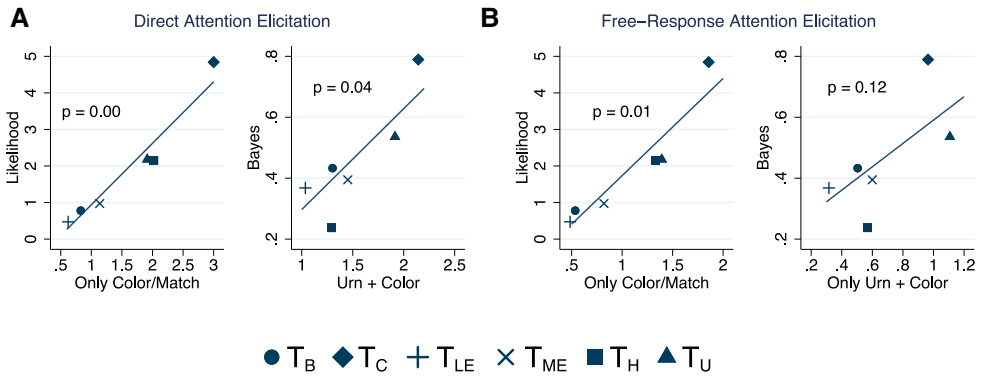

FIGURE 7

Treatment effects on beliefs and attention. The x-axis is the fraction of participants in each treatment attending to colour and/or match (left figure within each panel) and to urn + colour (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

5.7. *Model estimation*

We provide a structural test of our model by estimating it via maximum likelihood (details are in Supplementary Appendix C). This allows us to estimate latent prominence and the weight of contrast from observed probability estimates, and then 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

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

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.

Measured and model-implied attention are positively correlated. When beliefs move "as if" there was an increase in the salience of the signal, match, or the Bayes profile, measured attention on these profiles also increases. Second, contrast matters: its coefficient 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 colour 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 + colour) *decreases* from 22.1%

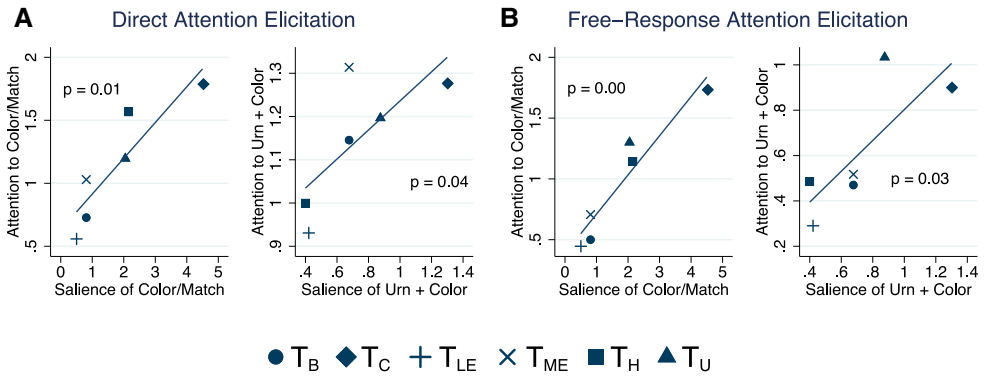

FIGURE 8

Measured versus revealed attention to features. The x-axis is the estimated salience of each attention profile (where we sum together the colour 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)

to 4.0% ($p < 0.01$). Consistent with interference, the salience of match also reduces attention to "only colour" (12.3% versus 6.9%, $p = 0.02$).

6. ADDITIONAL IMPLICATIONS OF SALIENCE-DRIVEN ATTENTION

We 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 and 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 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. Attention accounts for this phenomenon: one hypothesis is salient in the DM's mind, and so is more easily simulated than its alternative. One way to test for this mechanism is to change the prominence of a hypothesis in the description. In balls and urns, hypotheses are described as "what is the probability that the ball is drawn from A versus B?" The same question could be phrased as: "what is the probability that ball is drawn from A?" By leaving urn B implicit, the second phrasing 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 $\alpha = (\alpha_1, .., \alpha_O, \alpha_B)$.[22] When $\alpha_B = 1$ both hypotheses are attended to, which is the

22. 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.

case studied so far. When $\alpha_B = 0$, the DM fails to simulate H_B and solves the problem as:

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

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. 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",* $\Pr(H_A;\ \alpha_{A\cap g}) = \pi_A \cdot q$.

Neglect of H_B reduces the share of correct answers because 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 colour composition of B, which is not the case for the mechanism in Proposition 4.[23]

Second, and crucially, the "focal H_A" format creates an entirely "new mode", $\alpha_{A\cap g}$ anchored at $\pi_A \cdot q$. At this mode, which sharply identifies neglect of H_B, the DM attends to both statistical features (the selection of A and the drawing of a green ball from it), and replaces the original question with "what is the probability that a ball is green *and* from A"? These DMs simulate A by computing the joint probability $\pi_A \cdot q$ as in Equation (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 favourable 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 is true in our data.

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).

Keeping the alternative implicit is only 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, 1995) showed that more accurate inference is 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

23. 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.

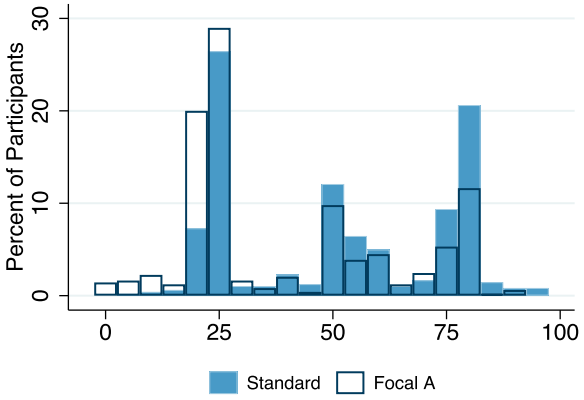

FIGURE 9
The figure shows the distribution of beliefs about $\Pr(A \mid g)$

this idea. In this format, in fact, there is a single statistical feature: "drawing a ball from U and of colour 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 argues 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.[24] 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 instead it is also necessary to draw 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).

The results are strongly in line with our model. In Panel A, compared with 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 20% becomes dominant. The benefit of the frequency format over the standard one is no longer clear: it leads many people to estimate A below its base rate despite the favourable signal.[25]

24. 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 B. Out of those, 15 are blue and 60 are green. There are many studies of the effect of training and communication of statistics (Visschers *et al.*, 2009; Gigerenzer, 2014; Operskalski and Barbey 2016).

25. 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 "colour" and "urn" features. Esponda *et al.* (2024)

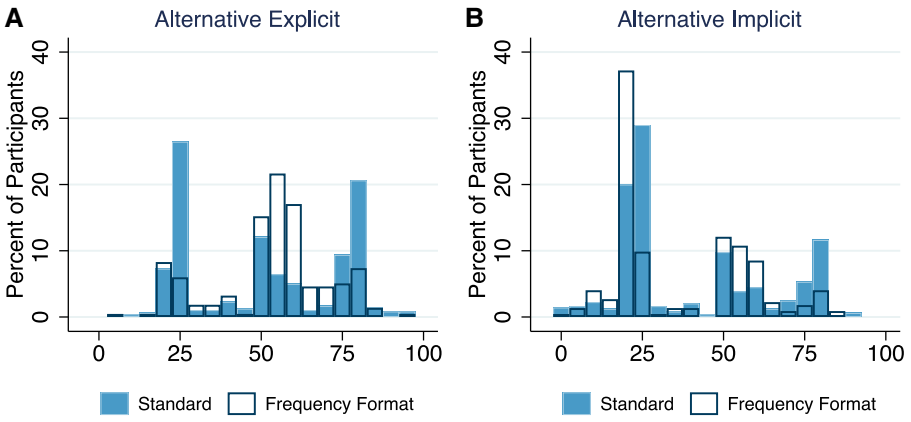

FIGURE 10
Balls and urns in baseline and frequency formats. Each panel shows the distribution of $\Pr(A \mid g)$

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. Psychology work on problem solving is consistent with this view: sometimes naturalistic settings and prior knowledge help, as in solving the Wason task (Wason, 1968); other times they impair problem solving because people fail to see unusual but useful properties of an object, as in the famous candle problem (Galinsky and Moskowitz, 2000). Systematically engaging with salience-driven attention, shaped by contrast and prominence, may help design architectures conducive to improved judgments.[26]

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 i.i.d. processes, Kahneman and Tversky (1972) and Benjamin *et al.* (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.

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.

26. We only considered prominence as a source of hypothesis neglect, but contrast may also play a role. Ba *et al.* (2024) show that overreaction to data increases when a neutral urn C with a 50–50 colour 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.

Consider this phenomenon through the lens of our model. The DM's estimate is shaped by the number $r \leq \min(K, n)$ of flips he attends to, captured by attention profile α_r. The latter pins down the representation $R_{\alpha_r}(H_i)$, which is the union of attended subsequences of length r of the hypothesis' atoms, $\omega \in H_i$.[27] 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_α. As before, ϵ_α is common to all profiles α in which flips are attended to, so it does not matter here. As we show in Supplementary 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 probability density function $\pi(K)$ in support $[1, \bar{K}]$. 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 versus H_2, more so when smaller values of K are more likely. As n increases, average beliefs converge to $\bar{\pi}(\bar{K})$.*

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 i.i.d. draws. Existing models have wrestled with reconciling the faulty reliance on the law of large numbers in the GF with an insufficient reliance on it in large samples (Benjamin *et al.*, 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 GF. 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 the 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 <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 GF, 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).

Table 5 presents the results. Consistent with our prediction, the coefficient on the size of the equivalence class is positive but >1, showing insensitivity, and is smaller for longer sequences $n = 4, 6$ compared with $n = 2$. Thus, salience-driven attention generates three observed behaviours: (1) the share of subjects committing the GF increases in sequence length n (contrast); furthermore, conditional on committing the fallacy (2) its severity increases with the size of a sequence's equivalence class based on sh (question substitution) but (3) less than proportionally to the latter's size (insensitivity). Property (3) follows from our model but to our knowledge has not been documented before.

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

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

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)

	(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	1,128	8,528	8,016	17,672
Individuals	282	533	501	1,316
R^2	0.20	0.10	0.06	0.37

Robust standard errors in parentheses. Data are restricted to participants for whom judged probabilities and balancedness of heads and tails are positively correlated.
** and *** indicate significance at the 5% and 1% levels, respectively.

the number of signals. 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 are favourable 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 favour 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 favourable 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 with 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.

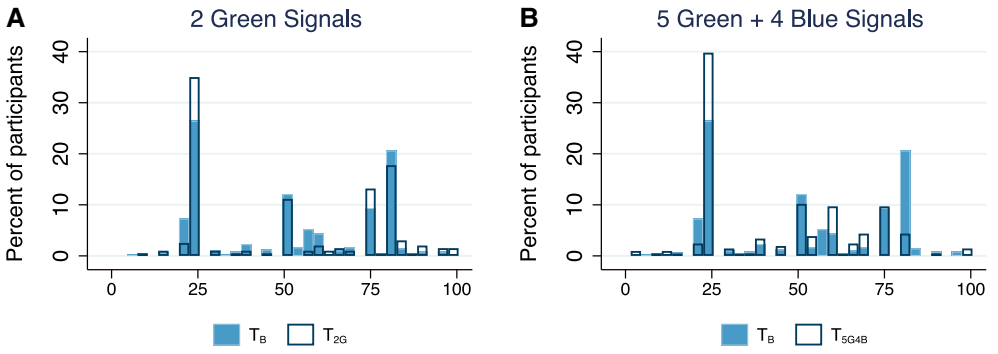

FIGURE 11

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)

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. Figure 11A shows the distribution of beliefs in these two treatments. Consistent with the insensitivity in (1), the average response is 52.6% (only 1.4 p.p. higher than in T_B, $p = 0.50$), which exhibits more average underreaction 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 (2) by eliciting beliefs after five green and four blue signals, under the same base rate $\pi_A = 0.25$ and the likelihood $q = 0.8$ as T_B. Figure 11B compares the resulting distribution of beliefs between T_B and T_{5G4B}. Consistent with prediction (2), 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 favour of the base rate.[28]

7. CONCLUSION

Understanding belief formation is critical to understanding economic behaviour. Over the past sixty years, psychologists and behavioural scientists have unveiled many systematic departures of beliefs from the standard Bayesian model (Benjamin *et al.*, 2019), including the GF, underreaction and overreaction in inference, and others. This evidence has led to a proliferation of bias-specific models, reflecting wide ranging and sometimes contradictory findings. This research has produced important insights but has also opened many doors, leaving a sense that anything goes.

Statistical problems are a useful laboratory to study belief formation and unveil unifying mechanisms, because in these problems: (1) there is a correct answer and (2) the given statistics offer anchors for detecting shifting attention between them. On this testing ground, our analysis shows that bias-specific models cannot account for two key empirical regularities: multimodality

28. We did not elicit attention to specific numbers and colours 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.

and instability. These phenomena point to a basic cognitive mechanism: salience-driven attention to the features of events. 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 these features can lead to different distorted representations of the hypothesis, which are 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.

This approach differs from a leading social science model viewing attention as a scarce resource that is *optimally* allocated to further the DM's goals (*e.g.* "rational inattention" in economics or "efficient coding" in psychology). While the scarcity of attention is uncontroversial, our analysis challenges its 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. Salience plays a key role to explaining such anomalies, in line with substantial research in psychology showing its role in attention allocation. Developing a deeper understanding of how goals shape salience and attention allocation is an important avenue for future work.

To that effect, an important direction is to integrate the roles of attention and selective memory. In the statistical problems we considered, all relevant data are put in front of subjects. Yet recalled past experiences arguably influence what features they attend to, representations, and estimates. One way in which the relevance of a witness statement in court draws attention is that it reminds the DM's of similar past experiences of people giving testimony. Briefly mentioning that a witness is unreliable cues the opposite reaction—we are indeed used to ignoring 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 could shed light on a range of issues, including why different people represent the same problem in different ways and make different choices, and why features that capture similarity to the data-generating process can be prominent (Bordalo *et al*., 2025b). In the field, such a theory of prominence would 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 *et al*., 2024), but also why learning can be sped up once neglected relevant features are made prominent (Hanna *et al*., 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 *et al*. (2023, 2025a) 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 (Conlon and Patel, 2025; Jiang *et al*., 2024; Bordalo *et al*., 2025a). Attention-driven representations can add a crucial ingredient to this theory: which cue in the environment is noticed and triggers retrieval. For example, the salient losses or failure of an individual bank may draw investors' attention, causing them to retrieve past episodes of financial meltdown, and to neglect the rarity of cataclysmic events.

The combination of memory and salience-driven attention is also relevant for consumer choice. Bordalo *et al.* (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 and prominence but also surprise, can expand our understanding of heterogeneity and instability of observed choices. Based on past experiences, 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 attention to features raises the promise of a general theory of intuitive judgments in both naturalistic and abstract settings.

Supplementary Data

Supplementary data are available at *Review of Economic Studies* online.

Data Availability

The data and code underlying this research is available on Zenodo at https://dx.doi.org/10.5281/zenodo.14036127.

Acknowledgements. 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, Florian Zimmermann, and four anonymous referees for helpful comments. N.G. thanks the European Research Council (grant no. 101097578) for financial support.

REFERENCES

AUGENBLICK, N., LAZARUS, E. and THALER, M. (2025), "Overinference from Weak Signals and Underinference from Strong Signals", *The Quarterly Journal of Economics*, **140**, 335–401.

BA, C., BOHREN, A. and IMAS, A. (2024), "Over- and Underreaction to Information" (Working Paper No. 24-030, PIER).

BENJAMIN, D. (2019), "Errors in Probabilistic Reasoning and Judgment Biases", in *Handbook of Behavioral Economics: Applications and Foundations 1* (North-Holland: Elsevier) 69–186.

BENJAMIN, D., BODOH-CREED, A. and RABIN, M. (2019), "Base Rate Neglect: Foundations and Implications" (Mimeo).

BENJAMIN, D., MOORE, D. and RABIN, M. (2017), "Biased Beliefs about Random Samples: Evidence from Two Integrated Experiments" (Working Paper No. 23927, NBER).

BENJAMIN, D. J., RABIN, M. and RAYMOND, C. (2016), "A Model of Nonbelief in the Law of Large Numbers", *Journal of the European Economic Association*, **14**, 515–544.

BORDALO, P., BURRO, G., COFFMAN, K., *et al.* (2025a), "Imagining the Future: Memory, Simulation, and Beliefs", *The Review of Economic Studies*, **92**, 1532–1563.

BORDALO, P., CONLON, J., GENNAIOLI, N., *et al.* (2023), "Memory and Probability", *The Quarterly Journal of Economics*, **138**, 265–311.

BORDALO, P., GENNAIOLI, N., LANZANI, G., *et al.* (2025b), "A Cognitive Theory of Reasoning and Choice" (Working Paper No. 33466, NBER).

BORDALO, P., GENNAIOLI, N. and SHLEIFER, A. (2012), "Salience Theory of Choice under Risk", *The Quarterly Journal of Economics*, **127**, 1243–1285.

—— —— —— (2013), "Salience and Consumer Choice", *The Journal of Political Economy*, **121**, 803–843.

—— —— —— (2022), "Salience", *Annual Review of Economics*, **14**, 521–544.

CAMERER, C. (1987), "Do Biases in Probability Judgment Matter in Markets? Experimental Evidence", *American Economic Review*, **77**, 981–997.

—— (1990), "Do Markets Correct Biases in Probability Judgment? Evidence from Market Experiments", in KAGEL, J. H. and GREEN, L. (eds) *Advances in Behavioral Economics* (Vol. **2**) (Norwood, NJ: Ablex Publishing Company) 125–172.

CHETTY, R., LOONEY, A. and KROFT, K. (2009), "Salience and Taxation: Theory and Evidence", *American Economic Review*, **99**, 1145–1177.

CLANCY, K., BARTOLOMEO, J., RICHARDSON, D., *et al.* (1981), "Sentence Decision Making: The Logic of Sentence Decisions and the Sources of Sentence Disparity", *Journal of Criminal Law and Criminology*, **72**, 524–554.

CONLON, J. (2025), "Attention, Information, and Persuasion" Working Paper.

CONLON, J. and PATEL, D. (2025), "What Jobs Come to Mind? Stereotypes about Fields of Study" Working Paper.

DE BRUIN, W. B., FISCHHOFF, B., MILLSTEIN, S., *et al.* (2000), "Verbal and Numerical Expressions of Probability: "It's A Fifty–Fifty Chance", *Organizational Behavior and Human Decision Processes*, **81**, 115–131.

DESIMONE, R. and DUNCAN, J. (1995), "Neural Mechanisms of Selective Visual Attention", *Annual Review of Neuroscience*, **18**, 193–222.

DOHMEN, T., FALK, A., HUFFMAN, D., *et al.* (2009), "The Non-Use of Bayes Rule: Representative Evidence on Bounded Rationality" (Working Paper No. 038, METEOR).

EDWARDS, W. (1968), "Conservatism in Human Information Processing", in KAHNEMAN, D., SLOVIC, P. and TVERSKY, A. (eds) *Judgment under Uncertainty: Heuristics and Biases* (Cambridge: Cambridge University Press) 359–369.
ENKE, B. (2020), "What You See is All There is", *The Quarterly Journal of Economics*, **135**, 1363–1398.
ENKE, B. and GRAEBER, T. (2023), "Cognitive Uncertainty", *Quarterly Journal Economics*, **138**, 2021–2067.
ENKE, B. and ZIMMERMANN, F. (2019), "Correlation Neglect in Belief Formation", *The Review of Economic Studies*, **86**, 313–332.
ESPONDA, I., VESPA, E. and YUKSEL, S. (2024), "Mental Models and Learning: The Case of Base-Rate Neglect", *American Economic Review*, **114**, 758–782.
EVERS, E. R. K., IMAS, A. and KANG, C. (2022), "On the Role of Similarity in Mental Accounting and Hedonic Editing", *Psychological Review*, **129**, 777–789.
FAN, T., LIANG, Y. and PENG, C. (2024), "The Inference-Forecast Gap in Belief Updating." SSRN 3889069.
GABAIX, X. (2014), "A Sparsity-based Model of Bounded Rationality", *The Quarterly Journal of Economics*, **129**, 1661–1710.
—— (2019), "Behavioral Inattention", in *Handbook of Behavioral Economics: Applications and Foundations 1* (Vol. **2**) (North-Holland: Elsevier) 261–343.
GAGNON-BARTSCH, T., RABIN, M. and SCHWARTZSTEIN, J. (2023), *Channeled Attention and Stable Errors* (Boston: Harvard Business School).
GALINSKY, A. and MOSKOWITZ, G. (2000), "Counterfactuals as Behavioral Primes: Priming the Simulation Heuristic and Consideration of Alternatives", *Journal of Experimental Social Psychology*, **36**, 384–409.
GIGERENZER, G. (1996), "On Narrow Norms and Vague Heuristics: A Reply to Kahneman and Tversky", *Psychological Review*, **3**, 592–596.
—— (2014), *Risk Savvy: How to Make Good Decisions* (New York: Viking).
GIGERENZER, G. and HOFFRAGE, U. (1995), "How to Improve Bayesian Reasoning Without Instruction: Frequency Formats", *Psychological Review*, **102**, 684–704.
GRAEBER, T. (2023), "Inattentive Inference", *Journal of the European Economic Association*, **21**, 560–592.
GRETHER, D. (1980), "Bayes Rule as a Descriptive Model: The Representativeness Heuristic", *The Quarterly Journal of Economics*, **95**, 537–557.
GRIFFIN, D. and TVERSKY, A. (1992), "The Weighing of Evidence and the Determinants of Confidence", *Cognitive Psychology*, **24**, 411–435.
GUYON, I. and ELISSEEFF, A. (2003), "An Introduction to Variable and Feature Selection", *Journal of Machine Learning Research*, **3**, 1157–1182.
HANNA, R., MULLAINATHAN, S. and SCHWARTZSTEIN, J. (2014), "Learning through Noticing: Theory and Evidence from a Field Experiment", *The Quarterly Journal of Economics*, **129**, 1311–1353.
JIANG, Z., LIU, H., PENG, C., *et al.* (2024), "Investor Memory and Biased Beliefs: Evidence from the Field" (Working Paper No. 33226, NBER).
KAHNEMAN, D. and FREDERICK, S. (2002), "Representativeness Revisited: Attribute Substitution in Intuitive Judgment", in GILOVICH, T., GRIFFIN, D., KAHNEMAN, D. (eds) *Heuristics and Biases: The Psychology of Intuitive Judgment* (New York: Cambridge University Press) 49–81.
KAHNEMAN, D. and TVERSKY, A. (1972), "Subjective Probability: A Judgment of Representativeness", *Cognitive Psychology*, **3**, 430–454.
KHAW, M. W., LI, Z. and WOODFORD, M. (2021), "Cognitive Imprecision and Small-Stakes Risk Aversion", *The Review of Economic Studies*, **88**, 1979–2013.
KLEINBERG, J., LAKKARAJU, H., LESKOVEC, J., *et al.* (2018), "Human Decisions and Machine Predictions", *The Quarterly Journal of Economics*, **133**, 237–293.
KLEINBERG, J., LIANG, A. and MULLAINATHAN, S. (2017), "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, June 26–30. Cambridge, MA, 125–126.
KRUSCHKE, J. (2008), "Models of Categorization", in *The Cambridge Handbook of Computational Psychology* (Cambridge University Press) 267–301.
LEWIS, M. (2003), *Moneyball* (New York: W.W. Norton).
LI, X. and CAMERER, C. (2022), "Predictable Effects of Visual Salience in Experimental Decisions and Games", *The Quarterly Journal of Economics*, **137**, 1849–1900.
LUDWIG, J. and MULLAINATHAN, S. (2024), "Machine Learning as a Tool for Hypothesis Generation", *The Quarterly Journal of Economics*, **139**, 751–827.
MCCLELLAND, J. and RUMELHART, D. (1981), "An Interactive Activation Model of Context Effects in Letter Perception: I. An Account of Basic Findings", *Psychological Review*, **88**, 375–407.
NICKERSON, R. (1998), "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises", *Review of General Psychology*, **2**, 175–220.
NOSOFSKY, R. (1988), "Similarity, Frequency, and Category Representations", *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **14**, 54–65.
NOSOFSKY, R., SANDERS, C., MEAGHER, B., *et al.* (2018), "Toward the Development of a Feature-Space Representation for a Complex Natural Category Domain", *Behavior Research Methods*, **50**, 530–556.
OPERSKALSKI, J. and BARBEY, A. (2016), "Risk Literacy in Medical Decision-Making", *Science*, **352**, 413–414.

RABIN, M. (2002), "Inference by Believers in the Law of Small Numbers", *The Quarterly Journal of Economics*, **117**, 775–816.
RABIN, M. and VAYANOS, D. (2010), "The Gambler's and Hot-Hand Fallacies: Theory and Applications", *The Review of Economic Studies*, **77**, 730–778.
REMINGTON, R. W., JOHNSTON, J. C. and YANTIS, S. (1992), "Involuntary Attentional Capture by Abrupt Onsets", *Perception & Psychophysics*, **51**, 279–290.
REUTSKAJA, E., NAGEL, R., CAMERER, C. F., *et al.* (2011), "Search Dynamics in Consumer Choice under Time Pressure: An Eye-Tracking Study", *American Economic Review*, **101**, 900–926.
SCHWARTZSTEIN, J. (2014), "Selective Attention and Learning", *Journal of the European Economic Association*, **12**, 1423–1452.
SELFRIDGE, O. (1955), "Pattern Recognition and Modern Computers", in Proceedings of the March 1–3, 1955, Western Joint Computer Conference, New York. Association for Computer Machinery (ACM), 91–93.
SIMON, H. (1957), *Models of Man* (New York: Wiley).
SIMS, C. (2003), "Implications of Rational Inattention", *Journal of Monetary Economics*, **50**, 665–690.
TORGERSON, W. (1952), "Multidimensional Scaling: I. Theory and Method", *Psychometrika*, **17**, 401–419.
TVERSKY, A. (1977), "Features of Similarity", *Psychological Review*, **84**, 327–352.
TVERSKY, A. and GATI, I. (1982), "Similarity, Separability, and the Triangle Inequality", *Psychological Review*, **89**, 123–154.
TVERSKY, A. and KAHNEMAN, D. (1983), "Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment", *Psychological Review*, **90**, 293–315.
VISSCHERS, V., MEERTENS, R., PASSCHIER, W., *et al.* (2009), "Probability Information in Risk Communication: A Review of the Research Literature", *Risk Analysis: An International Journal*, **29**, 267–287.
WASON, P. (1968), "Reasoning About a Rule", *The Quarterly Journal of Experimental Psychology*, **20**, 273–281.
WOODFORD, M. (2003), "Imperfect Common Knowledge and the Effects of Monetary Policy", in AGHION, P., FRYDMAN, R., STIGLITZ, J., WOODFORD, M. (eds) *Knowledge, Information and Expectations in Modern Macroeconomics* (Princeton: Princeton University Press) 25–58.
—— (2020), "Modeling Imprecision in Perception, Valuation, and Choice", *Annual Review of Economics*, **12**, 579–601.