MEMORY AND PROBABILITY*

PEDRO BORDALO
JOHN J. CONLON
NICOLA GENNAIOLI
SPENCER Y. KWON
ANDREI SHLEIFER

In many economic decisions, people estimate probabilities, such as the likelihood that a risk materializes or that a job applicant will be a productive employee, by retrieving experiences from memory. We model this process based on two established regularities of selective recall: similarity and interference. We show that the similarity structure of a hypothesis and the way it is described (not just its objective probability) shape the recall of experiences and thus probability assessments. The model accounts for and reconciles a variety of empirical findings, such as overestimation of unlikely events when these are cued versus neglect of noncued ones, the availability heuristic, the representativeness heuristic, conjunction and disjunction fallacies, and over- versus underreaction to information in different situations. The model yields several new predictions, for which we find strong experimental support. *JEL Codes:* D83, D91, C91.

I. INTRODUCTION

It is well known that memory plays an important role in belief formation. Tversky and Kahneman (1973) show that when instances of a probabilistic hypothesis are easier to recall, the hypothesis is judged to be more likely, a finding they call the availability heuristic. When prompted to think about an unlikely event, such as dying in a tornado, people overestimate its frequency (Lichtenstein et al. 1978). They also attach a higher probability to an event if its description is broken down into constituent parts, which facilitates retrieval of instances (Fischoff, Slovic, and Lichtenstein 1978). More broadly, beliefs depend on recalled personal experiences, such as stock market crashes

*We are grateful to Ben Enke, Drew Fudenberg, Sam Gershman, Thomas Graeber, Cary Frydman, Lawrence Jin, Yueran Ma, Fabio Maccheroni, Sendhil Mullainathan, Salvo Nunnari, Dev Patel, Kunal Sangani, Jesse Shapiro, Josh Schwartzstein, Adi Sunderam, and Michael Woodford for helpful comments. Julien Manili provided outstanding research assistance. Gennaioli thanks the Italian Ministry of Education, University and Research for financial support (PRIN 2017 Prot. 2017CY3SSY).

The Quarterly Journal of Economics (2023), 265–311. https://doi.org/10.1093/qje/qjac031.
Advance Access publication on August 29, 2022.

(Malmendier and Nagel 2011), and not just statistical information. Despite this evidence, a systematic analysis of the role of human memory in belief formation is lacking.

It is also well known that beliefs depart from rationality in a variety of ways that shape behavior. Sometimes unlikely events are overestimated, as when consumers overpay for insurance (Sydnor 2010; Barseghyan et al. 2013) or bet in long-shot lotteries (Chiappori et al. 2019). Other times, unlikely events are underestimated, as when investors neglect tail risk (Gennaioli, Shleifer, and Vishny 2012). Adding to the clutter, in finance there is abundant evidence of both over- and underreaction to news. Beliefs overestimate the future earnings of individual firms and of the market after periods of rapid earnings growth (Bordalo et al. 2019b, 2022c), leading to long-run return reversals, but underestimate the impact of other news, such as earnings surprises, leading to return momentum (Chan, Jegadeesh, and Lakonishok 1996; Bouchaud et al. 2019; Kwon and Tang 2021). This bewildering diversity of biases is puzzling and has led skeptics to minimize the evidence on beliefs and stick to rationality.

In this article, we show that building a theory of belief formation based on the psychology of human memory helps reconcile seemingly contradictory biases and generates new predictions. In our theory, a decision maker (DM) evaluates a probabilistic hypothesis by sampling instances of its occurrence from memory. In line with memory research (Kahana 2012), retrieval is shaped by frequency, similarity, and interference. The model unifies apparently contradictory phenomena based on the similarity between a hypothesis and experiences in the database.

When assessing an unlikely hypothesis, the DM overestimates it because he oversamples similar events from memory. However, when assessing a broad hypothesis, the DM does not think about its unlikely components, because the latter are dissimilar to the hypothesis itself. This mechanism explains why people systematically fail to imagine spending shocks that may hit them in the future and hence undersave (Augenblick et al. 2022), but at the same time overpay for insurance against a specific shock when it is described to them (Kunreuther and Pauly 2010).

This mechanism generates overreaction if the data point to a hypothesis that is sufficiently similar to these data compared with the alternative hypothesis, which makes instances of the former hypothesis easy to retrieve. When both hypotheses are similar

to the data, underreaction arises. This principle microfounds and generalizes the diagnostic expectations model of belief formation (Bordalo, Gennaioli, and Shleifer 2018) and unifies conflicting evidence of over- and underreaction to news in financial markets. It also explains why social stereotypes often exhibit a kernel of truth (Bordalo et al. 2016), but it also implies that they tend to be especially inaccurate for minorities.

To see how the model works, consider a DM assessing the probability of a hypothesis H_1 relative to a disjoint alternative H_2. The DM estimates the frequency of H_1 by the ease with which instances of it are retrieved compared with instances of H_2. The DM does not use statistical data such as base rates. For concreteness, suppose that H_1 = "cause of death is flood" compared to H_2 = "other causes of death." Similarity-based recall produces two effects. First, it implies that when thinking about H_1, instances of floods are easier to retrieve than those of earthquakes, because the former are more similar to H_1 than the latter. Second, it implies that the DM may also retrieve the similar yet irrelevant "accidental drowning," which belongs to the alternative H_2 = "other causes of death," or even "survival in a flood," which is a nonlethal experience. This is the phenomenon of interference, which produces hypothesis-inconsistent thoughts, potentially leading to underestimation. The same forces are at play when the DM thinks about the alternative H_2.

These memory effects may cause overestimation versus underestimation of unlikely events, depending on similarity. An event H_1 = "cause of death is flood" is overestimated because memory oversamples similar events relative to their rare occurrence and because the alternative hypotheses H_2 = "causes of death other than flood" consists of many dissimilar events that are difficult to retrieve, so they face strong interference. But the latter also implies that rare events in H_2 such as "botulism"—which are not explicitly assessed—are underestimated. They are too dissimilar from the average cause of death and so do not come to mind. Unlike Kahneman and Tversky's (1979) prospect theory, our model makes predictions for when low-probability events are over- or underestimated, helping explain conflicting risk attitudes documented in the field.

Strikingly, our model also produces biases in conditional probability assessments that are typically attributed to representativeness (Kahneman and Tversky 1973). Consider the well-known base-rate neglect. When given the data that Steve is shy and

withdrawn, subjects think he is more likely to be a librarian than a farmer, neglecting the fact that farmers are far more numerous than librarians. Similarity explains the mistake: the data "shy and withdrawn" is similar to a librarian (many librarians have this personality), but not to a farmer (many farmers are outgoing). The "farmer" hypothesis faces stronger interference, causing overreaction to data. This mechanism accounts for the conjunction fallacy (Tversky and Kahneman 1974, 1983). Similarity also creates limits to representativeness: when the data are fairly similar to both hypotheses, underreaction prevails. In financial markets, the data "positive earnings surprise," while statistically predictive of good future outcomes, occur frequently enough before average or bad future outcomes that they are fairly similar to those. The model predicts that, in this case, underreaction should prevail.

We test the new predictions of our model using a novel experimental design in which participants see 40 images that differ in content and in some cases also in color. Subjects then assess the probability that a randomly selected image has a certain property. To do so, they only need to recall what they saw. We manipulate the subjects' database of experiences and the cues they face when assessing a hypothesis. We also measure the recall of experiences. We find support for our predictions for how over- and underestimation of unlikely events can be switched on and off by modulating similarity and interference. We also generate over- and underreaction to data by varying the strength of the signal and the likelihood of the hypothesis. Across all treatments, recall of experiences and probability judgments are strongly correlated.

Recent research explores the role of memory in belief formation (Mullainathan 2002; Wachter and Kahana 2019; Bordalo et al. 2020a; Enke, Schwerter, and Zimmermann 2020). Some see this phenomenon as efficient information processing (Tenenbaum and Griffiths 2001; Azeredo da Silveira, Sung, and Woodford 2020; Dasgupta et al. 2020; Dasgupta and Gershman 2021). We instead start with well-documented regularities in recall and show how they unify the representativeness and availability heuristics (Tversky and Kahneman 1974). Due to similarity and interference, representative experiences are more "available," or accessible, for recall.

Bordalo et al. (2020a) present and experimentally test a model of memory-based beliefs in which the representativeness heuristic follows from a context-dependent similarity function (Tversky 1977), meaning that the similarity of an experience to a hypothesis

is higher if that experience is less likely in alternative hypotheses. We use a standard similarity function and obtain this effect as a special case of the broader role of interference in recall. This approach yields many new results. It yields biases attributed to the availability heuristic and new predictions on the underestimation of heterogeneous hypotheses, the over- versus underestimation of unlikely events, and the coexistence of under- and overreaction to data. We experimentally test these novel implications by extending the design in Bordalo et al. (2020a).[1]

We describe our model of similarity-based recall and probability judgments in Section II. Section III characterizes the departures of probability estimates from statistically correct beliefs. Section IV presents experimental results, and Section V covers economic applications. Section VI concludes.

II. The Model

A decision maker's (DM) memory database E consists of $N > 1$ experiences, accumulated either through personal events or via communication or media reports. An experience e is described by $F > 1$ features, each of which takes a value in $\{0,1\}$.

In our running example, we consider a database of potential causes of death. Here a subset of features captures different potential causes: f_1 may identify "car accident," f_2 "flood," f_3 "heart attack," and so on. One feature, which we denote by f_d, indicates whether the event was lethal. There are superordinate features, such as f_{d+1} = "disease," f_{d+2} = "natural disaster," and so on, which take the value of 1 for the relevant subsets of possible death events. Experiences are vectors of features. For instance, lethal heart attacks have $f_1 = f_2 = 0$, $f_3 = f_d = f_{d+1} = 1$, and f_{d+2}

1. In psychology, Sanborn and Chater (2016) present a model of beliefs based on Bayesian memory sampling. The MINERVA-DM model (Dougherty, Gettys, and Ogden 1999) features similarity-based recall and noisy encoding but does not allow for interference. These models cannot account for representativeness or the conjunction fallacy without ad hoc ancillary assumptions. In Billot et al. (2005), the probability of an elementary event is estimated based on its similarity to other events in the database, but they do not study judgment biases, and their model generates neither the conjunction nor the disjunction effect. In Johnson, Häubl, and Keinan (2007), buyers and sellers sample different aspects of a good from memory depending on how they are cued with different queries. While they focus on explaining the endowment effect rather than probability biases, they also emphasize interference and similarity to the cue in shaping retrieval.

$= 0$. Nonlethal heart attacks have the same feature values except for $f_d = 0$. Additional features may include the characteristics of people involved, such as their age or gender, or contextual factors such as the time and emotion associated with the experience. The set of features is sufficiently large that no two experiences are exactly identical.

We focus on the case in which the experiences in the database reflect the objective frequency of events (that of different causes of death in our example). In principle, the database could be person-specific (e.g., people from New York may hear of fewer experiences of death from tornado than do people from Des Moines) and could also be affected by repetition, rehearsal, and prominence of events (e.g., people may hear of more experiences of airplane crashes than of diabetes due to greater news coverage of the former). Furthermore, the database may include statistical information, as is the case in many experimental settings (Benjamin 2019). The database could be influenced by selective attention. A past smoker concerned with lung cancer could encode many events of this disease (Schwartzstein 2014). We leave such extensions to future work.

The DM forms beliefs about the relative frequency of two disjoint hypotheses H_1 and H_2, which are subsets of the database E. For instance, the DM may assess the frequency of death by H_1 = "natural disaster" versus H_2 = "all other causes." These hypotheses partition the subset of causes of death, identified by $f_d = 1$, on the basis of the "natural disaster" feature $f_{d+2} = 1$ versus $f_{d+2} = 0$. As we describe later, the DM makes his assessment by extracting a sample from his database. Critically, sampling is shaped by similarity and interference, in line with memory research (Kahana 2012; Bordalo et al. 2020a). Next we present our formalization of similarity.

II.A. Similarity

A symmetric function $S(u, v) : E \times E \to [0, \bar{S}]$ measures the similarity between any two experiences u and v in the database. It reaches its maximum $\bar{S}$ at $u = v$. Similarity between two experiences increases in the number of shared features. For instance, a death from a tornado is more similar to that from flooding than either is to death from diabetes, because the former are caused by a natural disaster rather than an illness. Different features may be differently weighted based on their importance or salience.

Episodes of a heart attack are similar to each other even if they occur in different contexts. We rely on general intuitions about similarity, not on a particular functional form. A rich literature measures subjective similarity between objects and connects it to observable features (Tversky 1977; Nosofsky 1992; Pantelis et al. 2008).

We define the similarity between two subsets of the database $A \subset E$ and $B \subset E$ to be the average pairwise similarity of their elements,

$$S(A, B) = \sum_{u \in A} \sum_{v \in B} S(u, v) \frac{1}{|A|} \frac{1}{|B|} . \tag{1}$$

$S(A, B)$ is symmetric and increases in feature overlap between members of A and B. The similarity between two disjoint subsets of E is positive if their elements share some features.

We use equation (1) to define four important objects. The first is the similarity $S(e, H_i)$ between a single experience $e \in E$ and a hypothesis H_i. It increases in the extent to which e shares features with the average member of H_i. Obviously, e = "flood" is similar to H_1 = "natural disaster," while e = "diabetes" is very dissimilar to it. The second object is the self-similarity of hypothesis H_i, $S(H_i, H_i)$. It measures the homogeneity of H_i. Consider H_1 = "natural disaster": a tornado in Tulsa is fairly similar to a tornado in Little Rock, but neither is as similar to an earthquake in California, which reduces the self-similarity of H_1. The third object is "cross-similarity" between hypotheses $S(H_1, H_2)$. In H_1 = "natural disaster", a death from a flood is similar to a death from accidental drowning in H_2, which raises $S(H_1, H_2)$. The fourth and final object is the cross-similarity between H_i and the rest of the database, $\bar{H} = E \backslash H_i \cup H_j$, denoted by $S(H_i, \bar{H})$. When assessing the frequency of different causes of death, $\bar{H}$ is the set of nonlethal events. In H_1 = "natural disaster," a death from flood is similar to the event of surviving a flood in $\bar{H}$, which raises $S(H_1, \bar{H})$. Throughout, we focus on the case in which a hypothesis is more similar to itself than to other parts of the database, $S(H_i, H_i) \geqslant \max\{S(H_i, H_j),\ S(H_i, \bar{H})\}$.[2]

2. This condition can be violated if H_1 has two opposite clusters and H_2 is in the middle. Consider a database with two generic features, and suppose that the DM assesses hypotheses $H_1 \equiv \{(1,0),(0,1)\}$ and $H_2 \equiv \{(1,1)\}$. Here members of

II.B. Memory Sampling

Our formalization of similarity-based sampling and its mapping with beliefs builds on two assumptions. The first formalizes cued recall.

ASSUMPTION 1. *Cued Recall: When cued with hypothesis H_i, the probability $r(e, H_i)$ that the DM recalls experience e is proportional to the similarity between e and H_i. That is,*

$$r(e, H_i) = \frac{S(e, H_i)}{\sum_{u \in E} S(u, H_i)}. \tag{2}$$

In the numerator of equation (2), sampling is shaped by similarity to the cue H_i. When thinking about deaths from H_i = "natural disasters," it is relatively easy to recall e = "deaths from floods," due to similarity. The denominator in equation (2) captures interference: all experiences $u \in E$ compete for retrieval, so they inhibit each other. When we think about death from H_i = "natural disasters," the mind may retrieve experiences of different yet frequent lethal events such as e = "death from a heart attack." If $S(u, v)$ is constant, sampling is frequency-based, so $r(e, H_i) = \frac{1}{N}$.

Interference reflects the fact that we cannot fully control what we recall.[3] It is a well-established regularity in memory research going back to the early twentieth century (Jenkins and Dallenbach 1924; McGeoch 1932; Underwood 1957).[4] Our application of interference to probability estimates is new. We show that it produces biases linked to the availability and representativeness heuristics.

Our second assumption is that, given the probability of recall function $r(e, H_i)$, probability judgments are formed according to the following two-stage sampling process:

H_1 disagree along all features, while H_2 agrees with one of them, so $S(H_1,H_1) < S(H_1,H_2)$.

3. Interference need not happen consciously. Recall failures may manifest as "mental blanks," inability to recall anything when thinking about H_i, or as "intrusions," namely, recall of hypothesis-inconsistent experiences $u \notin H_i$.

4. For example, recall from a target list of words suffers intrusions from other lists studied at the same time, particularly for words that are semantically related to the target list, resulting in a lower likelihood of retrieval and longer response times (Shiffrin 1970; Anderson and Spellman 1995; Lohnas, Polyn, and Kahana 2015). A related phenomenon is "false memories" (Brown, Buchanan, and Cabeza 2000).

ASSUMPTION 2. *Sampling and Counting.*

Stage 1: For each hypothesis H_i, the DM samples $T \geqslant 1$ experiences from E with replacement according to $r(e, H_i)$. Denote by R_i the number of successful recalls of experiences in H_i.

Stage 2: The DM estimates the probability of H_i, denoted $\hat{\pi}(H_i)$, as the share of successful recalls of H_i out of all successful recalls of the hypotheses considered:

$$\hat{\pi}(H_i) = \frac{R_i}{R_1 + R_2} \tag{3}$$

Intuitively, the DM draws two random samples, one for each hypothesis.[5] He then counts the number of successes in recalling each H_i, discarding intrusions, and finally estimates the probability of a hypothesis as its relative share of successful recall attempts.

The assumption that each hypothesis is sampled separately (stage 1) is realistic when the different hypotheses are prominently presented to the DM, which is the case in our experiments. It may be violated if the DM must represent an entire distribution without being cued with specific values (e.g., the age distribution of deaths). In this case, some outcomes may fail to come to mind. The assumption that the DM assesses probabilities by counting "successes" in the drawn samples (stage 2) is realistic in one-shot estimation problems but may fail in repeated settings, because the DM may learn about the selected nature of the recalled samples. Such learning is unlikely to be perfect, for it itself is subject to memory limitations. Relatedly, the sample size T may be optimized based on the DM's thinking effort. We also assume that when counting "successes," the DM recognizes whether a retrieved memory is consistent with any of the hypotheses. In practice, the DM may have a noisy recollection of a given experience or may distort recalled experiences self-servingly, as documented in the literature on the hindsight bias (Roese and Vohs 2012). These might be promising extensions of the model.

Finally, in our model the DM forms beliefs by counting the retrieved experiences consistent with each hypothesis and by discarding intrusions. Our model can be extended to account for situations in which the person assesses a novel scenario, so probability estimates do not involve only counting. In this case, sam-

5. Sampling with replacement has two interpretations. The first is that the sample size is small relative to N. The second is that repeated recall of certain events makes them more prominent in mind, affecting beliefs. This is consistent

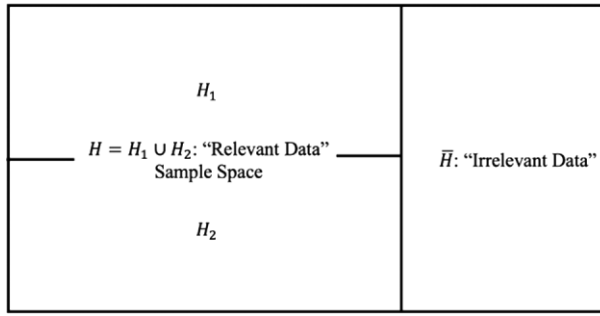

FIGURE I
Memory Database and Sample Space

pling from memory can be accompanied by the simulation of the novel scenario, as documented by a substantial body of work in psychology (Kahneman and Tversky 1981; Schacter, Addis, and Buckner 2007, Schacter et al. 2012). For example, when assessing the likelihood that a person is a feminist bank teller, a DM who has never met one may simulate the hypothesis using memories of people who are similar. Bordalo et al. (2022a) incorporate simulation into a model of memory-based beliefs and show it helps account for puzzling patterns in beliefs about COVID lethality.[6]

We view our model as the simplest way to introduce similarity into a sampling model. We judge its success by its ability to account for well-known biases, including strong violations of consistency such as partition dependence and the conjunction fallacy, and for recall data.

II.C. Beliefs

To understand the forces shaping belief formation, consider Figure I.

---

with the finding that unique events such as a stock market crash persistently affect beliefs (Malmendier and Nagel 2011).

6. Our model can also be enriched by allowing for (i) sampling to be influenced also by the most recently recalled item, (ii) the DM to count intrusions from $u \in H_j$, and (iii) retrieval to be driven by factors other than similarity. For instance, an experience may be more memorable if it is extreme or surprising (Kahneman et al. 1993), or if it is similar to experiences in other contexts, for example, names of celebrities are more easily remembered (Tversky and Kahneman 1973).

The hypotheses H_1 and H_2 identify three subsets of experiences in E.[7] They also subdivide E into relevant versus irrelevant experiences. $H = H_1 \cup H_2$ is the set of relevant experiences. In statistics, H is the sample space. The DM forms his subjective beliefs over it. Experiences in $\bar{H} = E \backslash H$, are "irrelevant," because they are inconsistent with either hypothesis. When thinking about a hypothesis H_i, similarity causes the DM to focus recall on subset H_i, but by similarity and frequency, sampling may erroneously slip to H_j and $\bar{H}$.

Denote by $\pi(H) = \frac{|H|}{|E|}$ the frequency of relevant data in the database and by $\pi(\bar{H}) = \frac{|\bar{H}|}{|E|}$ the frequency of irrelevant data. Denote by $\pi(H_i) = \frac{|H_i|}{|H|}$ the true relative frequency of H_i in the relevant data H, that is, the correct probability.[8] The total probability that the DM successfully recalls experiences of H_i when thinking about H_i is then given by:

$$
\begin{aligned}
r(H_i) &= \sum_{e \in H_i} r(e, H_i) \\
&= \frac{\sum_{e \in H_i} S(e, H_i)}{\sum_{u \in H_i} S(u, H_i) + \sum_{u \in H_j} S(u, H_i) + \sum_{u \in \bar{H}} S(u, H_i)} \\
&= \frac{\pi(H_i)\pi(H)}{\pi(H_i)\pi(H) + \frac{S(H_i, H_j)}{S(H_i, H_i)} \cdot \pi(H_j)\pi(H) + \frac{S(H_i, \bar{H})}{S(H_i, H_i)} \cdot \pi(\bar{H})}. \qquad (4)
\end{aligned}
$$

In psychology, $r(H_i)$ is known as the retrieval fluency of H_i. In the denominator of equation (4), it is ceteris paribus easier to recall a more frequent hypothesis. If similarity is constant, fluency only depends on frequency, $r(H_i) = \frac{\pi(H_i)}{\pi(H)}$.[9]

7. In a slight abuse of notation, we refer to H_i both as a given hypothesis, for example, "cause of death is flood," and the subset of experiences in E consistent with hypothesis H_i.

8. More precisely, $\pi(H_i)$ is the probability of H_i conditional on the relevant data H. To ease notation, we do not refer to $\pi(H_i)$ as $\pi(H_i|H)$, until we later study conditional beliefs in which the relevant data H is restricted to a subset D.

9. In equation (4) we use the equality $\sum_{e \in H_i} S(e, H_j) = S(H_i, H_j)|H_i|$ which follows from equation (1). Note that similarity does not matter when the DM either samples all data (i.e., $S(H_i, H_j) = S(H_i, H_i) = S(H_i, \bar{H})$), or all relevant data (which occurs when $S(H_i, H_j) = S(H_i, H_i)$, $S(H_i, \bar{H}) = 0$), with equal probability regardless of the cue. In both cases, the expression in equation (4) becomes proportional to $\pi(H_i)$, so that beliefs are unbiased.

Similarity shapes sampling in two ways: cuing and interference. Cuing means that thinking about $H_i =$ "flood" cues selective recall of deaths from floods. This effect is stronger when self-similarity $S(H_i, H_i)$ is higher. Cuing implies that the retrieval fluency of H_i is higher than its frequency, especially for unlikely hypotheses. People rarely experience floods and earthquakes compared to heart attacks, so cuing $H_1 =$ "natural disasters" boosts their retrieval.

Interference is captured by the denominator of equation (4) and works in two ways. First, H_i faces "interference from the alternative hypothesis" H_j. When thinking about deaths from $H_i =$ "flood," the mind may retrieve deaths due to causes similar to flood, such as "accidental drownings" or other natural disasters, that belong to $H_j =$ "other causes of death." In Figure I, this corresponds to "vertical" intrusions from H_j. Such intrusions are more common when the two hypotheses are more similar, $S(H_i, H_j)$ is higher. Second, H_i faces "interference from irrelevant data" $\bar{H}$. When thinking about deaths from $H_i =$ "flood," the mind may retrieve experiences of "surviving floods" that belong to nonlethal events in $\bar{H}$. In Figure I, this corresponds to horizontal intrusions from $\bar{H}$. This effect also hinders sampling of H_i, the more so the higher is cross-similarity $S(H_i, \bar{H})$.

We describe the probabilistic assessment $\hat{\pi}(H_i)$ in equation (3). By Assumption 2, the number of successes in recalling each hypothesis H_i follows a binomial distribution: $R_i \sim Bin(T, r(H_i))$. Beliefs $\hat{\pi}(H_i)$ are thus stochastic and characterized as follows.

PROPOSITION 1. *As $T \mapsto \infty$ the distribution of the estimated odds of H_i relative to H_j converges in distribution to a Gaussian with mean and variance*:

$$\mathbb{E}\left[\frac{\hat{\pi}(H_i)}{\hat{\pi}(H_j)}\right] = \frac{r(H_i)}{r(H_j)}, \tag{5}$$

$$\mathbb{V}\left[\frac{\hat{\pi}(H_i)}{\hat{\pi}(H_j)}\right] = \frac{1}{T}\left[\frac{r(H_i)}{r(H_j)}\right]^2\left[\frac{1-r(H_j)}{r(H_j)} + \frac{1-r(H_i)}{r(H_i)}\right]. \tag{6}$$

In equation (5), the DM attaches a higher probability to hypotheses with relatively high retrieval fluency, as in Tversky and Kahneman's (1973) availability heuristic. If similarity does not drive recall, for example, $S(u, v)$ is constant, beliefs are frequency

based. In this case, average odds in equation (5) are correct, $\frac{r(H_i)}{r(H_j)} = \frac{\pi(H_i)}{\pi(H_j)}$, but beliefs display noise in equation (6) due to sampling variance.

When similarity matters, biases arise. We study this case by assuming $S(H_i, H_i) > \max\{S(H_i, H_j),\ S(H_i, \bar{H})\}$. We focus on biases in average beliefs. For noise, the model implies that when two hypotheses are easy to recall, $r(H_1)$ and $r(H_2)$ are high, the DM uses a larger sample so belief variability declines. In Online Appendix B we test this and other predictions.

III. Judgment Biases

Section III.A shows how similarity affects interference from the alternative hypothesis, yielding biases related to the availability heuristic. Section III.B incorporates interference from irrelevant data, and shows that it accounts for the representativeness heuristic. Section III.C shows that these two forces can unify over- and underreaction of beliefs to data.

III.A. Similarity and Interference from the Alternative Hypothesis

To study interference from the alternative hypothesis, we restrict to the case in which the database E coincides with the relevant data for assessing H_1 and H_2 (or equivalently that similarity falls very sharply when moving outside H). In our example, this means that the DM only samples causes of death and there is no intrusion from unrelated events. Furthermore, we assume that T is high enough that average odds are characterized by equation (5).

Lichtenstein et al. (1978) document the overestimation of cued low-probability events, such as death from botulism or a flood, and underestimation of cued and likely causes, such as heart disease. The average assessed odds in equation (5) produce this phenomenon.

Proposition 2. *Holding $S(H_i, H_j)$ fixed, the estimate $\hat{\pi}(H_1)$ increases in the objective frequency $\pi(H_1)$. Overestimation, that is, $\hat{\pi}(H_1) > \pi(H_1)$, occurs if and only if the hypothesis is sufficiently unlikely, $\pi(H_1) < \pi^*$, where threshold π^* is*

defined by:

$$\frac{\pi^*}{1-\pi^*} \equiv \frac{1-\frac{S(H_1,H_2)}{S(H_1,H_1)}}{1-\frac{S(H_1,H_2)}{S(H_2,H_2)}}. \tag{7}$$

If both hypotheses are equally self-similar, $S(H_1, H_1) = S(H_2, H_2)$*, then* $\pi^* = 0.5$.

Overestimation of an unlikely hypothesis is due to cued recall of its instances, which occurs because the self-similarity of H_i is higher than its cross-similarity with H_j.[10] When thinking about H_1 = "floods," the DM selectively retrieves deaths due to floods, oversampling this rare event compared to H_2 = "other causes of death." Similarity thus creates insensitivity to frequency, a tendency for beliefs to be smeared toward 50:50.

Kahneman and Tversky's (1979) probability weighting function also features insensitivity to true frequency when weighting objective probabilities.[11] Our model applies to the construction of subjective probabilities and implies that an unlikely event may be over- or underestimated, due to interference. In sharp contrast with KT's probability weighting function, in our model unlikely events are prone to be neglected when they are not directly cued. To see this, consider the frequency with which a DM thinking about H_2 = "causes other than flood" samples elements of its subset H_{21} = "tornado" $\subset H_2$ compared with other elements in H_2. Such relative frequency, given by $\frac{r(H_{21},H_2)}{r(H_2)}$, is the belief the agent implicitly puts on tornadoes compared with H_2.

COROLLARY 1. *A subhypothesis* $H_{21} \subset H_2$ *is undersampled compared to its true frequency in* H_2 *if and only if* $S(H_{21}, H_2) < S(H_2, H_2)$*. Denote* $H_{22} = H_2 \backslash H_{21}$*. Holding fixed the similarity subhypotheses* $(S(H_{21}, H_{21}), S(H_{22}, H_{22}), S(H_{21}, H_{22}))$*, there*

10. The probability of a hypothesis $\pi(H_i)$ can be varied while holding similarities $S(H_i, H_j)$ fixed by keeping the distribution of similarity within hypotheses constant. Formally, this can be obtained by increasing the frequency of each $e_i \in H_i$ proportionally. Equation (1) is in fact homogeneous of degree zero with respect to such change in E.

11. Recent work microfounds this function based on the salience of lottery payoff (Bordalo, Gennaioli, and Shleifer 2012), noisy perception of numerical probabilities (Frydman and Jin 2022; Khaw, Li, and Woodford 2021), and cognitive uncertainty (Enke and Graeber 2019). In the Online Appendix, we show that recall is strongly correlated with a measure of subjective uncertainty.

is a threshold π^{**} such that H_{21} is undersampled compared to its true frequency in H_2 if and only if $\pi(H_{21}) < \pi^{**}$.

Roughly speaking, a noncued event H_{21} is neglected if it is less similar to the cued hypothesis H_2 to which it belongs compared to the hypothesis's average member, $S(H_{21}, H_2) < S(H_2, H_2)$. This depends in part on the event's frequency: the rarer is H_{21}, the more atypical it is of the cued H_2 and hence the more dissimilar it is to the latter. When thinking about H_2 = "causes other than flood," we may recall the likely "heart attack," not the unlikely "tornado." Underestimation of noncued events has important implications for economic choice. In Section V, we show how this logic helps explain systematic patterns of undersaving for retirement, arising from the neglect of heterogeneous, noncued expenses.

A second implication of similarity is that an event can be overestimated if homogeneous but underestimated if heterogeneous, regardless of its likelihood.

COROLLARY 2. *Holding fixed $\pi(H_1)$, as the events in H_1 become more homogeneous, that is, $S(H_1, H_1)$ increases, the probability assessment $\hat{\pi}(H_1)$ increases. If $S(H_1, H_1) > S(H_2, H_2)$, the threshold of Proposition 2 satisfies $\pi^* > 0.5$, and H can be overestimated even if it is likely.*

When H_1 becomes more self-similar, it is easier to recall. As a result, it is less likely that, when thinking about it, the mind slips to its alternative hypothesis H_2. According to equation (5), this increases the estimation of H_1, even if its objective probability stays constant.

This result indicates that cued unlikely events are prone to overestimation because they are often more self-similar than their alternative. When cued by H_1 = "flood," it is easy to imagine instances of this disaster, because they are similar to each other. By contrast, the alternative H_2 = "causes other than flood" is very heterogeneous and hence hard to imagine. This creates strong interference for H_2, hindering its assessment.

Tversky and Kahneman (1983) asked one group of subjects to assess the share of H_1= "words ending with _n_" in a certain text. Another group of subjects was asked to assess the probability of H_{11} = "words ending with _ing". Remarkably, subjects attached a lower probability to H_1 than to H_{11}, despite the fact that H_{11} is a subset of H_1. Similarity accounts for this phenomenon: instances of H_{11} = "words ending with _ing" share many features, such as

being gerunds or denoting similar activities, which brings many examples to mind. In contrast, H_1 = "words ending with _n_" includes many words that do not share these features (and often do not share many features with each other). This reduction in self-similarity makes it harder to recall words in H_1, causing its underestimation compared to its subset H_{11}.[12]

A third implication, following from Corollaries 1 and 2, is partition dependence. The total likelihood of death is estimated to be lower for "natural causes" than for "cancer, heart attack, or other natural causes" (Tversky and Koehler 1994). Many famous studies document this phenomenon (Benjamin 2019).[13] In our model, it arises because partitioning a hypothesis into more specific subevents increases its overall self-similarity, reducing interference. To see this, suppose that the alternative hypothesis H_2 is explicitly partitioned into H_{21} and H_{22}. The subsets are equally (i) likely, $\pi(H_{21}) = \pi(H_{22})$, (ii) self-similar, $S(H_{21}, H_{21}) = S(H_{22}, H_{22})$, and (iii) cross-similar to H_1, $S(H_{21}, H_1) = S(H_{22}, H_1)$. These conditions nest three hypotheses (H_1, H_{21}, H_{22}) in the binary hypotheses case, connecting to Proposition 2. We then obtain:

PROPOSITION 3. *Partitioning the alternative hypothesis H_2 into H_{21} and H_{22} is equivalent to increasing the self-similarity of H_2 if and only if:*

$$S(H_{21}, H_{21}) > S(H_{21}, H_{22}). \tag{8}$$

In this case, partitioning H_2 reduces $\hat{\pi}(H_1)$, and the more so the higher is $\frac{S(H_{21}, H_{21})}{S(H_{21}, H_{22})}$.

The assessment of a given hypothesis H_1 = "flood" is reduced when its alternative is specified as H_{21} = "natural causes" and H_{22} = "nonnatural causes other than flood," compared with when

12. To check this intuition, we ran a simple online survey. Respondents indeed rate randomly generated groups of "_ing" words as being more similar to each other than groups of _n_ words. Results are available on request.

13. For example, Fischhoff, Slovic, and Lichtenstein (1978) famously show that when assessing the cause of a car's failure to start, mechanics judge "ignition" more likely when residual causes were partitioned into "fuel" and "other." Sloman et al. (2004) show, in contrast, that death by "pneumonia, diabetes, cirrhosis or any other disease" is estimated to be less likely than death by "any disease." This is consistent with an extension of our model in which atypical cues such as "cirrhosis" focus attention on a narrow subset, interfering with the retrieval of more common diseases. A similar pattern occurs in free recall tasks (Sanborn and Chater 2016).

it is specified as H_2 = "causes other than flood." Cuing H_{21} and H_{22} fosters retrieval of alternatives to flood, which reduces the assessment of H_1 = "flood." Tversky and Koehler's (1994) "support theory" offers an explanation based on the idea that people evaluate events using a subadditive "support function." In our model, partition dependence comes from similarity in recall.

In sum, with similarity-based sampling, the DM evaluates a hypothesis by retrieving instances of it, but in doing so finds it hard not to think about the alternative hypothesis. Such interference reconciles well-known biases including overestimation of cued rare events, underestimation of rare events that are not cued, and availability effects in which the similarity structure of hypotheses and their description affect beliefs. Online Appendix A0 summarizes the main biases explained by our model and the required conditions on the similarity function.

III.B. Biases due to Interference from Irrelevant Experiences

Until now, we ruled out interference from irrelevant data by assuming that the database E coincides with the relevant data $H = H_1 \cup H_2$. Suppose, however, that the DM must condition H_1 and H_2 on data D, which identifies a subset $D \subset H$. For concreteness, the DM assesses deaths by H_1 = "accident" versus H_2 = "sickness" in the specific group of D = "young." The DM samples the events $H_1 \cap D$ = "accidents among the young" and $H_2 \cap D$ = "sickness among the young" using the retrieval fluencies $r(H_1 \cap D)$ and $r(H_2 \cap D)$. These retrieval fluencies are still defined by equation (4) with the change in notation $H = D$ and $\bar{H} = \bar{D}$.

Critically, now irrelevant experiences from $\bar{D} = E \backslash D$ can interfere, in our example those of $\bar{D}$ = "older" people. We show that this kind of interference produces effects typically explained using the representativeness heuristic. To visualize such interference, Figure II depicts the database E, where the size of each region roughly corresponds to true frequencies.

When thinking about $H_1 \cap D$ = "accident among the young," two kinds of interference are at work. First, as in our prior analysis, there is vertical intrusion of memories of young people dying from sickness (i.e., from $H_2 \cap D$), due to similarity among young people. Second, there is horizontal intrusion of irrelevant experiences of older people dying from accidents (i.e., from $H_1 \cap \bar{D}$), because of the similarity along the H_1 = "accident" dimension. Similarly, when thinking about $H_2 \cap D$ = "sickness among the

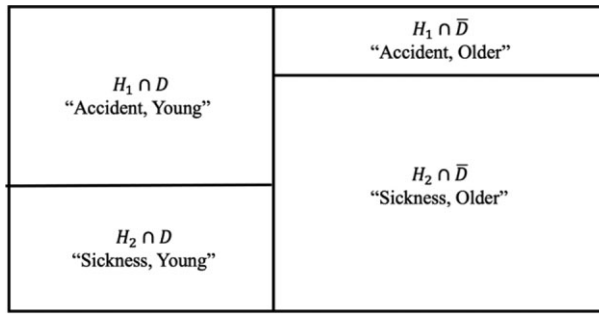

FIGURE II

Visualizing Conditional Assessments

young," the DM faces vertical intrusion from "accidents among the young" and horizontal intrusions from the irrelevant "sickness among the older."

Interference from $\bar{D}$ may affect one hypothesis more than the other. In our example, the deaths of older people interfere more with thinking about "sickness" because the bulk of the elderly die from sickness, not from accidents. Thus, when thinking about young people dying from sickness, many old people dying from sickness intrude, while intrusions are few when thinking about accidents. This effect can cause overestimation of $H_1 \cap D =$"accident among the young."

Formally, suppose that there are only two features (in our case the cause of death, accident versus sickness, and age, young versus older). The DM assesses the distribution of the first feature (cause of death) conditional on a value of the other (young). Suppose furthermore that similarity takes the functional form: $S(e, e') = \delta^{\sum_i |f_i - f_i'|}$, so it decreases by a factor of δ for each differing feature. We denote the conditional probability estimate obtained using equation (4) by $\hat{\pi}(H_i|D)$, and we compare it to the true conditional probability $\pi(H_i|D)$.

PROPOSITION 4. *For $\delta < 1$, the DM overestimates the probability of H_1 conditional on D, $\hat{\pi}(H_1|D) > \pi(H_1|D)$, if and only if:*

$$\pi(H_1|D)\pi(D) + \delta\pi\left(H_1|\bar{D}\right)\pi(\bar{D}) < \frac{\pi(D) + \delta\pi(\bar{D})}{2}. \tag{9}$$

The first term on the left side is standard: overestimation is more likely when the true conditional probability $\pi(H_1|D)$ is low,

in line with Section III.A. The second term is new: the conditional hypothesis is overestimated also if its frequency in the irrelevant data, $\pi(H_1|\bar{D})$, is low. In this case, H_1 is less similar to the irrelevant data $\bar{D}$ than H_2. Thus, H_1 faces less interference than H_2 from irrelevant data, which promotes overestimation of the former.

Consider this effect in Figure II. H_1 = "accident" is a common cause of death for the young ($\pi(H_1|D)$ is high), so interference from the alternative hypothesis promotes its underestimation. At the same time, when considering young people dying from sickness, many instances of the old dying from sickness intrude ($\pi(H_1|\bar{D})$ is low). This can cause overestimation of H_1 = "accident" for the young, even if for them it is the more likely cause of death.

Intrusion of irrelevant data sheds light on Kahneman and Tversky's (1973) representativeness heuristic, including the so-called conjunction fallacy in the Linda problem. Subjects are told that Linda was an activist in college, so D = "activist." Some are then asked the probability that she is currently a H_1 = "bank teller," others that she is a H_{11} = "feminist bank teller." Strikingly, feminist bank teller is rated likelier than bank teller, even though $H_{11} \subset H_1$. According to Proposition 4, this occurs for two reasons. First and foremost, H_{11} = "feminist bank teller" is much less similar to the group of $\bar{D}$ = "nonactivists" than H_1 = "bank teller." Intuitively, among "nonactivists" there are many fewer feminist bank tellers than bank tellers, $\pi(H_1|\bar{D}) > \pi(H_{11}|\bar{D})$. Thus, H_{11} = "feminist bank teller" faces less interference from irrelevant data than H_1 = "bank teller," which promotes overestimation of H_{11}. Second, H_1 = "bank teller" is likelier than H_{11} = "feminist bank teller," $\pi(H_1|D) > \pi(H_{11}|D)$, which also promotes underestimation of H_1 relative to H_{11}. Both effects create this conjunction fallacy.

Tversky and Kahneman (1983, 296) define representativeness as follows: "an attribute is representative of a class if it is very diagnostic; that is, the relative frequency of this attribute is much higher in that class than in a relevant reference class." Gennaioli and Shleifer (2010) formalize this idea by assuming that the conditional probability $\pi(H|D)$ is overestimated if the likelihood ratio $\frac{\pi(H|D)}{\pi(H|\bar{D})}$ is high. The conditioning data $\bar{D}$ in the denominator captures the "reference class" in the above definition. Bordalo et al. (2016) use this formula to model social stereotypes, Bordalo, Gennaioli, and Shleifer (2018) use it to model diagnostic expectations.

Intrusion from irrelevant data provides a foundation for representativeness and the reference class $\bar{D}$ in a way that squares with Kahneman and Tversky's (1973) broad intuition that similarity judgments affect beliefs. When $\pi(H_1|\bar{D})$ is low, H_1 is dissimilar to the irrelevant data $\bar{D}$. As a result, H_1 suffers less interference from $\bar{D}$ than does H_2, so experiences in $H_1 \cap D$ are easier to retrieve, causing overestimation of $\pi(H_1|D)$. In Section V.C we show that this mechanism explains "kernel of truth" stereotypes (Bordalo et al. 2016) and yields new predictions.

One advantage of our approach is to identify limits to representativeness, which are due to strong intrusion from the alternative hypothesis. We now show how the interaction between these forces throws new light on the conflicting evidence of over- and underreaction.

III.C. Underreaction and Overreaction to Data

Work from the lab and the field documents conflicting distortions in belief updating. There is evidence that people overestimate the probability of events in light of data which is informative about them, a finding often explained by the representativeness heuristic (Kahneman and Tversky 1973). There is also evidence of underestimation in similar situations, often explained with inattention (Sims 2003; Coibion and Gorodnichenko 2012; Gabaix 2019). Memory helps unify this evidence, yielding conditions under which either phenomenon should occur.

To connect to this debate, we define over- and underreaction. We say that the DM overreacts to data D if (i) D is objectively informative about a hypothesis H_i, that is, $\pi(H_i|D) > \pi(H_i)$, and (ii) the DM overestimates that hypothesis, that is, $\hat{\pi}(H_i|D) > \pi(H_i|D)$. The DM underreacts otherwise. This definition captures the intuition of overreaction in many real-world settings in which the DM's prior belief and the likelihood function are unavailable, as with stereotypes (red-haired Irish) or the Linda problem.[14] Proposition 4 implies the following result.

14. In Online Appendix A4, we show that our definition is equivalent to saying that the DM overreacts if and only if an upward revision of his belief in response to the data ($\hat{\pi}(H_i|D) > \hat{\pi}(H_i)$) is associated with an overestimation, or negative prediction error ($\hat{\pi}(H_i|D) > \pi(H_i|D)$). This criterion is often used to detect over- and underreaction in the field, using data on revisions of expectations (Coibion and Gorodnichenko 2012, and Bordalo et al. 2019b). When priors and likelihoods are available, under- and overreaction are often defined in terms of sensitivities rather than levels, as is done in Grether (1980), that is, in terms of the difference between the elicited prior and posterior beliefs.

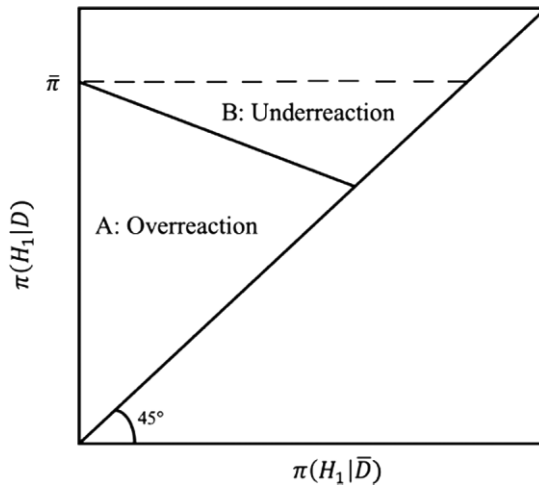

FIGURE III

Condition for Underreaction and Overreaction to Data

This figure depicts the region of $(\pi(H_1|D),\ \pi(H_1|\bar{D}))$ where the agent overreacts or underreacts to data D, where D is diagnostic of H $(\pi(H_1|D) > \pi(H_1|\bar{D}))$. Region A corresponds to overreaction, region B to underreaction.

COROLLARY 3. *Suppose that D is informative about H_1. If the true probability $\pi(H_1|D)$ is higher than a threshold $\bar{\pi} > 0.5$, the DM underreacts to D. If $\pi(H_1|D) < \bar{\pi}$, the DM overreacts to D if $\pi(H_1|D)$ or $\pi(H_1|\bar{D})$ are sufficiently low, and underreacts to D otherwise.*

In Figure III, the data are informative about H_1 in the region above the 45° line.[15] In region A, H_1 is overestimated, whereas in region B it is underestimated, as per equation (9).

Consider different cases, starting from the two most extreme ones. In the lower left corner of region A, overreaction is strong because interference is low from the irrelevant data $(\pi(H_1|\bar{D})$ is low) and from the alternative hypothesis $(\pi(H_1|D)$ is low). The resulting overreaction takes the form of base rate neglect. In Tversky and Kahneman (1974), people overestimate the chances that Steve, a "shy and withdrawn person with a passion for detail," is

15. This is equivalent to $\pi(H_1|D) > \pi(H_1)$. The characterization is identical for H_2 when $\pi(H_1|D) < \pi(H_1)$. In fact, overreaction is given by either $\hat{\pi}(H_1|D) < \pi(H_1|D) < \pi(H_1)$ or $\hat{\pi}(H_1|D) > \pi(H_1|D) > \pi(H_1)$.

a librarian rather than a farmer, even though farming is a much more common occupation, especially among men. Overreaction occurs because librarians are relatively rare ($\pi(H_1|D)$ is low) and because many farmers are neither shy nor have a passion for detail, so farmers are more similar to irrelevant data than librarians ($\pi(H_1|\bar{D})$ is low, which implies that $\pi(H_2|\bar{D})$ is high).

At the other extreme, in the upper part of Figure III, when $\pi(H_1|D) > \bar{\pi}$, interference from the alternative hypothesis is very strong. Here underreaction occurs and takes the form of general conservatism and aversion to extreme beliefs (Griffin and Tversky 1992; Benjamin 2019). Even though the data point to H_1, cuing the unlikely alternative H_2 causes its overestimation.

In the intermediate region of Figure III, H_1 is moderately likely. Whether over- or underreaction prevails depends on the signal's strength. Overreaction occurs when the signal is strong, in the upper part of region A. An investor may overestimate the probability that a firm has H_1 = "strong fundamentals" if it has experienced D = "rapid earnings growth." This occurs provided firms with strong fundamentals rarely exhibit lackluster growth, that is, $\pi(H_1|\bar{D})$ is low. In this case, rapid earnings growth makes it easy to think about strong fundamentals. It is instead harder to think about H_2 = "weak fundamentals," because these firms often produce $\bar{D}$ = "not rapid earnings growth." The DM overreacts because H_1, although likely, faces much less interference from irrelevant data than H_2. This logic offers a microfoundation for diagnostic expectations.

On the other hand, if the hypothesis is moderately likely but the signal is weak, there is underreaction, due to high interference from irrelevant data $\pi(H_1|\bar{D})$. Suppose that the DM evaluates a firm and the data are not "rapid growth" but rather D = "positive earnings surprise." Even firms with strong fundamentals may have negative earnings surprises, so $\pi(H_1|\bar{D})$ is higher than in the previous example. This creates interference for H_1, potentially causing its underestimation and underreaction. If D points to a fairly likely hypothesis, beliefs underreact to weakly diagnostic data and overreact to strongly diagnostic data.

To summarize, similarity and interference in human recall naturally account for a range of well-documented biases due to Kahneman and Tversky's availability and representativeness heuristics and shed light on conflicting evidence on under- and overreaction.

IV. Experiments

We assess our key predictions in two "pure recall" experiments in which we modulate similarity and interference by exogenously varying subjects' databases and cues. Experiment 1 studies the role of interference from the alternative hypothesis. Experiment 2 additionally studies interference from irrelevant data. In both experiments, subjects first go through a controlled set of experiences in which they see a series of images, and then they make a probabilistic assessment about them. To do so, they only need to recall the images they saw earlier. Relative to conventional designs, which provide subjects with statistical information (e.g., Edwards 1982; Enke and Graeber 2019) or ask hypothetical questions about naturalistic situations (Kahneman and Tversky 1973), our approach (i) allows us to control the memory database, (ii) avoids anchoring to given numerical probabilities, and (iii) enables us to measure recall of specific experiences and thus assess whether recall and probability estimates go hand in hand.

Subjects were recruited from Bocconi University undergraduates on the experimental economics email list. They could participate in both experiments, which occurred four months apart, and completed the experiments remotely due to COVID restrictions. They earned a €4 Amazon gift card, plus a bonus if their answer to one randomly chosen question was accurate.[16] Experiments were preregistered, including hypotheses and sample sizes, on the AEA RCT Registry, with ID AEARCTR-0006676. Online Appendix C provides more details about both surveys.

IV.A. Experiment 1: Testing Interference from the Alternative Hypothesis

This experiment tests three key implications of interference from the alternative hypothesis.

Prediction 1: Memory creates a tendency to overestimate cued unlikely hypotheses, and overestimation is stronger for rarer hypotheses (Proposition 2).

16. If the chosen question was a probability estimate, they earned €2 if their answer was within 5 percentage points of the truth. If it was a free recall task, each correctly/incorrectly recalled word increased/decreased subjects' chance of winning the bonus by 10 percentage points (bounded by 0 and 1). The bonus provides easy-to-understand incentives compared with other schemes such as binarized scoring rules, which distort truth telling (Danz, Vesterlund, and Wilson 2020).

Prediction 2: Holding objective frequencies constant, the assessed probability of a hypothesis increases when its alternative is more heterogeneous/less self-similar (Corollary 2).

Prediction 3: Holding objective frequencies constant, the assessed probability of a hypothesis decreases if its alternative is partitioned into two more self-similar subsets (Proposition 3).

Participants are told that they will see 40 words, one by one in a random order. They are told that they will then be asked questions about the words and that answering correctly will raise their chances of winning a bonus payment. They are not told what the questions will be. Participants answer three comprehension questions that ask them to redescribe each piece of the instructions. Eighty-nine percent of respondents answer all three questions correctly. The results we present are unchanged if we exclude the 11% who answered at least one question incorrectly.

In all treatments, some of the words are animals and some are not, though participants are not informed of this ahead of time. In three treatments they are then asked the following question: "Suppose the computer randomly chose a word from the words you just saw. What is the percent chance that it is. . .

an animal? ____%
anything else? ____%"

The two probabilities must add up to 100%. Afterward, participants are asked to list up to 15 animals and then up to 15 other words that they remember seeing. In all treatments, all exhibited words are relevant to answering the question. Thus, there is no interference from irrelevant data.

The four treatments to test Predictions 1–3 are:

T1: 20% of the words are animals; 80% of the words are names (half male and half female).
T2: 40% of the words are animals; 60% of the words are names (half male and half female).
T3: 40% of the words are animals; the remaining words do not belong to any common category, and hence are very dissimilar to one another.
T4: The distribution of words is as in T2, but subjects are asked about the probability of animals, men's names, and women's names. Assessments must add up to 100%.

TABLE I
TREATMENTS IN EXPERIMENT 1

Treatment	Sample size	Distribution of images	Examples	Elicited belief
T1	244	20% animals, 80% names	Lion, John, moose, rat, Margaret, deer, Edward, Nancy, wolf . . .	P(animal) versus P(other)
T2	244	40% animals, 60% names	Paul, John, moose, rat, Margaret, deer, Laura, Nancy, Edward . . .	P(animal) versus P(other)
T3	241	40% animals, 60% heterogeneous	Lion, sled, moose, rat, pure, deer, half, good, wolf . . .	P(animal) versus P(other)
T4	234	40% animals, 60% names	Lion, John, moose, rat, Margaret, deer, Edward, Nancy, wolf . . .	P(animal) versus P(men) versus P(women)

These experimental treatments are summarized in Table I.[17]

Comparing T1 and T2 offers a test for Prediction 1: we expect overestimation of $\hat{\pi}(animal)$ especially in T1, when animals are objectively rarer. By comparing T2 and T3 we can test Prediction 2: compared with T2, $\hat{\pi}(animal)$ should be higher in T3, because

17. Three remarks on our experiments. First, heterogeneous words in T3 were chosen using a random word generator, eliminating words that we deemed too similar to each other (e.g., mayor, elected, town). Second, in the recall task in T4, participants are asked to list up to 15 examples each of animals, men's names, and women's names. Third, in addition to treatments T1–T4, we ran a treatment T5, where we replaced women's names in T1 with ocean animals (e.g., shark, starfish, dolphin). Participants are then asked the probability of "land animals" (in T1, all animals are land animals) and "anything else." In the recall task, participants are asked to list examples of "land animals" and "other words" that they recall seeing. By increasing cross-similarity $S(H_1, H_2)$, this treatment should exert an ambiguous effect on assessments, but it should reduce the ability to recall examples of H_1 = "land animals." Though the recall data appear consistent with this hypothesis, there was an unexpected confusion about what counted as a land animal: over a quarter of respondents list at least one ocean animal in the free recall task when asked to list land animals. For comparison, no respondents list names when prompted to recall animals in T1. We are therefore less confident in the data from T5 and exclude it from the main analysis. The Online Appendix describes this issue and the results from this treatment in greater detail.

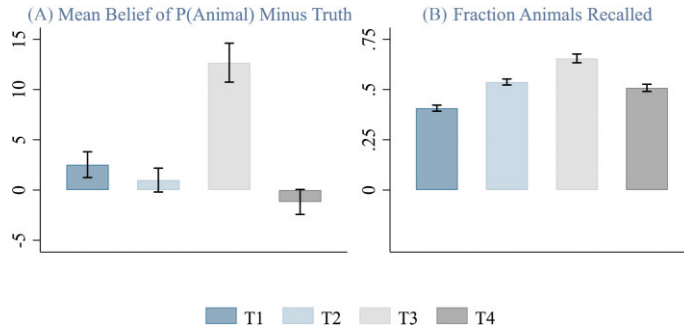

FIGURE IV

Results from Experiment 1

This figure shows mean belief of the probability of animals (Panel A) and the mean fraction of recalled words that were animals (Panel B) in Experiment 1. Bands show 95% confidence intervals. The distribution of words for each treatment are: T1: 20% animals, 40% men's names, 40% women's names, T2: 40% animals, 30% men's names, 30% women's names, T3: 40% animals, 60% heterogeneous words, T4: 40% animals, 30% men's names, 30% women's names.

the alternative hypothesis (nonanimals) is very heterogeneous. By comparing T4 and T2, we can test Prediction 3: in T4 the alternative hypothesis is split into two more self-similar subhypotheses (men's names and women's names), so $\hat{\pi}(animal)$ should be lower than in T2. Last, the treatment effects on the recall task should mirror those on $\hat{\pi}(animal)$. This is not necessarily due to a causal effect of recalled examples on probability estimates, as both outcomes may be products of retrieval fluency.

IV.B. Experiment 1 Results

Figure IV shows the treatment effects. Panel A reports the under- or overestimation of $\hat{\pi}(animal)$ compared to the truth. Panel B reports the share of animals among all recalled examples.[18]

Consistent with Prediction 1, there is a tendency to overestimate $\hat{\pi}(animal)$, especially in T1, where animals are only 20%

18. Throughout the analysis that follows, we look at treatment effects on the number of correctly recalled words. About 18% of the answers to recall questions (which were free text entry) are not in fact words that were shown to participants or are words corresponding to other hypotheses. Unless otherwise noted, results look very similar if we instead use the number of recall entries (regardless of whether they were correct or incorrect) for a category.

of words: overestimation of animals (that is, mean belief minus truth) is 2.5 percentage points in T1 and 1 percentage point in T2, and only the former is significantly different from zero at conventional levels ($p < .01$ and $p = .10$, respectively). Also, the overestimation in T1 is marginally statistically different from that in T2 ($p = .09$).

The result of T3 is striking: consistent with Prediction 2, when we replace people's names with heterogeneous words while keeping the true frequency of "animal" constant at 40%, the overestimation of $\hat{\pi}(animal)$ increases from 1 percentage point in T2 to 12.7 percentage points in T3 ($p < .01$). Thus, overestimation depends not only on actual frequency but also on how self-similar the alternative hypothesis is. This effect can dominate attenuation to 50:50: in T3, $\hat{\pi}(animal)$ overshoots 50% ($p = .01$). The role of similarity in recall emerges as a powerful force in probabilistic assessments.

Finally, when partitioning "nonanimals" into the finer subhypotheses "men's names" and "women's names" in T4, the assessment $\hat{\pi}(animal)$ falls by 2.1 percentage points compared with T2 ($p = .013$) and is even underestimated relative to the truth ($p = .060$). Similarity-based recall implies that the more specific cues in the partition of H_2 can turn overestimation of an unlikely hypothesis H_1 (as in T2) into its underestimation (in T4).

The treatment effects on recall, shown in Figure IV, Panel B, mirror those on beliefs. Significantly fewer (40% versus 54%) of recalled words are animals in T1 compared to T2 ($p < .01$), because there are objectively fewer animal words in the former. In T3, where nonanimals are heterogeneous words, recall of animals jumps to 66%, significantly higher than in T2 ($p < .01$). Finally, in T4, where men's and women's names are separated out, significantly fewer recalled words are animals (50%) than in T2 ($p = .02$).[19]

One might worry that our results are driven by differences in attention or encoding of words across treatments, rather than by the intended retrieval mechanism. However, until the words are presented, all treatments are identical to participants, so our

19. Although the treatment effects on recall and probability are aligned qualitatively, the exact magnitudes need not align. Indeed, the magnitude of the effect on recall seems to be greater than the effect on probability estimation. In general, the explicitly recalled samples and the internal recall fluency used in probability judgments may not be the same.

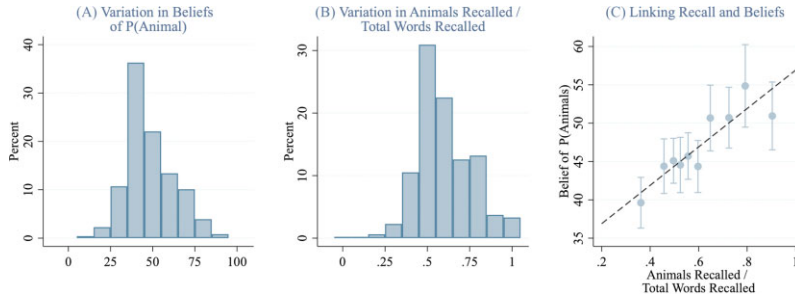

FIGURE V

The Relationship between Recall of Examples and Beliefs

Panel A shows the distribution of beliefs about the probability of animals in Experiment 1. Panel B shows the distribution of the number of animals recalled divided by the total number of words recalled. Panel C bins the data by deciles of animals recalled divided by total number of recalled words (x-axis) and shows mean beliefs of the probability of animals (y-axis). The dashed line shows the OLS line of best fit. Bands show 95% confidence intervals. All panels restrict the data to T2 and T3.

treatment effects cannot be attributable to differences in what participants were told they would see or be asked. In addition, in T1, T2, and T3, the same question is asked, and the only difference between treatments is the words that are presented. Comparing T2 and T4, we see that the distribution of words is identical, and only the question-cue is changed. Differences in cuing/retrieval are thus likely driving our results.[20]

We conclude the analysis of Experiment 1 by looking at the link between beliefs and recall at the individual level.

Figure V pools the T2 and T3 treatments, where the true distribution of words includes 40% animals and 60% nonanimals (results look similar if we include the other treatments). There is substantial heterogeneity in beliefs (Panel A) and in the fraction of recalled words that are animals (Panel B). Crucially, beliefs and recall are highly correlated: respondents who recall relatively more animals estimate the probability of drawing an animal to be higher, adding credence to our interpretation that beliefs and free

20. The recall data reveal some primacy effects whereby words that were (randomly) presented earlier in the sequence are more likely to be recalled. See Online Appendix C for more details, including robustness to controlling for such effects.

recall are both dependent on retrieval fluency (Panel C).[21] In the Online Appendix, we also show that, in line with equation (6), participants who recall more words (a proxy for T) also have less variable beliefs (Kahneman, Sibony, and Sunstein 2021).

IV.C. Experiment 2: Interference from Irrelevant Data

We designed Experiment 2 to test the implications of interference from irrelevant data. Participants are told that they will be shown 40 images, each of which is either a word or a number, and either orange or blue. Participants are not told how many images would be of each color, but in all treatments 20 images are orange and 20 are blue. Participants are told that their bonus will depend on their answer to questions about the images but are not told what the question will be. After seeing the images, one by one in a random order, participants in all treatments are asked: "Suppose the computer randomly chose an image from the images you just saw. It is orange. What is the percent chance that it is a word?"

Participants must thus assess the probability $\hat{\pi}(w|o)$ that an image is a word conditional on the data that it is orange. Participants answer by clicking on a slider that ranges from 0% to 100%.[22] They are then asked to list up to 10 orange words that they recall seeing.

In this experiment, a subset of experiences—blue words and blue numbers—are irrelevant for assessing the distribution of orange images. Crucially, as subjects try to recall orange words (numbers), the irrelevant blue words (numbers) may come to mind and interfere. Of course, interference from the alternative hypothesis is also at play: when thinking about orange words, orange numbers may also come to mind, causing smearing toward 50:50.

To identify interference from irrelevant data, we fix the share of orange images that are words, $\pi(w|o)$, and vary the share of blue images that are words, $\pi(w|b)$. In *Low* treatments, no blue image is a word, $\pi(w|b) = 0$, in *High* treatments all blue images are words, $\pi(w|b) = 1$, and in *Middle* treatments either half or 30%

21. The correlation is not causal, but it is also not mechanical: subjects are separately asked the percent chance that a randomly chosen word is an animal and then to recall up to 15 examples of each hypothesis.

22. The slider begins with no default, so that participants have to click somewhere on the slider and then move the draggable icon (that appears where they first click) to indicate their answer.

of blue images are words, $\pi(w|b) = 0.5,\ 0.3$. Our model predicts that higher $\pi(w|b)$ should reduce both the estimated $\pi(w|o)$ and recall of orange words.

We study how interference from irrelevant data interacts with that from the alternative hypothesis. To do so, we vary the share of orange images that are words, $\pi(w|o)$, that is, the correct answer. In *Neutral* treatments the true answer is 50%, $\pi(w|o) = 0.5$. In *Intermediate* treatments the true answer is 55%, $\pi(w|o) = 0.55$. In *Common* treatments the true answer is 70%, $\pi(w|o) = 0.7$. Due to growing interference from the alternative hypothesis, treatments with higher $\pi(w|o)$ should see a stronger tendency toward underestimating orange words, potentially even if there is very little interference from irrelevant data, namely, even if $\pi(w|b)$ is very low.

In the same experimental setting, Bordalo et al. (2020a) show that increasing the association of irrelevant data with a category causes a lower probability estimate for that category, in line with our treatments varying $\pi(w|b)$ here. The novelty of Experiment 2 is to contrast this force with interference from the alternative hypothesis by varying $\pi(w|o)$ and setting $\pi(w|o) \geqslant 0.5$. This is key, for it helps assess whether interference from the alternative hypothesis may be a source of underreaction to data as described in Figure III.[23]

Table II describes all treatments, identified by the acronym of the true answer N(eutral), I(ntermediate), C(ommon), and interference from irrelevant data, L(ow), M(iddle), and H(igh).

IV.D. Experiment 2 Results

Figure VI, Panel A reports, for each treatment, the difference between the average assessment $\hat{\pi}(w|o)$ and the true fraction $\pi(w|o)$ of orange images that are words. Panel B reports the average number of orange words recalled by subjects in each treatment.

Consistent with our model, stronger interference from irrelevant data (higher $\pi(w|b)$) reduces the assessment $\hat{\pi}(w|o)$ that an orange image is a word, across all treatments ($p < .01$ in each case). When normatively irrelevant blue words are more numerous, interference in recall of orange words is stronger. Recall data

23. We did not include an *IM* treatment due to sample size limitations.

TABLE II

TREATMENTS IN EXPERIMENT 2

Treatment	Distribution	Distribution of irrelevant data	Sample size (N)	Elicited belief
Neutral	50% orange words, 50% orange numbers	NL: 0% blue words	147	P(word | orange)
		NM: 50% blue words	146	
		NH: 100% blue words	151	
Intermediate	55% orange words, 45% orange numbers	IL: 0% blue words	158	P(word | orange)
		IH: 100% blue words	154	
Common	70% orange words, 30% orange numbers	CL: 0% blue words	154	P(word | orange)
		CM: 30% blue words	149	
		CH: 100% blue words	144	

Notes. This table describes the treatments in Experiment 2. For all treatments, the L and H subtreatments consist of 0% and 100% blue words, respectively. The *Neutral* and *Common* treatments also have an M subtreatment, which is 50% blue words for *Neutral* and 30% for *Common*.

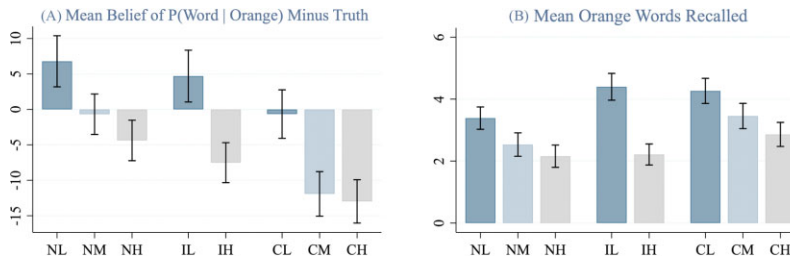

FIGURE VI

Testing Prediction 4

Panel A shows the average belief that the randomly drawn image is a word conditional on it being orange minus the true conditional probability. Panel B shows the average number of correctly recalled orange words. In the L treatments, all blue images are numbers. In the H treatments, all blue images are words. In the M treatment when 70% of orange images are words (CM), 30% of blue images are words. In the M treatment when 50% of orange images are words (NM), 50% of blue images are also words. Bands show 95% confidence intervals.

in Panel B support this mechanism: subjects recall fewer correct orange words when $\pi(w|b)$ is higher.[24]

In line with predictions, overestimation of $\hat{\pi}(w|o)$ arises only if orange words are rare enough, namely, in the *Neutral* and *Intermediate* treatments. In the *Common* treatments, when $\pi(w|o) = 0.7$, there is no overestimation of $\hat{\pi}(w|o)$ even in the extreme case of $\pi(w|b) = 0$.[25]

In sum, Experiment 2 is consistent with two key predictions of Proposition 4. First, underestimation of a likely hypothesis can be turned into overestimation if the recall of the alternative hypothesis faces strong interference from irrelevant yet sufficiently similar experiences. This is evident from the switch from underestimation to overestimation of $\pi(w|o) = 0.55$ as we move from treatment IH to IL (and the consistent drop in the recall of orange words).

Second, if the hypothesis is very likely (as in the *Common* treatments), overestimation disappears because interference from the alternative hypothesis becomes very strong. This occurs even if interference from irrelevant data $\pi(w|b)$ is low, as evident in treatments CL and CM. Treatment CM shows another key prediction of our model: when orange is a weaker signal of the image being a word, beliefs underreact to the orange data, $\hat{\pi}(w|o) < \pi(w|o)$, while they are well calibrated when the signal is strong in CL, here $\hat{\pi}(w|o) \approx \pi(w|o)$. In the other treatments, when the data are indicative of a rare hypothesis, beliefs overreact, as predicted by our theory.[26] Balancing

24. This result, unlike the others in this section, looks different if we focus only on the number of words that participants list in the recall tasks (as opposed to counting the number of correctly recalled words). Participants actually list more words as being orange in high interference subtreatments, though significantly fewer correct orange words. We think that this occurs because it is much easier to guess words that may have been orange in treatments in which there are more words overall. This issue does not arise in Experiment 1 (which occurred chronologically after Experiment 2) because there we focus on categories for which it is difficult to incorrectly list a word as being in the wrong category.

25. These results cannot be explained by the fact that subjects misinterpret our request for P(word | orange) as asking for either P(orange) or P(orange word). If so, their answers should not depend on the distribution of blue words. If they interpreted the question as asking for P(word), the effect should be opposite to what we observe.

26. In NH, IH, and CH, the data $D =$ "*orange*" is informative of $H_2 =$ "number." Because in our treatments $\pi(n|o) < 0.5$, here we are in the lower part of Region A, in which the data point to an unlikely hypothesis. Consistent with the model,

interference from the alternative hypothesis and from irrelevant data accounts for both over- and underreaction of beliefs to data. Regularities in selective memory unify different biases in probability judgments.

V. Applications: Similarity And Interference In Economic Decisions

The mechanisms of memory speak to many economic settings. In this section, we discuss some of them, such as saving decisions, the pricing of insurance and Arrow-Debreu securities, and labor market stereotypes. We start with a general setup. A DM evaluates an action a, which yields payoff $u(a)$ today and a state-contingent payoff $u_s(a)$ tomorrow, with the probability of state $s \in \{1, 2\}$ given by $\pi_1(a)$ and $\pi_2(a)$, respectively. The expected utility of action a is:

$$V(a) = u(a) + \sum_{s\in\{1,2\}} u_s(a)\pi_s(a). \tag{10}$$

The action a could be the decision to save, to purchase a security with state-contingent payoffs, or to hire a worker with a particular skill.

Equation (10) highlights a key feature of standard models: the expected utility of an action only depends on its influence on the payoff in each state s and its objective probability. In particular, payoff probabilities are sufficient statistics for valuation, and the different contingencies in which the same payoff is delivered do not matter.[27] Similarity in recall breaks down this invariance in two important ways. First, the subjective probability of a state depends not only on its objective probability but also on the similarity of experiences associated with each state, which can be influenced by its description. More homogeneous states are easier to retrieve and receive greater decision weights. Second, the DM's beliefs and actions also depend on the interference from experiences irrelevant to the decision at hand, such as those regarding a counterfactual group or action. In the following applications, we

in these treatments we see overreaction: $\hat{\pi}(n|o)$ is overestimated or equivalently $\hat{\pi}(w|o)$ is underestimated.

27. This is also true for prospect theory and its extensions (e.g., Kőszegi and Rabin 2006), where the subjective decision weights and the reference points depend only on the objective probabilities and the payoffs.

study the implications of these two violations of invariance: our results on savings and Arrow-Debreu security prices highlight the former, and those on social stereotypes and labor market discrimination highlight the latter.

V.A. *Similarity, Interference, and Savings Decisions*

A growing body of work connects undersaving to cognitive mistakes rather than to present bias. Consumers systematically underestimate future expenditures, a phenomenon known as the "planning fallacy" (Peetz and Buehler 2009). In particular, they fail to account for idiosyncratic events such as a speeding ticket, a medical need, or a car repair (Sussman and Alter 2012). A *Wall Street Journal* column advises that retirement spending averages $400 more per month than expected because of surprising outlays: "These are bills outside what we normally would expect: the garage door spring and cable that snapped and had to be replaced; the family member who asked for financial help; the X-rays and dentists' fee for a sudden toothache; the small tree in our yard that, it turned out, was dying and needed to be removed; the storm that damaged the screens on our porch; the stone that hit and cracked our windshield; the request from a charity that we felt we needed to honor. The list goes on" (Ruffenach 2022). Failure to account for such events causes undersaving. Augenblick et al. (2022) link such mispredictions of unusual events to savings and spending by farmers in Zambia, which leads to hunger prior to the harvest.

A notable feature of the events described in the press and academic studies is the extreme diversity of unexpected spending shocks. Our model delivers this phenomenon as a form of systematic forgetting caused by the dissimilarity/heterogeneity of such shocks. Suppose that a is current saving out of normal income Y, so that $u(a) = u(Y - a)$. Future utility is then $u_1(a) = u(Y + a)$ under normal conditions and $u_2(a) = u(Y - L + a)$ under an expenditure shock $L > 0$. $u(.)$ is an increasing and concave Von Neumann-Morgenstern utility function.

The shock hits with exogenous probability $\pi_2 < 0.5$. It can arise from $N > 1$ mutually exclusive causes s_{2i}, $i = 1, \ldots, N$, each occurring with probability $\pi_{2,i}$, where $\sum_i \pi_{2i} = \pi_2$. With expected utility in equation (10), only the total probability π_2 of the shock matters and the DM's optimal choice is simple: savings increase in π_2, $\frac{da}{d\pi_2} > 0$. As idiosyncratic spending needs become more

frequent, the DM transfers more resources to the high marginal utility "shock" state.

Consider a DM with limited memory. When thinking about saving, the DM estimates the probability of $H_2 =$ "s_2" by retrieving each shock s_{2i} experienced in the past. In his database, s_{2i} is encoded in terms of its cause and the income loss it was associated with. Formally, s_{2i} is a vector of $N+1$ features. The first feature takes the value 1 because loss L was borne. Feature $i+1$ takes value 1 because the loss was of type i. All other features are 0. The normal income state s_1 is then a vector of $N+1$ zeroes. Vectors are encoded with their frequencies $(\pi_{2i}), \pi_1$.

Similarity among experience vectors shapes retrieval. The self-similarity of an experience is maximal and equal to 1, $S(s_1, s_1) = S(s_{2,i}, s_{2,i}) = 1$. Any pair of other experiences, by contrast, differ along two features. The normal state s_1 and any shock state $s_{2,i}$ differ along the occurrence of the loss L and its cause i. Shocks $s_{2,i}$ and $s_{2,j}$ differ on their causes: i and j. Denote by $\Delta \in [0, 1]$ the drop in similarity entailed by two dissonant features. We then have $S(s_1, s_{2,i}) = S(s_{2,i}, s_{2,j}) = 1 - \Delta$ for $i = 1, \ldots, N$ and $i \neq j$. This in turn implies:

$$S(s_1, s_2) = 1 - \Delta, \tag{11}$$

$$S(s_2, s_2) = 1 - \Delta(1 - C), \tag{12}$$

where $C = \sum_i \left(\frac{\pi_{2,i}}{\pi_2}\right)^2$ is the Herfindhal index of concentration of the aggregate shock state s_2 across the N different causes of expenditure shocks.

Equation (12) embeds the key interference mechanism. The DM's ability to estimate the overall probability of the income loss L depends on how heterogeneous the loss state is. If s_2 is fully concentrated on a single cause, $C = 1$, self-similarity is maximal, $S(s_2, s_2) = 1$, and s_2 is easy to recall. If s_2 is fully dispersed among infinitely many idiosyncratic spending needs, $C \to 0$, then the self-similarity of s_2 is minimal, so experiences of income loss are harder to retrieve.

By plugging equations (11) and (12) into equation (5), the total estimated probability of experiencing an expenditure shock

is given by:

$$\hat{\pi}_2(\pi_2, C) = \frac{\pi_2(1-\Delta\pi_2)}{(1-\pi_2)\left[1-\frac{\Delta C}{1-\Delta(1-C)}(1-\pi_2)\right]+\pi_2(1-\Delta\pi_2)}. \quad (13)$$

In the absence of similarity-driven distortions, $\Delta = 0$, the DM is well calibrated: $\hat{\pi}_2(\pi_2, C) = \pi_2$. If $\Delta > 0$, his belief is distorted: $\hat{\pi}_2(\pi_2, C) \neq \pi_2$. More specifically, if the causes of income loss are sufficiently heterogeneous and dispersed,

$$C < C^* \equiv \frac{(1-\Delta)\pi_2}{1-(1+\Delta)\pi_2}, \quad (14)$$

the DM underestimates the frequency of the shock, $\hat{\pi}_2(\pi_2, C) < \pi_2$, and overestimates it otherwise.

Given that savings increase in the estimated probability $\hat{\pi}_2(\pi_2, C)$, similarity has important implications. If rainy days were due to a single cause, $C = 1$, the DM would overestimate their likelihood and oversave, consistent with overweighting of unlikely risks in Kahneman and Tversky's probability weighting function. But when rainy days come for many different reasons, C is low, recall of each specific reason faces a lot of interference, which causes forgetting. As a result, the DM underestimates the likelihood of s_2 and undersaves, consistent with the evidence. Savings decisions no longer satisfy the invariance of equation (10) and now depend on the similarity of anticipated future expenditures.

Similarity also generates "framing" effects: the DM will save more if specific shocks are described to him, because this frame boosts retrieval (as in Corollary 2). Augenblick et al. (2022) find that such intervention indeed increases savings, and Peetz et al. (2015) show that it increases predicted future spending. This mechanism also accounts for the "planning fallacy" (Kahneman and Tversky 1979), in which people systematically underestimate the time required to complete a task. The causes of delay are all different, which hinders their recall.

V.B. *Similarity and Asset Prices*

Selective memory has implications for pricing financial assets. Suppose that, rather than saving, the DM can purchase insurance against future income shocks. Formally, the DM chooses the quantity a of Arrow-Debreu securities on shock i to buy, where the security pays off L if a loss in state s_{2i} materializes. Denote the price of such a claim by P_i. If Arrow-Debreu securities are in zero net supply, the rational equilibrium price of insurance against

shock i is given by:

$$P_i^r = \mu \cdot \pi_{2,i} \cdot L,$$

where $\mu \equiv \frac{u'(Y-L)}{u'(Y)} > 1$ is the DM's marginal rate of substitution. Furthermore, prices are additive under rationality: the price of buying a broad insurance contract against any income loss of L is equal to the sum of the prices of all Arrow-Debreu claims, given by $P^r = \mu \cdot \pi_2 \cdot L$.

Selective memory creates a wedge between these prices. Suppose that all shocks are equally likely, $\pi_{2,i} = \frac{\pi_2}{N}$, and send $N \mapsto \infty$. The shock state is then fully dispersed, $C \mapsto 0$, and the price of the broad insurance contract is given by:

$$P_b = \mu \cdot \hat{\pi}_2(\pi_2, 0) \cdot L = \mu \cdot L \cdot \left(\frac{1 - \Delta\pi_2}{1 - \Delta\pi_2^2} \right) \pi_2, \tag{15}$$

which is less than the rational price P^r for any $\Delta > 0$. Intuitively, the DM fails to retrieve the different shocks insured by the broad contract and so undervalues that contract.

Consider instead the price of insuring any loss by buying Arrow-Debreu claims. The market price of doing so is the price of N identical claims, each one fully concentrated on s_{2i} and paying with probability $\frac{\pi_2}{N}$. That is, $P_i = \mu \cdot L \cdot \hat{\pi}_2(\frac{\pi_2}{N}, 1)$, so the total price of all claims is:

$$\lim_{N \to \infty} \mu \cdot L \cdot N \cdot \hat{\pi}_2 \left(\frac{\pi_2}{N}, 1 \right) = \mu \cdot L \cdot \left(\frac{1}{1 - \Delta} \right) \pi_2. \tag{16}$$

In stark contrast with the broad contract, the individual claims are overvalued compared to their rational price $P^r = \mu \cdot L \cdot \pi_2$. This is again due to similarity: as the DM thinks about each specific shock, he focuses on its occurrence, overestimating the insurance payout rate.

Similarity causes market prices to be subadditive, again breaking down the invariance of equation (10) with respect to fine state descriptions. Similarity explains why people are reluctant to buy broad (e.g., health) insurance but overpay for insuring specific unlikely risks, as documented for extended warranties (Abito and Salant 2019), flight insurance (Eisner and Strotz 1961), and specific diseases/causes of death (Johnson et al. 1993; Kunreuther and Pauly 2010). Relatedly, people are more likely to buy insurance after a disaster hits and gradually cancel the insurance if the policy has not paid out over time (Kunreuther and Pauly 2010). This is in line with the intuition that the possibility of disaster is cued and hence retrieved only right after its occurrence. More

broadly, by generating subadditive prices, selective memory can have an important impact on asset markets.[28]

V.C. *Minority Stereotypes and Illusory Correlation*

There is growing interest in economics in understanding social stereotypes, which shape discrimination in gender assessments (Bordalo et al. 2019a), labor markets (Neumark 2018), education (Carlana 2019), judicial decisions (Arnold, Dobbie, and Yang 2018), and politics (Bonomi, Gennaioli, and Tabellini 2021; Bordalo, Tabellini, and Yang 2020). Our model accounts for the "kernel of truth" model of stereotypes in Bordalo et al. (2016), but also helps explain additional findings from social psychology, which note that stereotypes are often directed at minorities (Hilton and von Hippel 1996) and may arise even in the absence of any group differences, as an illusory correlation (Sherman, Hamilton, and Roskos-Ewoldsen 1989).[29]

Before we present the analysis, consider the following example. A board of directors considers a female candidate for a CEO position. Suppose that most CEOs are competent, but that the vast majority of current CEOs are male. When considering a female candidate, the hypothesis that she is competent suffers strong interference from the large number of male CEOs, who dominate these positions. As a consequence of such interference from irrelevant data, the hypothesis that a female CEO candidate is competent might be underestimated. Interference from very common but irrelevant data becomes a source of an illusory correlation and stereotypes.

Formally, an employer decides whether to hire a worker from a minority group G, based on his beliefs about the worker's productivity. In terms of equation (10), hiring the worker ($a = 1$) yields high utility $u_1(a) = \theta_H a$ if the worker is productive, which occurs with probability $\pi_{H,G}$, and low utility $u_2(a) = -\theta_L a$ if the worker

28. For instance, this mechanism may help explain why investors are slower to evaluate news about a company that is "complicated" and consists of many heterogeneous subsidiaries, relative to "pure play" businesses (e.g., Cohen and Lou 2012), or why in a spinoff the parent is valued less than the equity carve out it owns (Lamont and Thaler 2003).

29. This effect, originally documented in the context of erroneous clinical judgments (Chapman 1967) is robustly produced in experiments, and has been proposed as a mechanism for negative views on minorities as well as for beliefs in nonsocial settings, for example, that bad weather is correlated with joint pain (Jena et al. 2017).

is unproductive (with probability $\pi_{L,G} = 1 - \pi_{H,G}$). For simplicity, we assume no utility cost in hiring ($u(a) = 0$). A rational DM hires the worker if the probability he is unproductive is low enough, $\pi_{L,G} < \pi^* \equiv \frac{\theta_H}{\theta_H + \theta_L}$.

A DM with selective memory forms belief $\hat{\pi}_{L,G}$ by sampling his past experiences. The memory database encodes two features: whether a worker is productive, H or L, and his social group, G or $\bar{G}$. Experiences that share only one feature have similarity $1 - \Delta$, whereas those differing in both features have similarity $1 - 2\Delta$. The extent of interference depends on the prevalence of the groups, denoted by p_G and $p_{\bar{G}} = 1 - p_G$, respectively, as well as of the low type in $\bar{G}$, denoted by $\pi_{L,\bar{G}}$. Using equation (5), the estimated odds that a G worker has low productivity is:

$$\frac{\hat{\pi}_{L,G}}{\hat{\pi}_{H,G}} = \frac{\pi_{L,G}}{\pi_{H,G}} \cdot \frac{p_G \Delta \left(1 + \pi_{H,G} - \pi_{H,\bar{G}}\right) + \Delta \pi_{H,\bar{G}} + (1 - 2\Delta)}{p_G \Delta \left(1 + \pi_{L,G} - \pi_{L,\bar{G}}\right) + \Delta \pi_{L,\bar{G}} + (1 - 2\Delta)}. \tag{17}$$

If $\Delta > 0$, the DM overestimates the probability that the worker from G is a low type if and only if:

$$\pi_{L,G} < \varphi \equiv \frac{\frac{\pi_{LG}}{\pi_{L\bar{G}}}}{p_G \frac{\pi_{LG}}{\pi_{L\bar{G}}} + p_{\bar{G}}}. \tag{18}$$

The DM has a negative stereotype of the worker from G if the share of experiences with low types from this group ($\pi_{L,G}$) is smaller than a threshold φ. In line with Proposition 4, low-frequency events tend to be overestimated. This threshold increases in the likelihood ratio of low types in G relative to $\bar{G}$, $\frac{\pi_{LG}}{\pi_{L\bar{G}}}$. The likelihood ratio is high when most members of the majority group $\bar{G}$ are high types ($\pi_{L\bar{G}}$ is low), so that they strongly interfere with the recall of high types in the minority group G. In turn, this causes an overestimation of low types in G. Stereotypes are an example of the second type of violation of the invariance in equation (10): the retrieval of irrelevant experiences in $\bar{G}$ intrude and distort assessments.

Interference provides a memory foundation for the stereotypes model of Bordalo et al. (2016), which relies on the likelihood ratio but also generates new predictions. In particular, it predicts that stereotypes should be stronger for a minority group because the strength of interference from $\bar{G}$ is especially high when the majority dominates the database. Formally, provided $\pi_{L,G} \geqslant \pi_{L,\bar{G}}$, equation (18) is easier to meet if p_G is low.

This effect can produce minority stereotypes even if different types are equally frequent in both groups, a phenomenon

known as illusory correlation (Sherman, Hamilton, and Roskos-Ewoldsen 1989). To see this, set $\pi_{L,G} = \pi_{L,\bar{G}} = \pi_L$ in equation (17) to obtain:

$$\frac{\hat{\pi}_{L,G}}{\hat{\pi}_{H,G}} = \frac{\pi_L}{\pi_H} \cdot \frac{p_G\Delta + 1 - \Delta - \Delta\pi_L}{p_G\Delta + 1 - \Delta - \Delta(1 - \pi_L)}. \tag{19}$$

When low types are rare, $\pi_L < 0.5$, their frequency is overestimated if and only if the group is a minority, $p_G < 0.5$. Notably, the stereotype emerges even though the share of low types in the two groups is the same. Recall of high types in G is inhibited by the many high types from $\bar{G}$ that flood the DM's memory database, while the reverse interference from G to $\bar{G}$ is far weaker.

VI. Conclusion

We have presented a model of memory-based probability judgments, with two main ingredients: (i) databases of experiences, and (ii) cues that trigger selective recall of these experiences. Recall is driven by similarity of the experiences to the cues, which include hypotheses and data. Similarity helps retrieve relevant experiences but also invites interference from experiences inconsistent with the hypothesis at hand (but similar to it). The new insight is that a hypothesis is underestimated when, compared to its alternative, it is more vulnerable to interference because it is more heterogeneous, more likely, or more similar to irrelevant data.

This notion that probability estimates are shaped by content (as captured by feature similarity) and not just by objective frequency accounts for and reconciles a wide range of seemingly inconsistent experimental and field evidence, including availability and representativeness heuristics proposed by Tversky and Kahneman (1974), overestimation of the probabilities of unlikely hypotheses, conjunction and disjunction fallacies in experimental data, and under- and overreaction to information. We tested several novel predictions of the model using an experimental design in which we control both the memory database and the cues subjects receive, and we found strong supportive evidence. Finally, we showed how memory-based beliefs shed light on several economic applications, linking undersaving and subadditivity of prices to failure to forecast heterogeneous states of the world under a broad cue, and minority stereotypes to interference from the larger majority group.

Our analysis opens the gates for many research directions, and in conclusion we list three we find particularly promising. First, probability judgments can pertain to events not yet experienced by the DM, such as forecasts of the future, or to events that are described in terms of statistics or data-generating processes (Benjamin 2019). Memory plausibly plays a central role in these settings as well. With respect to forecasts, a significant literature in psychology shows that the mental simulation of future events is intimately linked to memory processes (Dougherty, Gettys, and Thomas 1997; Brown, Buchanan, and Cabeza 2000). People combine past experiences with simulated ones (Kahneman and Miller 1986; Schacter, Addis, and Buckner 2007; Biderman, Bakkour, and Shohamy 2020), with the ease of simulation also driven by perceived similarity (Woltz and Gardner 2015). In this way, memory shapes forecasts. Bordalo et al. (2022a) incorporate simulation into the model presented here and apply it to studying beliefs about COVID when it was a novel threat. The model explains strong regularities in such beliefs, including the fact that through interference, experiencing nonhealth adversities leads to less pessimism about COVID lethality for the general population.

In addition, individuals often have both statistical and experiential information, such as in the literature on the description-experience gap in risky choice (Hertwig and Erev 2009). This research suggests an interaction between the two sources of information in generating beliefs, where statistical information may also act as a cue for retrieving semantic content from memory.

Expanding the model may lead to new predictions. One important direction is to better understand the drivers of retrieval, here summarized by a similarity function. Different people may interpret the same cue differently, depending in part on differences in their experiences, on their perceptions of similarity or attention to features of the stimulus, or on chance. The attention channel can be important. In an experiment with U.S. federal judges by Clancy et al. (1981), judges adjudicated a set of hypothetical criminal cases with multiple attributes. The authors found that different judges attended to different attributes of the case and proposed radically different sentences. Such heterogeneous responses may naturally occur if a DM's perceived similarity depends on the range of past experiences, or if these experiences influence the mental model that the decision maker uses (Schwartzstein 2014).

Another theoretical extension concerns learning and its distortions. For example, in our approach signals about an event

prime recall of previous experiences of the event itself, which may create a form of confirmation bias (Nickerson 1998).

Finally, our analysis focuses on the role of memory in probability estimates, but the applications of cued recall based on similarity to belief formation are much broader. The principles we described in this article can be applied to many problems, including consumer choice, advertising, persuasion, political positioning, and product branding.

University of Oxford, United Kingdom
Harvard University, United States
Bocconi University, Italy
Harvard University, United States
Harvard University, United States

Supplementary Material

Supplementary material is available at the *Quarterly Journal of Economics* online.

Data Availability

Data and code replicating the tables and figures in this article can be found in Bordalo et al. (2022b) in the Harvard Dataverse, https://doi.org/10.7910/DVN/CJUCHR.

References

Abito, Jose Miguel, and Yuval Salant, "The Effect of Product Misperception on Economic Outcomes: Evidence from the Extended Warranty Market," *Review of Economic Studies* 86 (2019), 2285–2318, https://doi.org/10.1093/restud/rdy045.

Anderson, Michael, and Barbara Spellman, "On the Status of Inhibitory Mechanisms in Cognition: Memory Retrieval as a Model Case," *Psychological Review*, 102 (1995), 68–100, https://doi.org/10.1037/0033-295x.102.1.68.

Arnold, David, Will Dobbie, and Crystal Yang, "Racial Bias in Bail Decisions," *Quarterly Journal of Economics*, 133 (2018), 1885–1932, https://doi.org/10.1093/qje/qjy012.

Augenblick, Ned, Kelsey Jack, Supreet Kaur, Felix Masiye, and Nicholas Swanson, "Budget Neglect in Consumption Smoothing: A Field Experiment on Seasonal Hunger" UC Berkeley working paper, 2022.

Azeredo da Silveira, Rava, Yeji Sung, and Michael Woodford. "Optimally Imprecise Memory and Biased Forecasts," NBER Working Paper no. 28075, 2020, https://doi.org/10.3386/w28075.

Barseghyan, Levon, Francesca Molinari, Ted O'Donoghue, and Joshua C. Teitelbaum, "The Nature of Risk Preferences: Evidence from Insurance Choices," *American Economic Review*, 103 (2013), 2499–2529, https://doi.org/10.1257/aer.103.6.2499.

Benjamin, Daniel, "Errors in Probabilistic Reasoning and Judgment Biases," *Handbook of Behavioral Economics: Applications and Foundations*, vol. 2, Douglas B. Bernheim, Stefano DellaVigna and David Laibson, Eds. (San Diego: Elsevier Science & Technology, 2019), Ch. 2, 69–186, https://doi.org/10.1016/bs.hesbe.2018.11.002.
Biderman, Natalie, Akram Bakkour, and Daphna Shohamy, "What Are Memories For? The Hippocampus Bridges Past Experience with Future Decisions," *Trends in Cognitive Sciences*, 24 (2020), 542–556, https://doi.org/10.1016/j.tics.2020.04.004.
Billot, Antoine, Itzhak Gilboa, Dov Samet, and David Schmeidler, "Probabilities as Similarity-Weighted Frequencies," *Econometrica*, 73 (2005), 1125–1136, http://www.jstor.org/stable/3598817.
Bonomi, Giampaolo, Nicola Gennaioli, and Guido Tabellini, "Identity, Beliefs, and Political Conflict," *Quarterly Journal of Economics*, 136 (2021), 2371–2411, https://doi.org/10.1093/qje/qjab034.
Bordalo, Pedro, Giovanni Burro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer, "Imagining the Future: Memory, Simulation, and Beliefs about Covid," Harvard University working paper, 2022a.
Bordalo, Pedro, John J. Conlon, Nicola Gennaioli, Spencer Y. Kwon, and Andrei Shleifer, "Replication Data for: 'Memory and Probability'," Harvard Dataverse, 2022b, https://doi.org/10.7910/DVN/CJUCHR.
Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, Frederik Schwerter, and Andrei Shleifer, "Memory and Representativeness," *Psychological Review*, 128 (2020a), 71–85, https://doi.org/10.1037/rev0000251.
Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer, "Stereotypes," *Quarterly Journal of Economics*, 131 (2016), 1753–1794, https://doi.org/10.1093/qje/qjw029.
———, "Beliefs about Gender," *American Economic Review*, 109 (2019a), 739–773, https://doi.org/10.1257/aer.20170007
Bordalo, Pedro, Nicola Gennaioli, Yueran Ma, and Andrei Shleifer, "Overreaction in Macroeconomic Expectations," *American Economic Review*, 110 (2020b), 2748–2782, https://doi.org/10.1257/aer.20181219.
Bordalo, Pedro, Nicola Gennaioli, Rafael La Porta, and Andrei Shleifer, "Diagnostic Expectations and Stock Returns," *Journal of Finance*, 74 (2019b), 2839–2874, https://doi.org/10.1111/jofi.12833.
Bordalo, Pedro, Nicola Gennaioli, Rafael La Porta, and Andrei Shleifer, "Belief Overreaction and Stock Market Puzzles," Harvard University working paper, 2022c.
Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer, "Salience Theory of Choice under Risk," *Quarterly Journal of Economics*, 127 (2012), 1243–1285, https://doi.org/10.1093/qje/qjs018.
———, "Diagnostic Expectations and Credit Cycles," *Journal of Finance*, 73 (2018), 199–227, https://doi.org/10.1111/jofi.12586.
———, "Memory, Attention, and Choice," *Quarterly Journal of Economics*, 135 (2020), 1399–1442, https://doi.org/10.1093/qje/qjaa007.
Bordalo, Pedro, Marco Tabellini, and David Y. Yang, "Issue Salience and Political Stereotypes," NBER Working Paper no. 27194, 2020, https://doi.org/10.3386/w27194.
Bouchaud, Jean-Philippe, Philipp Krüger, Augustin Landier, and David Thesmar, "Sticky Expectations and the Profitability Anomaly," *Journal of Finance*, 74 (2019), 639–674, https://doi.org/10.1111/jofi.12734.
Brown, Norman, Lori Buchanan, and Roberto Cabeza, "Estimating the Frequency of Nonevents: The Role of Recollection Failure in False Recognition," *Psychonomic Bulletin and Review*, 7 (2000), 684–691, https://doi.org/10.3758/BF03213007.
Carlana, Michela, "Implicit Stereotypes: Evidence from Teachers' Gender Bias," *Quarterly Journal of Economics*, 134 (2019), 1163–1224, https://doi.org/10.1093/qje/qjz008.
Chan, Louis, K.C., Narasimhan Jegadeesh, and Josef Lakonishok, "Momentum Strategies," *Journal of Finance*, 51 (1996), 1681–1713, https://doi.org/10.1111/j.1540-6261.1996.tb05222.x.

Chapman, Loren J., "Illusory Correlation in Observational Report," *Journal of Verbal Learning and Verbal Behavior*, 6 (1967), 151–155, https://doi.org/10.1016/S0022-5371(67)80066-5.
Chiappori, Pierre-Andre, Bernard Salanié, François Salanié, and Amit Gandhi, "From Aggregate Betting Data to Individual Risk Preferences," *Econometrica*, 87 (2019), 1–36, https://doi.org/10.3982/ECTA11165.
Clancy, Kevin, John Bartolomeo, David Richardson, and Charles Wellford, "Sentence Decision-Making: The Logic of Sentence Decisions and the Extent and Sources of Sentence Disparity," *Journal of Criminal Law and Criminology*, 72 (1981), 524–554.
Cohen, Lauren, and Dong Lou, "Complicated Firms," *Journal of Financial Economics*, 104 (2012), 383–400, https://doi.org/10.1016/j.jfineco.2011.08.006.
Coibion, Olivier, and Yuriy Gorodnichenko, "What Can Survey Forecasts Tell Us about Information Rigidities?" *Journal of Political Economy*, 120 (2012), 116–159, https://doi.org/10.1086/665662.
Danz, David, Lise Vesterlund, and Alistair Wilson, "Belief Elicitation: Limiting Truth Telling with Information on Incentives," NBER Working Paper no. 27327, (2020), https://doi.org/10.3386/w27327.
Dasgupta, Ishita, and Samuel J. Gershman, "Memory as a Computational Resource," *Trends in Cognitive Sciences*, 25 (2021), 240–251, https://doi.org/10.1016/j.tics.2020.12.008.
Dasgupta, Ishita, Eric Schulz, Joshua B. Tenenbaum, and Samuel J. Gershman, "A Theory of Learning to Infer," *Psychological Review*, 127 (2020), 412–441, https://doi.org/10.1037/rev0000178.
Dougherty, Michael R. P., Charles F. Gettys, and Rickey P. Thomas, "The Role of Mental Simulation in Judgments of Likelihood," *Organizational Behavior and Human Decision Processes*, 70 (1997), 135–148, https://doi.org/10.1006/obhd.1997.2700.
Dougherty, Michael R. P., Charles F. Gettys, and Eve E. Ogden, "MINERVA-DM: A Memory Processes Model for Judgments of Likelihood," *Psychological Review*, 106 (1999), 180–209, https://doi.org/10.1037/0033-295X.106.1.180.
Edwards, Ward, "Conservatism in Human Information Processing," in *Judgment under Uncertainty: Heuristics and Biases,* Daniel Kahneman, Paul Slovic and Amos Tversky, eds. (Cambridge: Cambridge University Press, 1982), 359–369, originally published in *Formal Representation of Human Judgment*, Benjamin Kleinmuntz, ed. (New York: John Wiley and Sons, Inc. 1968). https://doi.org/10.1017/CBO9780511809477.026.
Eisner, Robert, and Robert H. Strotz, "Flight Insurance and the Theory of Choice," *Journal of Political Economy*, 69 (1961), 355–368, http://www.jstor.org/stable/1828645.
Enke, Benjamin, and Thomas Graeber, "Cognitive Uncertainty," NBER Working Paper no. 26518, 2019, https://doi.org/10.3386/w26518.
Enke, Benjamin, Frederik Schwerter, and Florian Zimmermann, "Associative Memory and Belief Formation," NBER Working Paper no. 26664, 2020, https://doi.org/10.3386/w26664.
Fischhoff, Baruch, Paul Slovic, and Sarah Lichtenstein, "Fault Trees: Sensitivity of Estimated Failure Probabilities to Problem Representation," *Journal of Experimental Psychology: Human Perception and Performance*, 4 (1978), 330–344, https://doi.org/10.1037/0096-1523.4.2.330.
Frydman, Cary, and Lawrence Jin, "Efficient Coding and Risky Choice," *Quarterly Journal of Economics*, 137 (2022), 161–213, https://doi.org/10.1093/qje/qjab031.
Gabaix, Xavier, "Behavioral Inattention," *Handbook of Behavioral Economics: Applications and Foundations*, vol. 2, Douglas B. Bernheim, Stefano DellaVigna, and David Laibson, eds. (San Diego: Elsevier Science & Technology 2019), Ch.4, 261–343, https://doi.org/10.1016/bs.hesbe.2018.11.001.
Gennaioli, Nicola, and Andrei Shleifer, "What Comes to Mind," *Quarterly Journal of Economics*, 125 (2010), 1399–1433, https://doi.org/10.1162/qjec.2010.125.4.1399.

Gennaioli, Nicola, Andrei Shleifer, and Robert Vishny, "Neglected Risks, Financial Innovation, and Financial Fragility," *Journal of Financial Economics*, 104 (2012), 452–468, https://doi.org/10.1016/j.jfineco.2011.05.005.
Grether, David M., "Bayes Rule as a Descriptive Model: The Representativeness Heuristic," *Quarterly Journal of Economics*, 95 (1980), 537–557, https://doi.org/10.2307/1885092.
Griffin, Dale, and Amos Tversky, "The Weighing of Evidence and the Determinants of Confidence," *Cognitive Psychology*, 24 (1992), 411–435, https://doi.org/10.1016/0010-0285(92)90013-R.
Hertwig, Ralph, and Ido Erev, "The Description–Experience Gap in Risky Choice," *Trends in Cognitive Sciences*, 13 (2009), 517–523, https://doi.org/10.1016/j.tics.2009.09.004.
Hilton, James, and William von Hippel, "Stereotypes," *Annual Review of Psychology*, 47 (1996), 237–271, https://doi.org/10.1146/annurev.psych.47.1.237.
Jena, Anupam B., Andrew R. Olenski, David Molitor, and Nolan Miller, "Association between Rainfall and Diagnoses of Joint or Back Pain: Retrospective Claims Analysis," *British Medical Journal*, 359 (2017), https://doi.org/10.1136/bmj.j5326
Jenkins, John G., and Karl M. Dallenbach, "Obliviscence during Sleep and Waking," *American Journal of Psychology*, 35 (1924), 605–612, https://doi.org/10.2307/1414040.
Johnson, Eric J., Gerald Häubl, and Anat Keinan, "Aspects of Endowment: A Query Theory of Value Construction," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33 (2007), 461–474, https://doi.org/10.1037/0278-7393.33.3.461.
Johnson, Eric J., John Hershey, Jacqueline Meszaros, and Howard Kunreuther, "Framing, Probability Distortions, and Insurance Decisions," *Journal of Risk and Uncertainty*, 7 (1993), 35–51, https://doi.org/10.1007/BF01065313.
Kahana, Michael, *Foundations of Human Memory* (New York: Oxford University Press, 2012.
Kahneman, Daniel, Barbara L. Fredrickson, Charles A. Schreiber, and Donald A. Redelmeier, "When More Pain Is Preferred to Less: Adding a Better End," *Psychological Science*, 4 (1993), 401–405, https://www.jstor.org/stable/40062570.
Kahneman, Daniel, and Dale T. Miller, "Norm Theory: Comparing Reality to its Alternatives," *Psychological Review*, 93 (1986), 136–153, https://doi.org/10.1037/0033-295X.93.2.136.
Kahneman, Daniel, Olivier Sibony, and Cass Sunstein, *Noise: A Flaw in Human Judgment* (New York: Little, Brown, 2021).
Kahneman, Daniel, and Amos Tversky, "On the Psychology of Prediction," *Psychological Review*, 80 (1973), 237–251, https://doi.org/10.1037/h0034747.
———, "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, 47 (1979), 263–292, https://doi.org/10.2307/1914185.
———, "The Simulation Heuristic," Stanford University Working Paper, 1981.
Khaw, Mel Win, Ziang Li, and Michael Woodford, "Cognitive Imprecision and Small-Stakes Risk Aversion," *Review of Economic Studies*, 88 (2021), 1979–2013, https://doi.org/10.1093/restud/rdaa044.
Kőszegi, Botond, and Matthew Rabin, "A Model of Reference-Dependent Preferences," *Quarterly Journal of Economics*, 121 (2006), 1133–1165, https://doi.org/10.1093/qje/121.4.1133.
Kunreuther, Howard, and Mark V. Pauly, "Insuring against Catastrophes," in *The Known, the Unknown, and the Unknowable in Financial Risk Management: Measurement and Theory Advancing Practice,* Francis X. Diebold, Neil A. Doherety and Richard J. Herring, eds. (Princeton, NJ: Princeton University Press, 2010, 210–238), https://doi.org/10.1515/9781400835287-011.
Kwon, Spencer Yongwook, and Johnny Tang, "Extreme Events and Overreaction to News," *available at SSRN* (2021), https://dx.doi.org/10.2139/ssrn.3724420.
Lamont, Owen A., and Richard H. Thaler, "Can the Market Add and Subtract? Mispricing in Tech Stock Carve-Outs," *Journal of Political Economy*, 111 (2003), 227–268, https://www.journals.uchicago.edu/doi/full/10.1086/367683.

Lichtenstein, Sarah, Paul Slovic, Baruch Fischhoff, Mark Layman, and Barbara Combs, "Judged Frequency of Lethal Events," *Journal of Experimental Psychology: Human Learning and Memory*, 4 (1978), 551–578, https://doi.org/10.1037/0278-7393.4.6.551.

Lohnas, Lynn J., Sean M. Polyn, and Michael J. Kahana, "Expanding the Scope of Memory Search: Modeling Intralist and Interlist Effects in Free Recall," *Psychological Review*, 122 (2015), 337–363, https://doi.org/10.1037/a0039036.

Malmendier, Ulrike, and Stefan Nagel, "Depression Babies: Do Macroeconomic Experiences Affect Risk Taking?," *Quarterly Journal of Economics*, 126 (2011), 373–416, https://doi.org/10.1093/qje/qjq004.

McGeoch, John A., "Forgetting and the Law of Disuse," *Psychological Review*, 39 (1932), 352–370, https://doi.org/10.1037/h0069819.

Mullainathan, Sendhil, "A Memory-Based Model of Bounded Rationality," *Quarterly Journal of Economics*, 117 (2002), 735–774, https://doi.org/10.1162/003355302760193887.

Neumark, David, "Experimental Research on Labor Market Discrimination," *Journal of Economic Literature* 56 (2018), 799–866, https://doi.org/10.1257/jel.20161309.

Nickerson, Raymond S., "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises," *Review of General Psychology*, 2 (1998), 175–220, https://doi.org/10.1037/1089-2680.2.2.175.

Nosofsky, Robert M., "Similarity Scaling and Cognitive Process Models," *Annual Review of Psychology*, 43 (1992), 25–53, https://doi.org/10.1146/annurev.ps.43.020192.000325.

Pantelis, Peter C., Marieke K. van Vugt, Robert Sekuler, Hugh R. Wilson, and Michael J. Kahana, "Why Are Some People's Names Easier to Learn Than Others? The Effects of Face Similarity on Memory for Face-Name Associations," *Memory and Cognition*, 36 (2008), 1182–1195, https://doi.org/10.3758/MC.36.6.1182.

Peetz, Johanna, and Roger Buehler, "Is There a Budget Fallacy? The Role of Savings Goals in the Prediction of Personal Spending." *Personality and Social Psychology Bulletin*, 35 (2009), 1579–1591, https://doi.org/10.1177/0146167209345160.

Peetz, Johanna, Roger Buehler, Derek J. Koehler, and Ester Moher, "Bigger Not Better: Unpacking Future Expenses Inflates Spending Predictions." *Basic and Applied Social Psychology*, 37 (2015), 19–30, https://www.tandfonline.com/doi/full/10.1080/01973533.2014.973109.

Roese, Neal J., and Kathleen D. Vohs, "Hindsight Bias," *Perspectives on Psychological Science*, 7 (2012), 411–426, https://doi.org/10.1177/1745691612454303.

Ruffenach, Glenn, "How Much More Will You Spend in Retirement Than Expected? My Rule: $400 a Month," *Wall Street Journal,* March 31, 2022, https://www.wsj.com/articles/how-much-more-will-you-spend-in-retirement-than-expected-11648685983.

Sanborn, Adam N., and Nick Chater, "Bayesian Brains without Probabilities," *Trends in Cognitive Sciences*, 20 (2016), 883–893, https://doi.org/10.1016/j.tics.2016.10.003.

Schacter, Daniel L., Donna Rose Addis, and Randy L. Buckner, "Remembering the Past to Imagine the Future: The Prospective Brain," *Nature Reviews Neuroscience*, 8 (2007), 657–661, https://doi.org/10.1038/nrn2213.

Schacter, Daniel L., Donna Rose Addis, Demis Hassabis, Victoria C. Martin, R. Nathan Spreng, and Karl K. Szpunar, "The Future of Memory: Remembering, Imagining, and the Brain," *Neuron*, 76 (2012), 677–694, https://doi.org/10.1016/j.neuron.2012.11.001.

Schwartzstein, Joshua, "Selective Attention and Learning," *Journal of the European Economic Association*, 12 (2014), 1423–1452, https://doi.org/10.1111/jeea.12104.

Sherman, Steven J., David L. Hamilton, and David R. Roskos-Ewoldsen, "Attenuation of Illusory Correlation," *Personality and Social Psychology Bulletin* 15 (1989), 559–571, https://doi.org/10.1177/0146167289154009.

Shiffrin, Richard, "Memory Search," in *Models of Human Memory*, Donald A. Norman, ed. (New York: Academic Press, 1970), 375–447.
Sloman, Steven, Yuval Rottenstreich, Edward Wisniewski, Constantinos Hadjichristidis, and Craig R. Fox, "Typical versus Atypical Unpacking and Superadditive Probability Judgment," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30 (2004), 573–582, https://doi.org/10.1037/0278-7393.30.3.573.
Sims, Christopher, "Implications of Rational Inattention," *Journal of Monetary Economics*, 50 (2003), 665–690, https://doi.org/10.1016/S0304-3932(03)00029-1.
Sussman, Abigail B., and Adam L. Alter, "The Exception is the Rule: Underestimating and Uverspending on Exceptional Expenses," *Journal of Consumer Research*, 39 (2012), 800–814, https://doi.org/10.1086/665833.
Sydnor, Justin, "(Over)insuring Modest Risks," *American Economic Journal: Applied Economics*, 2 (2010), 177–199, https://doi.org/10.1257/app.2.4.177.
Tenenbaum, Joshua B., and Thomas L. Griffiths, "Generalization, Similarity, and Bayesian Inference," *Behavioral and Brain Sciences*, 24 (2001), 629–640, https://doi.org/10.1017/s0140525x01000061.
Tversky, Amos, "Features of Similarity," *Psychological Review*, 84 (1977), 327–352, https://doi.org/10.1037/0033-295X.84.4.327.
Tversky, Amos, and Daniel Kahneman, "Availability: A Heuristic for Judging Frequency and Probability," *Cognitive Psychology*, 5 (1973), 207–232, https://doi.org/10.1016/0010-0285(73)90033-9.
———, "Judgment under Uncertainty: Heuristics and Biases," *Science*, 185 (1974), 1124–1131.
———, "Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment," *Psychological Review*, 90 (1983), 293–315, https://doi.org/10.1037/0033-295X.90.4.293.
Tversky, Amos, and Derek J., Koehler "Support Theory: A Nonextensional Representation of Subjective Probability," *Psychological Review*, 101 (1994), 547–567, https://doi.org/10.1037/0033-295X.101.4.547.
Underwood, Benton J., "Interference and Forgetting," *Psychological Review*, 64 (1957), 49–60, https://doi.org/10.1037/h0044616.
Wachter, Jessica A., and Michael Jacob Kahana, "A Retrieved-Context Theory of Financial Decisions," NBER Working Paper no. 26200, 2019, https://doi.org/10.3386/w26200.
Woltz, Dan J., and Michael K. Gardner, "Semantic Priming Increases Word Frequency Judgments: Evidence for the Role of Memory Strength in Frequency Estimation," *Acta Psychologica*, 160 (2015), 152–160, https://doi.org/10.1016/j.actpsy.2015.07.012.