MEMORY AND PROBABILITY

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

September 8, 2021

Abstract. People often estimate probabilities, such as the likelihood that an insurable risk will materialize or that an Irish person has red hair, by retrieving experiences from memory. We present a model of this process based on two established regularities of selective recall: similarity and interference. The model accounts for and reconciles a variety of conflicting empirical findings, such as overestimation of unlikely events when these are cued vs. neglect of non-cued ones, the availability heuristic, the representativeness heuristic, as well as over vs. underreaction to information in different situations. The model makes new predictions on how the content of a hypothesis (not just its objective probability) affects probability assessments by shaping the ease of recall. We experimentally evaluate these predictions and find strong experimental support.

[1] Saïd Business School, University of Oxford, Harvard University, Bocconi University and IGIER, Harvard University, and Harvard University. 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.

1. 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 et al. 1978). More broadly, beliefs depend on recalled personal experiences, such as stock market crashes (Malmendier and Nagel 2011), and not just on statistical information.

It is also well known that beliefs depart from rationality in a variety of ways, which shape economic behavior. Sometimes unlikely events are overestimated, for instance 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 downside risks (Gennaioli et al. 2012). Adding to the clutter, in finance there is extensive evidence of both over and underreaction to news. Beliefs overestimate the future prospects of individual firms and the aggregate market after periods of rapid earnings growth (Bordalo et al. 2019, 2020), leading to long-run return reversals, but underestimate the impact of other news, such as earnings surprises, leading to price drifts and momentum (Chan et al. 1996; Bouchaud et al. 2019; Kwon and Tang 2021). This bewildering diversity of biases has puzzled many economists, and led sceptics to minimize the direct evidence on beliefs and embrace rationality.

We present a model of memory and belief formation, and show that it helps reconcile seemingly contradictory biases and generates new predictions. We test these predictions in two new experiments and find empirical support.

In our theory, a decision maker (DM) assesses two disjoint hypotheses H_1 and H_2 by sampling his memory database E, which contains both personal and second-hand experiences. The DM assesses the relative frequency of H_1 by his ability to retrieve experiences consistent with it from the memory database, compared to the alternative H_2. In our stylized setting, the DM relies only on sampled experiences, and does not use statistical information such as base rates. Crucially, as the DM retrieves experiences of H_i, that hypothesis acts as a memory cue, shaping selective recall according to two well-established forces in the psychology of memory: similarity and interference (Kahana 2012). Consider for instance a DM trying to assess the probability of $H_1 =$ "cause of death is flood" compared to $H_2 =$ "other causes of death". Similarity means that when thinking about H_1, experiences of floods are easier to retrieve than experiences of earthquakes, because the former are more similar to the hypothesis than the latter. Interference means that, because recall cannot be fully controlled, hypothesis-inconsistent experiences may be erroneously retrieved, especially those similar to the hypothesis itself.

Two forms of interference prove important for shaping probability judgments. The first is interference from the alternative hypothesis. When thinking about $H_1 =$ "cause of death is flood", the DM may erroneously recall a similar cause of death such as "accidental drowning", an instances of the alternative hypothesis $H_2 =$ "other causes of death". The second force is interference from what we call "irrelevant data"—that is, from experiences inconsistent with either of the hypotheses at hand. When thinking about $H_1 =$ "cause of death is flood", the DM may erroneously recall the similar event "survival in a flood," which belongs to the irrelevant "experiences other than deaths." Both types of interference inhibit the successful recall of each hypothesis. The hypothesis that faces relatively less interference is oversampled and its probability estimate is inflated. Also, because recall is random, in our model probability estimates are noisy,

even given the same memory database (Kahneman et al. 2021).

This mechanism naturally generates several well-documented biases. Interference from the alternative hypothesis yields a striking new principle: heterogeneous hypotheses, those that consist of very dissimilar events such as $H_2 =$ "other causes of death," are underestimated. Because such hypotheses are composed of events that are not similar to each other, they are hard to retrieve, which lowers their assessment. This principle accounts for biases typically explained with the availability heuristic (Tversky and Kahneman 1973). First, it explains why we tend to overestimate unlikely events when they are prompted, and to neglect them when they are not prompted. Only in the former case does similarity boost recall of these events. Second, it offers one mechanism for partition dependence in beliefs: the likelihood attributed to H_2 = "cause of death is cancer, heart attack, or any other causes of death" is higher than that attributed to H = "other causes of death" (Tversky and Koehler 1994), because the partition makes each hypothesis-cue more similar to the respective instances, and hence less subject to interference.

In turn, interference from irrelevant data illuminates biases in conditional probability assessments, many of which are attributed to the representativeness heuristic (Kahneman and Tversky 1973). Consider for instance base-rate neglect. Given the data that a person is "shy and withdrawn," it may be hard to think about the hypothesis that he is H_1 = "a farmer" because many farmers with different personalities are retrieved. These experiences are irrelevant, but interfere with the recall of shy farmers, reducing the assessment of the hypothesis. We show that this same mechanism can account for the conjunction fallacy (Tversky and Kahneman 1974, 1983).

Finally, the interaction between the two forms of interference unifies instances of over and underreaction to data in conditional assessments. When the data cue an unlikely hypothesis, the model yields overreaction through base-rate neglect. When the hypothesis is fairly likely, the

model delivers overreaction if the data is quite informative and underreaction otherwise. This can help explain why overreaction in financial markets is observed after informative histories such as rapid earnings growth but not when signals are weak, such as a positive earnings surprise. If the hypothesis is highly likely, the model predicts underreaction to data, yielding a form of aversion to extreme beliefs (Griffin and Tversky 1992).

To test the predictions of our model, we introduce 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 possesses a certain property. To do so, subjects 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 that over and under-estimation of unlikely events can be switched on and off by modulating 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 work in economics explores the role of memory in belief formation (Mullainathan 2002; Bordalo et al 2020; Wachter and Kahana 2020; Enke et al. 2020). We unify the representativeness and availability heuristics (Tversky and Kahneman 1974) by showing that – due to similarity and interference – representative experiences are more "available," or accessible, for recall. We show that this approach micro-founds and generalize previous formalizations of representativeness (Gennaioli and Shleifer 2010; Bordalo et al. 2016; Bordalo et al. 2020), as well as the overreaction mechanism in Diagnostic Expectations (Bordalo et al. 2018).

In psychology, Sanborn and Chater (2016) present a model of beliefs based on Bayesian memory sampling. Unlike this approach, we start with well-established regularities in recall,

similarity and interference, and show that they deliver systematic biases. The Minerva-DM model (Dougherty et al. 1999) features similarity-based recall and noisy encoding, but does not allow for interference. As such, it cannot account for key biases such as representativeness or the conjunction fallacy without making ad hoc ancillary assumptions.

In Billot et al. (2005), the probability of an elementary event (rather than of broad hypotheses) is estimated based on its similarity to other events in the database. They do not study judgment biases and their model generates neither the conjunction fallacy nor the disjunction effect.

A related literature examines the link between memory and beliefs from the perspective of efficient information processing (Tenenbaum and Griffiths 2001; Dasgupta et al. 2020; Azeredo da Silveira et al. 2020; Dasgupta and Gershman 2021). We instead start with well-documented regularities in recall and interference, and show how they unify a range of evidence.

Kahneman and Tversky (1979) propose a probability weighting function in which small probabilities are overestimated in lottery choice. Our model applies to the construction of subjective beliefs from memory and explains why in the real world low probability events are often overestimated when cued and underestimated otherwise.

We describe our model of similarity-based recall and probability judgments in Section 2, characterize the departures of probability estimates from statistically correct beliefs in Section 3, and present the experimental results in Sections 4 and 5. Section 6 concludes.

2. 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", etc. One feature, which we denote by f_d, indicates whether the event was lethal or not. There are superordinate features, such as f_{d+1} ="disease", f_{d+2} ="natural disaster", etc, 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 $f_{d+2} = 0$. Non-lethal 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 than 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 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). The database could also 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" vs. 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 $f_{d+2} = 1$ vs. $f_{d+2} = 0$. In the language of probability, the union of the hypotheses $H = H_1 \cup H_2$ is the sample

space over which the DM forms his subjective beliefs.[2] Appendix A3 extends the model to more than two hypotheses. We refer to the sample space as the "relevant data" for the probabilistic assessment. We refer to $\overline{H} = E \backslash H$, the set of experiences inconsistent with either hypothesis, as the "irrelevant data". Figure 1 depicts this decomposition of the memory database E.

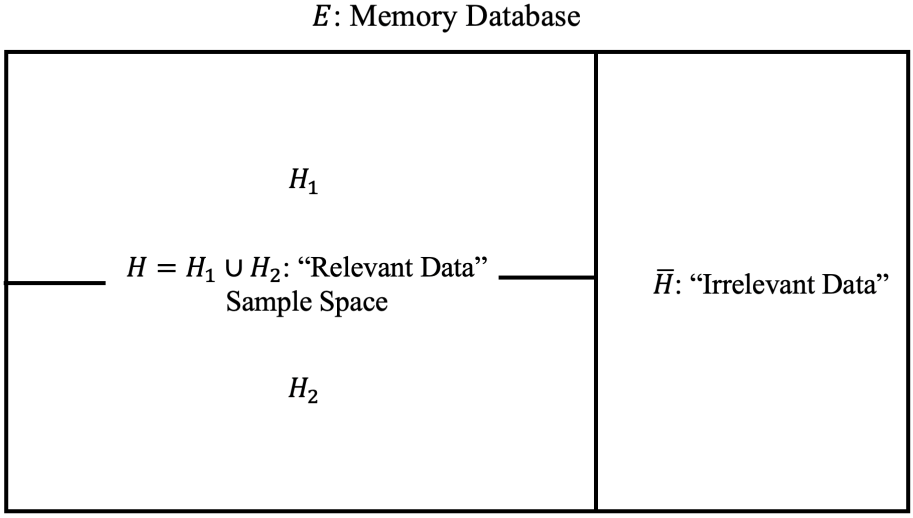


Figure 1: Memory Database and Sample Space

As we describe in detail below, the DM makes his assessment by extracting a sample from his database with replacement.[3] In a model in which sampling is only based on objective frequency, the DM draws any given experience in E with probability $1/N$. He then estimates the probability of H_i as the relative frequency of this hypothesis among all draws belonging to the relevant data H. In this model, due to the random nature of recall, the DM's beliefs would then be noisy but

[2] In a slight abuse of notation, we refer to H_i both as a given hypothesis, e.g. "cause of death is flood", and the subset of experiences in E consistent with hypothesis H_i.

[3] 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 with the fact that unique experienced events such as a stock market crash appear to persistently affect beliefs (Malmendier and Nagel 2011). Sanborn and Chater (2016) allow for more structured Bayesian approaches to frequency-based sampling, such as Markov Chain Monte Carlo. These, however, account neither for the role of similarity nor for systematic violations of consistency such as disjunction and conjunction fallacies.

average probabilistic assessments would be unbiased. We depart from this benchmark by allowing sampling to be influenced by similarity.

2.1 Similarity Based Sampling

In line with memory research (Kahana 2012; Bordalo et al. 2020), we allow recall to be influenced by similarity and interference. Each hypothesis acts as a cue for selective recall. When thinking about death from H_1 = "natural disasters", the DM recalls events similar to it: he samples, but not based on objective frequency alone.

To formalize this idea, we first define similarity. A symmetric function $S(u,v)\colon E \times E \to [0,U]$ measures the similarity between any two experiences u and v in the database. It is maximal, equal to U, when $u = v$. Similarity between two experiences increases in their number of shared features. For instance, a death from a tornado is more similar to a death from flooding than either is to death from diabetes, as the former are both caused by a natural disaster rather than an illness. Different features may be differently weighted, based on their functional importance or salience.[4] For instance, episodes of heart attack are similar to each other even if they occur in different contexts. For most of our analysis 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,

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

[4]One similarity function, parametrized by $\delta \in (0,1]$, $\{w_p\}$, is $S(u,v) = \delta^{\sum_p w_p |f_{up} - f'_{vp}|}$, where $w_p \geq 0$ for all p.

$S(A,B)$ 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.

From Equation (1) it is automatic to define $S(e,H_i)$, the similarity between a single experience e and hypothesis H_i. This is an important object to formalize cued recall. When cued by hypothesis H_i, the probability of sampling experience $e \in E$ increases in the similarity of e to H_i, and decreases with interference from other experiences in E, as we formalize below.

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

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

If similarity is constant, sampling is frequency-based—i.e., $r(e,H_i) = 1/N$. Compared to this benchmark, the numerator of (2) captures the fact that it is easier to recall e if it is more similar to the cue H_i. When thinking about deaths from $H_i =$ "natural disasters", it is relatively easy to recall deaths from floods, due to similarity. The denominator in (2) captures interference: all experiences $u \in E$ compete for retrieval, so they inhibit each other. Retrieval of e is especially inhibited by experiences that are similar to the cue. One source of interference is from other hypothesis-consistent experiences: when thinking about $H_i =$ "natural disasters", deaths from tornadoes interfere with recall of deaths from floods, because the former are also similar to H_i.

Crucially, in Equation (2) interference in the denominator also comes from hypothesis-inconsistent experiences $u \notin H_i$. There are two manifestations of this. The first is interference from the alternative hypotheses H_j: when thinking about deaths from $H_i =$ "natural disaster", the mind may retrieve experiences of deaths from other causes such as "terrorist attacks". The second is interference from "irrelevant data" $\overline{H} \equiv E\backslash H$, which are experiences outside the sample space and

hence inconsistent with either hypothesis. When thinking about deaths from H_i = "natural disasters", the mind may retrieve instances of survival in natural disasters.

Interference is due to the fact that we cannot fully control what we recall. In our model interference always comes from recalling a particular experience, but we do not claim that this process is consciously noticed. Recall failures may either manifest as "mental blanks", namely inability to recall anything when thinking about H_i, or as "intrusions", where the DM explicitly recalls experiences $u \notin H_i$ that are inconsistent with the hypothesis he is thinking about.

Similarity-based interference is well-established in memory research going back to the early 20th century (Jenkins and Dallenbach 1924; Keppel 1968; McGeoch 1932; Underwood 1957; Whitely 1927). 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 lower likelihood of retrieval and longer response times (Shiffrin 1970; Lohnas et al. 2015). In the "fan effect", Anderson (1974) shows that concepts associated with more items are more difficult to remember in response to any specific cue.[5] Our application of interference to probability estimates is new.

We assume probability judgments are formed according to the following two stage sampling process:

Assumption 2. Sampling, Interference and Counting

Stage 1: ["Train of thought for H_i"] Each hypothesis H_i cues sampling of $T \geq 1$ experiences from E according to $r(e_k, H_i)$. Denote by R_i the number of successful recalls of all $e \in H_i$.

[5] Two approaches have been proposed to account for this evidence: associative models (Anderson and Reder 1999) and inhibition models (Anderson and Spellman 1995). These models do not incorporate interference from irrelevant data. A robust phenomenon related to intrusions is that of "false memories" (Deese 1959; Roegiger and McDermott 1995). For example, recall from a list of words that are semantically related suffers intrusions from similar words not on the list, e.g. mis-remembering milk from a list that includes butter, cheese, and white (Brown et al 2000).

Stage 2: [Renormalization] The subjective probability of H_i, denoted $\hat{\pi}(H_i)$, is the share of successful counts for H_i out of all successful counts for both hypotheses:

$$\hat{\pi}(H_i) = \frac{R_i}{\sum_{j \leq 2} R_j} \tag{3}$$

Intuitively, the DM draws two random samples, one for each hypothesis. 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 above) is especially realistic when the DM is asked to assess binary hypotheses, or when the different hypotheses are prominently presented to him, which is the case in our experiments. It may be violated if the DM's task is to represent a distribution with many possible outcomes (e.g., the age distribution of deaths), because some possible outcomes may fail to come to mind, and so are not sampled.

The assumption that the DM assesses probabilities by counting "successes" in the drawn samples (Stage 2 above) is realistic in one-shot estimation problems, but may fail in repeated estimation problems, because the DM may learn about the selected nature of the recalled samples. Of course, 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 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. Future work should seek to endogenize the set of outcomes and hypotheses the DM samples and allow for learning and effort, which seem important for real world applications.[6]

[6] Our model can also be enriched by allowing: a) sampling to be influenced also by the most recently recalled item, b) the DM to count intrusions from $u \in H_j$, and c) retrieval to be driven by factors other than similarity. For instance,

2.2 Characterizing Beliefs

We examine the two stages of Assumption 2 in turn. We first derive the probability of recalling any experience $e \in H_i$ when the DM thinks about H_i. This is what psychologists call retrieval fluency of hypothesis H_i. Using Equation (2), we can show that this is given by:

$$r(H_i) = \sum_{e \in H_i} r(e, H_i) = \frac{\sum_{e \in H_i} S(e, H_i)}{\sum_{u \in H_i} S(u, H_i) + \sum_{u \in H_j} S(u, H_i) + \sum_{u \in \bar{H}} S(u, H_i)}$$

$$= \frac{\pi(H_i)}{\pi(H_i) + \frac{S(H_i, H_j)}{S(H_i, H_i)} \cdot \pi(H_j) + \frac{S(H_i, \overline{H})}{S(H_i, H_i)} \cdot \frac{\pi(\overline{H})}{\pi(H)}} \qquad (4)$$

In this equation, $\pi(H_i)$ is the true relative frequency of H_i in the relevant data H, i.e., the correct probability, $\pi(H)$ is the true probability of the relevant data compared to the entire database E, while $\pi(\overline{H})$ is the true probability of irrelevant data, namely outside of H.

The term $S(H_i, H_i)$ is the similarity of H_i to itself. We call this the "self-similarity" of H_i. It captures the homogeneity of this hypothesis, namely the extent to which its experiences share similar features. 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 $H_1 =$ "natural disaster". The "cross-similarity" terms $S(H_i, H_j)$ and $S(H_i, \overline{H})$ capture the similarities of H_i to the alternative hypothesis H_j and to irrelevant data $\overline{H}$, respectively. A death from flood in $H_1 =$ "natural disaster" is similar to a death from accidental drowning in H_2, which raises $S(H_1, H_2)$, and it is also similar to the event of surviving a flood in $\overline{H}$, which raises $S(H_1, \overline{H})$. In Equation (4), such cross-similarity terms interfere, reducing retrieval fluency of H_i.

---

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, e.g. names of celebrities are more easily remembered (Tversky and Kahneman 1973).

Define $\omega(H_i)$ as the ratio between retrieval fluency $r(H_i)$ and the frequency-based probability of drawing a member of the same hypothesis H_i, i.e. $\omega(H_i) = r(H_i)/\frac{|H_i|}{N}$.

Proposition 1 *If* $S(H_i, H_i) = S\left(H_i, H_j\right) = S\left(H_i, \overline{H}\right)$*, sampling is frequency based (*$\omega(H_i) = 1$*). If instead* $S(H_i, H_i) > \max\left[S\left(H_i, H_j\right), S\left(H_i, \overline{H}\right)\right]$ *there is oversampling of* H_i *given the cue* H_i *(*$\omega(H_i) > 1$*). The extent of oversampling* $\omega(H_i)$ *falls with:*

1. *the true frequency* $\pi(H_i)$ *of* H_i*.*
2. *the strength of interference from* H_j *and* $\overline{H}$*, measured by* $\frac{S\left(H_i, H_j\right)}{S(H_i, H_i)}$ *and* $\frac{S\left(H_i, \overline{H}\right)}{S(H_i, H_i)}$ *respectively.*

As in Equation (2), if similarity is constant, our model yields frequency-based sampling. If instead a hypothesis is more similar to itself than to the rest of the database, which is often a valid condition,[7] similarity-based recall entails oversampling. When cued by $H_i =$ "natural disasters", people scan their memories for earthquakes, floods, etc., and so oversample these events relative to their true likelihood. If all hypotheses being considered have high self-similarity, they are all oversampled.[8] Henceforth, we assume $S(H_i, H_i) > \max\left\{S\left(H_i, H_j\right), S(H_i, \overline{H})\right\}$.

Importantly, Property 1 says that oversampling is especially severe when thinking about objectively unlikely hypotheses. People rarely experience floods and earthquakes compared to heart attacks or accidents, so cueing $H_1 =$ "natural disasters" greatly boosts their retrieval.

Property 2 says that, even though similarity focuses sampling on H_i, the retrieval of H_i can be inhibited by interference, to an extent that depends on the content of the database (i.e., on its similarity structure). Retrieving many instances of $H_1 =$ "natural disasters" is difficult because

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

[8] When taking comparative statics with respect to the similarities $S\left(H_i, H_j\right), S(H_i, H_i)$, and $S(H_i, \overline{H})$, we hold the objective frequencies $\pi(H_i)$ constant.

this set contains very different events (e.g. earthquakes, floods), so $S(H_1, H_1)$ is low. Furthermore, $H_1 =$ "natural disasters" has high cross-similarities, which trigger recall of both inconsistent experiences such as "accidental drowning" and of irrelevant ones such as "survival in natural disasters". The strength of these effects is captured by $\frac{S(H_1,H_2)}{S(H_1,H_1)}$ and $\frac{S(H_1,\overline{H})}{S(H_1,H_1)}$.

Having explored the determinants of retrieval fluency, we now turn to the probabilistic assessment $\hat{\pi}(H_i)$ in Equation (3). In light of Assumption 2, the number of successes in recalling each hypothesis H_i follows a binomial distribution: $R_i \sim Bin\big(T, r(H_i)\big)$, where $r(H_i)$ is given by Equation (4). This implies that beliefs $\hat{\pi}(H_i)$ are stochastic, characterized as follows.

Proposition 2 *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 tends to attach a higher probability to hypotheses that have relatively high retrieval fluency, as in Tversky and Kahneman's (1973) availability heuristic. In light of Proposition 1, then, average beliefs can be distorted. In particular, a hypothesis facing stronger interference from its alternative or from irrelevant data will be underestimated. This intuition unifies the account of many probability judgments and yields new predictions.

Equation (6) shows that the retrieval fluency of different hypotheses also shapes the variability of beliefs. We defer the analysis of this aspect to Section 5. Here it suffices to say that, in general, when two hypotheses are easy to recall—i.e., when both $r(H_1)$ and $r(H_2)$ are high—

the variability of beliefs declines, because the DM benefits from a larger sample size. In Section 5 we test this and other predictions about noise in probabilistic assessments.

3. Judgment Biases

We next examine biases in probabilistic assessments. Section 3.1 deals with interference from the alternative hypothesis, which yields several biases related to the availability heuristic. Section 3.2 incorporates interference from irrelevant data, and shows that it accounts for biases related to the representativeness heuristic. Section 3.3 shows that these two forces can unify over and underreaction of beliefs to data.

3.1 Biases due to Interference from the Alternative Hypothesis

To focus here on interference from the alternative hypothesis, we restrict our attention 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.

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 3 *The estimate $\hat{\pi}(H_1)$ increases in the objective frequency $\pi(H_1)$. Overestimation, i.e., $\hat{\pi}(H_1) > \pi(H_1)$, occurs if and only if $\pi(H_1) < \pi^*$, 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. When thinking about H_1= "floods", the DM selectively retrieves deaths due to floods. When thinking about H_2= "other causes", he selectively retrieves other causes of death. Oversampling occurs for both hypotheses but, as shown in Proposition 1, it favours the less likely one, which would otherwise be less sampled. This causes insensitivity to true frequencies.

Similarity yields a fundamental new implication: it is the content of events in the database, and not only their frequency, that shapes the sensitivity to frequency.

Corollary 1. *As the events in H_1 become more homogeneous, i.e. $S(H_1, H_1)$ 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.*

In our model, as H_1 becomes more self-similar, it is easier to recall. As a result, it is less likely that, when thinking about it, recall 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 has one important implication: cued unlikely events are prone to overestimation not only due to their low probability, but also 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 fairly similar to each other. By contrast, the alternative $H_2 =$ "causes other than flood" is very heterogeneous, and contains causes of deaths similar to floods, like other natural disasters or accidental drownings. Thus, when thinking about of $H_2 =$ "causes other than flood", recall may slip back to H_1 = "flood", hindering the assessment of the former.

This mechanism has an additional prediction: unless directly cued, unlikely events are prone to be under-sampled and neglected. Consider the frequency with which a DM thinking about H_2 recalls experiences in its subset $H_{21} \subset H_2$, for example $H_2 =$ "causes other than flood" and

H_{21} = "tornado". By Equation (2), the likelihood of sampling experiences of H_{21} = "tornado" depends on how similar tornados are to the broader cue, H_2 = "causes other than floods," captured by $S(H_{21}, H_2)$. Denote the other experiences in H_2 by $H_{22} = H_2 - H_{21}$.

Corollary 2 *If $S(H_{21}, H_2) < S(H_{22}, H_2)$, then H_{21} is undersampled relative to its frequency in H_2. If the two subsets are equally self-similar, $S(H_{21}, H_{21}) = S(H_{22}, H_{22})$, then H_{21} is undersampled, namely $S(H_{21}, H_2) < S(H_{22}, H_2)$, if and only if it is less frequent, or $|H_{21}| < |H_{22}|$.*

A non-cued rare event is neglected when it is dissimilar from the cued hypothesis to which it belongs. This depends, among other things, on the event's frequency: the rarer is H_{21}, the more atypical it is of the cued hypothesis H_2 and hence the more dissimilar it is from H_2. Thus, when thinking about H_2 = "causes other than flood", we may recall the likely "heart attack", but not the unlikely "tornado". This is in stark contrast to Corollary 1, which predicts that the DM overestimates rare events when prompted to think about them. Similarly, when thinking about H = "murders in Michigan", people tend to neglect the non-cued subset H' = "murders in Detroit", because Detroit is dissimilar to the rest of Michigan. In contrast, people tend to overestimate H' = "murders in Detroit" when explicitly cued (Kahneman and Frederick 2002).

This principle also implies that, for given database E, the description of a hypothesis can affect assessments. 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). Our model accounts for this phenomenon: partitioning a hypothesis into more specific sub-events increases its overall self-similarity, reducing interference. To see this, consider again a DM assessing the same hypothesis H_2, but now H_2 is explicitly partitioned into H_{21} and H_{22}. For the purpose of

Proposition 4, assume 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, $S(H_{21}, H_1) = S(H_{22}, H_1)$.[9] We obtain:

Proposition 4 *Partitioning the alternative hypothesis H_2 into H_{21} and H_{22} is equivalent to changing the self-similarity of H_2 while leaving its true frequency unchanged. The self-similarity of H_2 increases if and only if:*

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

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

The assessment of H_1 = "flood" is reduced when its alternative is specified as H_{21} = "natural causes" and H_{22} = "non-natural causes other than flood", compared to when it is specified as H_2 = "causes other than flood". Cueing H_{21} and H_{22} fosters retrieval when thinking about alternatives to flood, which reduces the assessment of H_1 = "flood".

We summarize the above results using Figure 2, which plots the probability estimate $\hat{\pi}(H_1)$ against the truth $\pi(H_1)$ in three conditions, which correspond to Experiment 1 we present later.

[9] These conditions nest three hypotheses (H_1, H_{21}, H_{22}) in the binary hypotheses case, connecting to Proposition 2.

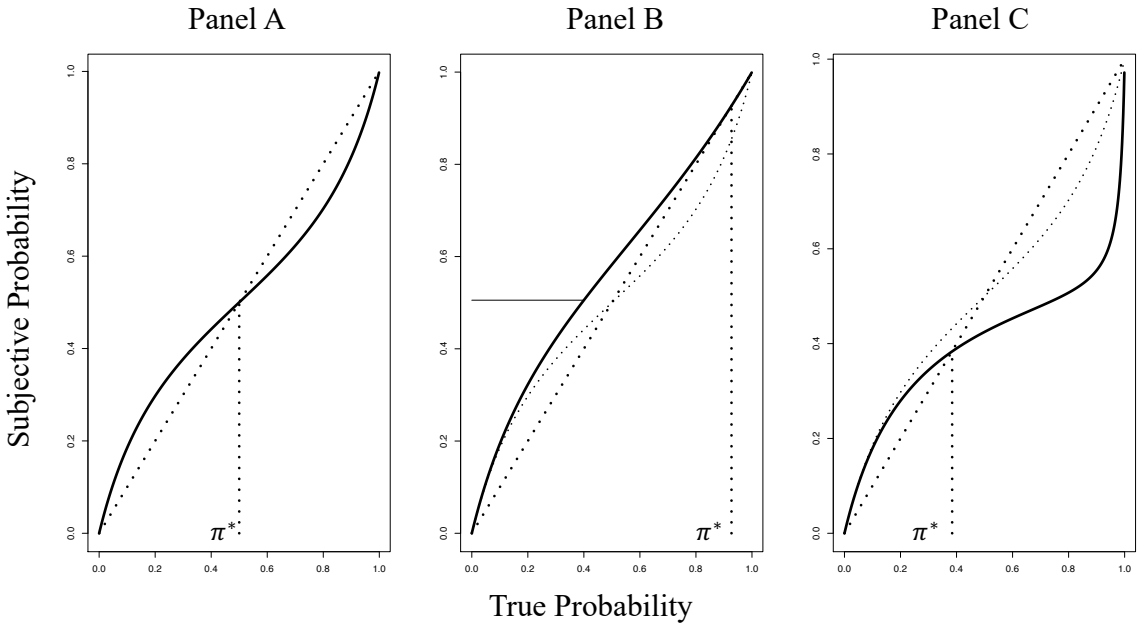


Figure 2: Over- and underestimation of hypotheses, and the role of similarity

Notes: This figure shows the estimate $\hat{\pi}(H_1)$ against the truth $\pi(H_1)$, holding fixed $S(H_1, H_1)$ and varying $S(H_2, H_2)$. π^* is the threshold where beliefs switch from under to overestimation. In Panel A, H_1 and H_2 are equally self-similar: $S(H_1, H_1) = S(H_2, H_2)$. In Panel B, H_2 is highly heterogenous: $S(H_2, H_2) \ll S(H_1, H_1)$. In Panel C, H_2 is much more self-similar than H_1: $S(H_2, H_2) \gg S(H_1, H_1)$.

Panel A shows the case of equally homogeneous hypotheses, $S(H_1, H_1) = S(H_2, H_2)$. Here $\pi^* = 0.5$ and probabilities of H_1 and H_2 are "smeared" toward 50:50. Kahneman and Tversky's (1979) probability weighting function features insensitivity to frequency in the domain of objective probabilities in lottery choice. Recent work provides foundations for this function based on the salience of lottery payoff (Bordalo et al. 2012), noisy perception of numerical probabilities (Khaw et al. 2020, Frydman and Jin 2020), and cognitive uncertainty (Enke and Graeber 2020). Our model applies instead to the construction of subjective probabilities of cued hypotheses from experience,

when objective probabilities are not given. Selective recall is a central mechanism in our approach, and a source of perceived uncertainty.[10]

Panel B illustrates our key prediction that "content matters". It plots the assessment $\hat{\pi}(H_1)$ when the self-similarity of hypothesis H_2 is reduced relative to that of H_1, or $S(H_2, H_2) < S(H_1, H_1)$. Compared to Panel A, the subjective probability curve shifts up, boosting the assessment of H_1 at any given true frequency $\pi(H_1)$. This effect can have dramatic consequences.

In a striking experiment, Tversky and Kahneman (1983) asked a group of subjects to assess the share of $H_1 =$ "words ending with _n_" in a text. They then asked another group of subjects to assess the probability of $H_1' =$ "words ending with _ing" in a text. Remarkably, subjects attached a lower probability to H_1 than to H_1', despite the lower objective frequency of the latter. Our model accounts for this phenomenon as a shift from Panel A to Panel B. Intuitively, instances of $H_1' =$ "words ending with _ing" share many features, such as being gerunds, denoting similar activities, etc, which makes it easy to bring many examples to mind. In contrast, $H_1 =$ "words ending with _n_" additionally includes many words which do not share these features (and which often do not share many features with each other). This reduction in self-similarity makes it harder to recall words in H_1, even though it is a superset of $H_1{}'$.

Finally, Panel C plots $\hat{\pi}(H_1)$ for the case in which the alternative hypothesis H_2 is partitioned into subsets H_{21} and H_{22}, that is, when the self-similarity of hypothesis H_2 is increased relative to that of H_1 (Proposition 4). Compared to Panel A, the estimation curve for H_1 drops, potentially turning its overestimation into underestimation. Several famous studies show such "partition dependence" of beliefs, whereby the probability assigned to an event decreases if its

[10] In the appendix, we provide evidence that recall is strongly correlated with a measure of subjective uncertainty.

alternative is partitioned (Benjamin 2019).[11] In Tversky and Koehler's (1994) "Support Theory", this phenomenon arises because people evaluate events using a sub-additive "support function". In our model, partition dependence comes from similarity in recall.[12]

With similarity-based sampling, the DM evaluates a hypothesis by selectively retrieving instances of it, but in doing so finds it hard not to think about the alternative hypothesis. Such interference reconciles well known biases such as smearing of probability judgments toward 50:50, underestimation of rare events that are not cued, and availability effects in which the normatively irrelevant content of hypotheses and their description affect judgments.

3.2 Biases due to Interference from Irrelevant Experiences

So far, interference from irrelevant data was ruled out by our assumption that the database E coincides with the relevant data $H = H_1 \cup H_2$. Suppose, however, that the DM must now condition H_1 and H_2 on data D, which identifies a subset $D \subset H$. Now irrelevant experiences in $\overline{D} \equiv H \backslash D$ can interfere provided that they are similar to those in D. We next show that interference from irrelevant data is crucial to understanding the effects of data-provision and conditional expectations. In particular, it produces overreaction effects typically explained using the representativenesss heuristic.

To gain intuition, consider our running example. The DM assesses the relative probability of different causes of death H_i, but he is now restricted to the set of young people, namely D = "young." For concreteness, the DM assesses deaths by H_1 = "accident" vs. H_2 = "sickness" among

[11] For example, Fischhoff et al. (1978) famously show that when assessing the cause of a car's failure to start, mechanics judges "ignition" more likely when alternative causes were partitioned into "ignition", "fuel", "other".
[12] In contrast to the disjunction fallacy, death by "pneumonia, diabetes, cirrhosis or any other disease" is estimated to be less likely than death by "any disease" (Sloman et al. 2004). This is consistent with a natural extension of our model in which atypical cues such as "cirrhosis" focus attention on a narrow subset, which interferes with the retrieval of more common diseases. A similar pattern occurs in free recall tasks (Slamecka 1968; Sanborn and Chater 2016).

the D = "young". To do so, he separately thinks about the events in $H_1 \cap D$ = "accidents among the young" and $H_2 \cap D$ = "sickness among the young". The retrieval fluencies $r(H_1 \cap D)$ and $r(H_2 \cap D)$ of these finer hypotheses are still captured by Equation (4), with the only difference that now the relevant data is D instead of H and the irrelevant data is $\overline{D}$ instead of $\overline{H}$. These recall fluencies produce, following Assumption 2, a *conditional* probability estimate $\hat{\pi}(H_i|D)$ that we compare to the true conditional probability $\pi(H_i|D)$.

Crucially, as the DM thinks about causes of death among $D =$ "young", causes of death for $\overline{D} =$ "older people" may interfere. To visualize this possibility, Figure 3 depicts the database E, where the size of each region roughly corresponds to true frequencies.

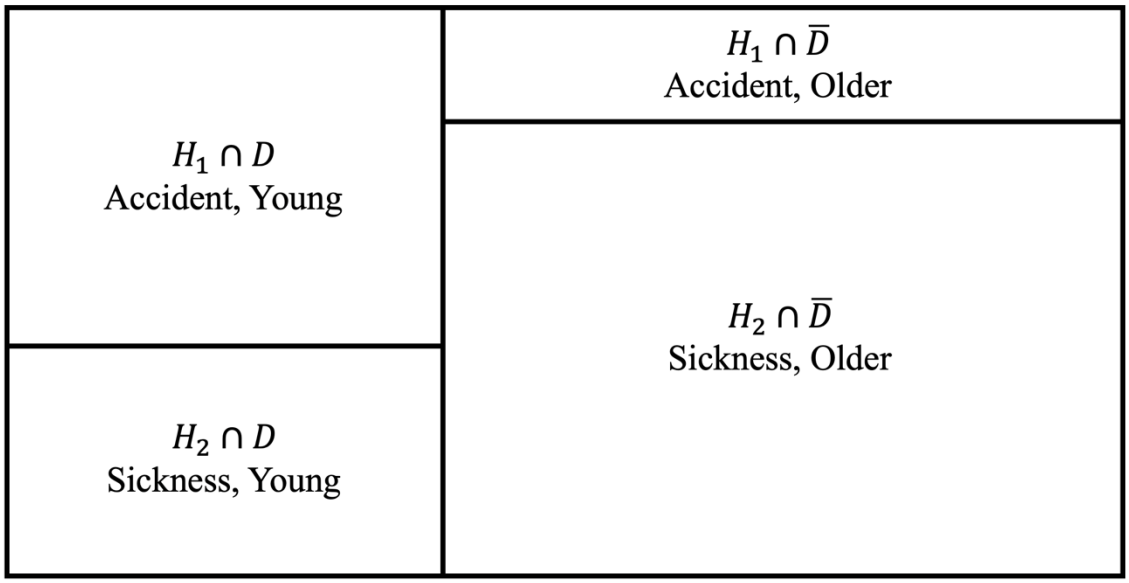


Figure 3: Visualizing conditional assessments

When thinking about $H_1 \cap D =$ "accident among the young", two forms of interference are at work. First, in line with our prior analysis, memories of young people dying from sickness (i.e. from $H_2 \cap D$) intrude, due to similarity among young people. Second, and crucially, experiences of older people dying from accidents can also intrude from $H_1 \cap \overline{D}$. This is

interference from the irrelevant data $\overline{D}$ = "older", which occurs due to the similarity along the H_1 = "accident" dimension. The same holds when thinking about $H_2 \cap D$ = "sickness among the young", which faces intrusion from "accidents among the young" and from irrelevant experiences of "sickness among the older".

The strength of intrusion from irrelevant data, here deaths among older people, depends on the similarity between a cause of death and the irrelevant data. A key driver of this similarity is the joint distribution of features: the greater the share of older people dying from sickness, the more similar is sickness (relative to accidents) to deaths of older people, and hence the stronger the interference from the latter. As we show next, this effect can significantly influence beliefs.

Formally, suppose that there are only two features, in our case the cause of death (accident vs. sickness) and age (young vs 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.

Proposition 5. *For $\delta < 1$, the DM overestimates the probability of H_1 conditional on D, namely $\hat{\pi}(H_1|D) > \pi(H_1|D)$, if and only if:*

$$\pi(H_1|D)\pi(D) + \delta\pi\left(H_1\middle|\overline{D}\right)\pi\left(\overline{D}\right) < \frac{\pi(D) + \delta\pi\left(\overline{D}\right)}{2}. \quad (9)$$

Overestimation is more likely when the true conditional probability $\pi(H_1|D)$ is low. This is due to interference from the alternative hypothesis which, as shown in Proposition 3, favours H_1 when it is relatively unlikely. Now, however, the conditional hypothesis is over-estimated also if its frequency in the irrelevant data, $\pi\left(H_1\middle|\overline{D}\right)$, is low. In this case, hypothesis H_1 is less similar to the irrelevant data $\overline{D}$ than H_2. It thus faces less interference, which promotes its overestimation.

Consider how this works in Figure 3. Because H_1= "accident" is a common cause of death for the young ($\pi(H_1|D)$ high), interference from the alternative hypothesis promotes its underestimation. At the same time, however, most older people die from sickness, not from accidents ($\pi(H_1|\overline{D})$ low). As a result, when thinking about young people dying from sickness, many instances of older people dying from sickness intrude. This can cause overestimation of H_1 = "accident" for the young, even if it is a likely cause of death for this group.

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, i.e. D = "activist". Some subjects are asked the probability that she is currently a H_1 = "bank teller", others that she is a H_1' = "feminist bank teller". The striking finding is that feminist bank teller is rated likelier than bank teller.

According to Proposition 5, this occurs for two reasons. First and foremost, H_1 = "bank teller" is much more similar to the group of $\overline{D}$ = "non-activists" than H_1' = "feminist bank teller", so it faces much more interference from irrelevant data. The similarity gap is due to the fact that, among non-activists, there are many more bank tellers than feminist bank tellers, $\pi\left(H_1|\overline{D}\right) > \pi\left(H_1'|\overline{D}\right)$. Second, H_1 = "bank teller" is likelier than H_1' = "feminist bank teller", $\pi(H_1|D) > \pi(H_1'|D)$, which also promotes underestimation of H_1 relative to H_1'. Interference from both the irrelevant data and the alternative hypothesis contribute to creating this conjunction fallacy.[13]

More broadly, Tversky and Kahneman (1983) propose the following definition of representativeness: "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."

[13] In the previous "availability" experiments on the share of words ending with "_n_" vs "_ing", there is also a conjunction fallacy, but it is due to interference from the alternative hypothesis.

GS (2010) propose that the conditional probability $\pi(H|D)$ is overestimated if the likelihood ratio $\frac{\pi(H|D)}{\pi(H|\overline{D})}$ is high. In this formula, the irrelevant data $\overline{D}$ captures the "reference class" in Tversky and Kahneman's definition, with respect to which the "relative frequency" is computed. This formalization is at the heart of Bordalo et al. (2018)'s model of diagnostic expectations.

Intrusion from irrelevant data offers a foundation for the reference class $\overline{D}$ and for the role of $\pi(H|\overline{D})$ in shaping the assessment of H conditional on D. When $\pi\left(H|\overline{D}\right)$ is low, the hypothesis H is dissimilar to the irrelevant data $\overline{D}$, which makes it easy to imagine the occurrence of consistent experiences in $H \cap D$. It takes a small step to say that in this case "H is representative of D". Our model, through Equation (9), says that in this case $\pi(H|D)$ is overestimated. This is in line with the GS (2010) model, which also yields overestimation when $\pi\left(H|\overline{D}\right)$ is low. [14] Our approach microfounds representativeness effects from memory, but also reconciles them with Kahneman and Tversky's (1973) broad intuition that similarity judgments affect beliefs.

One advantage of our approach is to identify limits to representativeness. When the true conditional probability $\pi(H|D)$ is high, intrusion from the alternative hypothesis is strong. Thus, according to Equation (9), the conditional probability may be underestimated, even if $\pi(H|\overline{D})$ is low. This mechanism throws new light on the conflicting evidence of over and underreaction.

3.3 Underreaction and overreaction to data

Survey and field evidence documents conflicting distortions in the reaction of beliefs to data. There is evidence that people over-estimate the probability of events in light of data

[14] Bordalo et al. (2016) argue that people overestimate the share of red-haired people in Ireland because "red hair" is much more common there than in the rest of the world. In the current model, when thinking about the hypothesis H_1 = "red hair" vs H_2 = "dark hair" conditional on the data D = "Irish", hypothesis H_2 faces strong interference from many dark haired people from outside Ireland (belonging to $\overline{D}$). H_1 is then overestimated because it faces much less interference from $\overline{D}$ and because it is unlikely (there are few red-haired people even in Ireland).

informative about them, a finding typically explained by the representativeness heuristic (e.g., Kahneman and Tversky 1973). But there is also evidence of under-estimation in similar situations, which is often explained with inattention (Sims 2003; Coibion and Gorodnichenko 2012; Gabaix 2018). Memory, and in particular the two forms of interference studied above, can help reconcile this conflicting evidence, yielding conditions as to when either phenomenon should occur.

To connect our model to this evidence, we define over and underreaction in terms of the level of beliefs. We say that the DM overreacts to data D if: 1) D is objectively informative about a certain hypothesis H_i, i.e. $\pi(H_i|D) > \pi(H_i)$, and 2) the DM overestimates that hypothesis, i.e. $\hat{\pi}(H_i|D) > \pi(H_i|D)$. 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 in cases of stereotypes (e.g. red haired Irish), or the Linda problem.[15]

To see our model's implications, consider overreaction to data informative about H_1. We start with requirement 2). Using Equation (9), Figure 4 separates the regions of over and underestimation of $\pi(H_1|D)$ in the space of true conditional probabilities $\left(\pi(H_1|D), \pi\left(H_1|\overline{D}\right)\right)$.

[15] When priors are measured and likelihoods are available, under and overreaction is often defined in terms of *sensitivities* rather than *levels*, as is done in Grether (1980), i.e. in terms of the difference between the elicited prior and posterior beliefs. For time series forecasts, a popular measure is the correlation between future forecast errors and current forecast revisions (Coibion and Gorodnichenko (2012) and Bordalo et al. (2019)).

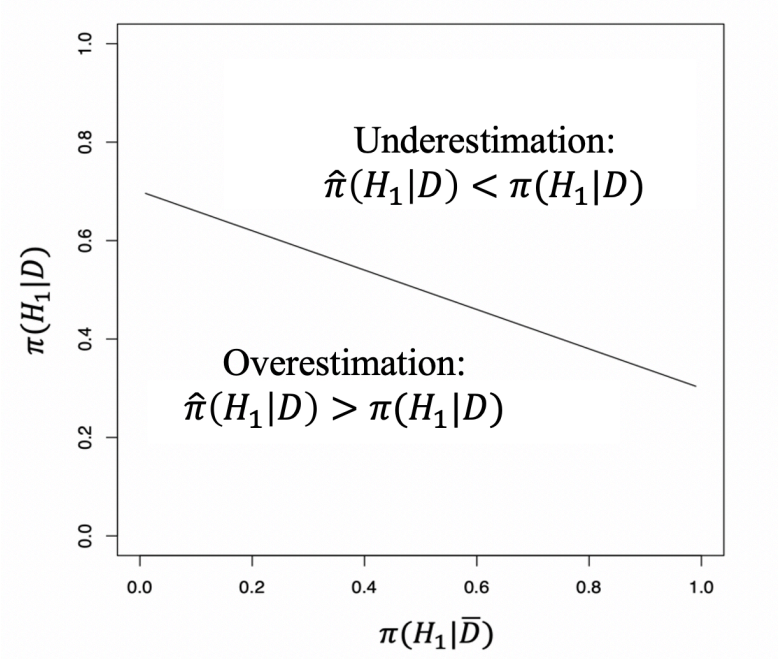


Figure 4: Condition for underestimation and overestimation of conditional beliefs

Notes: The DM overestimates the conditional probability of H_1 given D, $\hat{\pi}(H_1|D) > \pi(H_1|D)$, in the region below the solid line. He instead underestimates it in the region above the solid line.

As $\pi(H_1|D)$ increases, we move vertically from over to underestimation of the hypothesis H_1 conditional on D. This is due to interference from the alternative hypothesis, which lowers the assessment of more likely hypotheses. As $\pi(H_1|\overline{D})$ increases, we move horizontally from over-to-underestimation of the conditional hypothesis. This is driven by interference from irrelevant data.

Consider next requirement 2). In Figure 4, the region in which the data is informative about H_1 corresponds to the area above the 45° line, where $\pi(H_1|D) > \pi\left(H_1|\overline{D}\right)$.[16] In this region, the DM "overreacts" to D if he overestimates $\hat{\pi}(H_1|D)$, while he "underreacts" to D if he underestimates $\hat{\pi}(H_1|D)$. Corollary 3 and Figure 4 describe these regions.

[16] This is equivalent to $\pi(H_1|D) > \pi(H_1)$, as $\pi(H_1) = \pi(D) \cdot \pi(H_1|D) + \left(1 - \pi(D)\right) \cdot \pi(H_1|\overline{D})$ is a convex combination of $\pi(H_1|D)$ and $\pi(H_1|\overline{D})$.

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

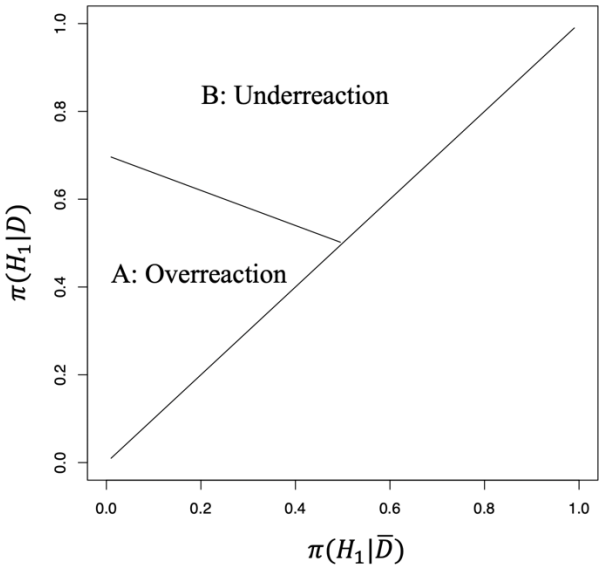


Figure 5. Condition for underreaction and overreaction to data

Notes: This figure depicts the region of $\left(\pi(H_1|D),\ \pi(H_2|\ \overline{D})\right)$ where the agent overreacts or underreacts to data D, where D is diagnostic of H ($\pi(H_1|D) > \pi(H_1|\overline{D})$). Region A corresponds to overreaction, region B to underreaction.

Whether over or underreaction prevails depends on the balance between interference from irrelevant data, which promotes overreaction when $\pi(H_1|\overline{D})$ is low, and interference from alternative hypotheses, which promotes underreaction when $\pi(H_1|D)$ is high.

Consider the different cases, starting with the situation in which overreaction can occur, namely $\pi(H_1|D) < \bar{\pi}$. The strongest case of overreaction occurs in the lower left corner of region A in Figure 5, in which both forms of interference are low for H_1, namely $\pi(H_1|D)$ and $\pi(H_1|\overline{D})$ are both low. This case sheds light on 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 a librarian rather than a farmer, even though farming is a much more common occupation, especially among men. Overreaction occurs not only because librarians are relatively rare ($\pi(H_1|D)$ is low), but also because many librarians are shy and have a passion for detail, while comparatively few farmers do ($\pi(H_1|\overline{D})$ is also low).

Overreaction is not a mere manifestation of base rate neglect, however, as it can arise even if the hypothesis is quite likely, potentially even if $\pi(H_1|D) > 0.5$. This occurs in the upper part of region A. For instance, an investor may overestimate the probability that a firm has H_1 = "strong fundamentals" if it has experienced D = "rapid earnings growth" or "rapid growth" for short. This occurs provided firms with strong fundamentals mostly exhibit rapid growth, namely $\pi(H_1|\overline{D})$ is low. In this case, when thinking about firms with strong fundamentals, it is easy to recall examples that have experienced rapid growth. At the same time, when thinking about firms with H_2 = "weak fundamentals," it is hard to recall examples that have experienced rapid growth, because weak firms often produce $\overline{D}$ = "not rapid growth". The DM overreacts because H_1, although likely, faces much less interference from irrelevant data than H_2. As mentioned above, this logic for overreaction offers a micro-foundation for diagnostic expectations.

Critically, however, the model also highlights two limits to overreaction. The first occurs when interference from irrelevant data $\pi(H_1|\overline{D})$ is sufficiently high, for instance due to a weak signal. Suppose that the DM judges the fundamentals of a firm and the data is not "rapid growth" but rather D = "positive earnings surprise". Clearly, firms with strong fundamentals may well have negative earnings surprises, so $\pi(H_1|\overline{D})$ is higher than in the previous example. When thinking about these firms, many examples of strong-fundamentals firms with negative earnings surprises come to mind and interfere, potentially causing underestimation of H_1 and hence

underreaction.[17] Thus, if D points to a fairly likely hypothesis, beliefs underreact to weakly diagnostic data and overreact to sufficiently diagnostic data.[18]

The second limit to overreaction occurs when the true probability of H_1 is very high, $\pi(H_1|D) > \bar{\pi}$, due to extremely informative data or a concentrated prior. This occurs in the upper left portion of region B of Figure 5. Here, H_1 suffers from very strong interference from H_2, which causes underreaction. This effect is consistent with the evidence for general conservatism or aversion to extreme beliefs (Griffin and Tversky 1992; Benjamin 2019).

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 the evidence on under and over-reaction.

4. Experiments

We now 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 studies the additional role of interference from irrelevant data. Subjects in both experiments first go through a controlled set of experiences in which they see a series of images, and then 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 1968; Enke and Graeber 2020) or ask hypothetical questions about naturalistic situations (Kahneman and Tversky

[17] For a given objective probability $\pi(H_1|D)$, underreaction occurs when the signal is sufficiently weak that in Figure 4 we move horizontally from the upper part of A to the lower part of B.

[18] In our model the relative strength of intrusion from irrelevant data is stronger when the data D is rare, namely when $\pi(D)$ is small compared to $\pi(\overline{D})$, see Equation (4) and (9). This fits the intuition of stronger overreaction for small populations, such as those of the Irish or of firms exhibiting strong growth.

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 to assess whether recall and probability estimates go hand in hand.

Subjects were recruited among Bocconi University undergraduates enrolled in the experimental economics email list. They could participate in both experiments, which occurred four months apart from each other (recruiting for each was conducted separately). Participants completed the experiments remotely due to Covid restrictions. They earned a 4 euro Amazon gift card, and could earn a 2 euro bonus if their answer to one randomly chosen question was within 5 percentage points of truth. If the randomly chosen question was a free recall task, each correctly/incorrectly recalled word increased/decreased subjects' chance of winning the 2 euro bonus by 10 percentage point, and each incorrect word reduced their chances by 10 percentage points (of course, their chances could not go below zero or above 1).[19] Experiments were pre-registered, including hypotheses and sample sizes, on the AEA RCT Registry, with ID AEARCTR-0006676. Appendix B provides more details about both surveys.

4.1 Experiment 1: Testing Interference from the Alternative Hypothesis

Participants see 40 words, some of which are animals and some are not, in a random order. 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? ____%"

[19] The bonus provides sharp and easily understood incentives, as compared to other schemes such as binarized scoring rules, which have been shown to distort truth telling (Danz et al., 2020).

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 a fourth treatment, we change the formulation of the question in a way we describe below. In all treatments, all exhibited words are relevant to answering the question. Thus, there is no interference from irrelevant data.

We next describe three predictions from the model and four treatments that allow us to test them. The predictions are:

Prediction 1: Memory creates a tendency to overestimate cued unlikely hypotheses, and overestimation is stronger for rarer hypotheses. (Proposition 3)

Prediction 2: Holding objective frequencies constant, the assessed probability of a hypothesis decreases when its alternative is more heterogeneous/less self-similar. (Corollary 1)

Prediction 3: Holding objective frequencies constant, the assessed probability of a hypothesis increases if its alternative is partitioned into two more self-similar subsets. (Proposition 4)

The treatments to test these predictions 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.[20]

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%.[21]

[20] These words were chosen using a random word generator, eliminating words that we deemed too similar to each other (e.g., Mayor, Elected, Town).

[21] In the free recall task, participants are also asked to list up to 15 examples each of animals, men's names, and women's names (so, three total recall tasks).

These experimental treatments are summarized in Table 1 below.[22]

Comparing T1 and T2 offers a test for Prediction 1: we should expect overestimation $\hat{\pi}(animal)$ in both treatments but especially in T1, when animals are objectively rarer. By comparing T2 and T3 we can test Prediction 2: compared to T2, $\hat{\pi}(animal)$ should be higher in T3, because the alternative hypothesis (i.e., non-animals) is very heterogeneous. By comparing T4 and T2 we can test Prediction 3: because the alternative to animals in T4 is split into two more self-similar sub-hypotheses (men's names and women's names), $\hat{\pi}(animal)$ should be lower in T4 than in T2. Lastly, 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, but may rather reflect the fact that both outcomes are products of retrieval fluency.

Treatment	Sample Size	Distribution of Images	Examples	Elicited Belief
T1	$N = 244$	20% Animals, 80% Names	Lion, John, Moose, Rat, Margaret, Deer, Edward, Nancy, Wolf…	P(Animal) vs P(Other)
T2	$N = 244$	40% Animals, 60% Names	Paul, John, Moose, Rat, Margaret, Deer, Laura, Nancy, Edward…	P(Animal) vs P(Other)

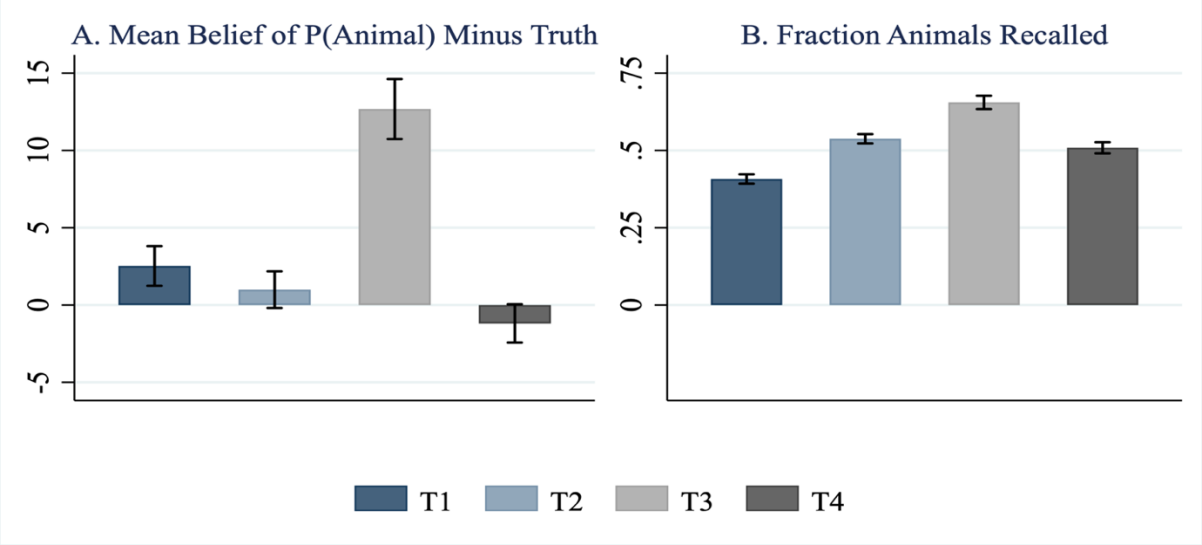
[22] In addition to treatments T1-T4, we ran a treatment T5, where we replaced women's names in T1 with ocean animals (e.g., "Shark", "Starfish", "Dolphin", etc.). Participants are then asked the probability of "Land Animals" (in T1, all animals are land animals) and "Anything else." In the recall task, 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 Appendix describes this issue and the results from this treatment in greater detail. Our main results in Section 4.5 and 5 hold qualitatively if we include T5.

T3	$N = 241$	40% Animals, 60% Heterogeneous	Lion, Sled, Moose, Rat, Pure, Deer, Half, Good, Wolf…	P(Animal) vs P(Other)
T4	$N = 234$	40% Animals, 60% Names	Lion, John, Moose, Rat, Margaret, Deer, Edward, Nancy, Wolf…	P(Animal) vs P(Men) vs P(Women)

Table 1: Treatments in Experiment 1

4.2 Experiment 1 Results

Figure 6 shows the treatment effects. Panel A reports the over-or-underestimation of $\hat{\pi}(animal)$ compared to the truth. Panel B reports the share of animals among all recalled examples.[23]



[23]Throughout the analysis to follow, 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.

Figure 6: Results from Experiment 1

Notes: 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.

Consistent with Prediction 1, there is a tendency to overestimate $\hat{\pi}(animal)$, especially in T1 where animals are only 20% of words: overestimation of animals (that is, mean belief minus truth) is 2.5 percentage points (pp) in T1 and 1 pp in T2, and only the former is significantly different from zero at conventional levels ($p < 0.01$ and $p = 0.10$, respectively). Furthermore, the overestimation in T1 is marginally statistically significantly different from that in T2 ($p = 0.09$).

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 pp in T2 to 12.7 pp in T3 ($p < 0.01$). There are two key messages here. First, overestimation depends not only on actual frequency but also on how self-similar the alternative hypothesis is. Second, this effect can dominate attenuation to 50:50: in T3, $\hat{\pi}(animal)$ overshoots 50% ($p = 0.01$). Interference in recall emerges as a powerful force in probabilistic assessments.

Finally, when partitioning "non-animals" into the finer sub-hypotheses "men's names" and "women's names" in T4, the assessment $\hat{\pi}(animal)$ falls by 2.1 percentage points compared to T2 ($p = 0.013$). In T4, $\hat{\pi}(animal)$ is underestimated relative to the truth ($p = 0.060$), while it is overestimated in T2 and T3 (though only statistically significantly so in the latter). 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 Panel B of Figure 6, in each case mirror those on beliefs. In T2, on average 54% of correctly recalled words are animals. Significantly fewer (40%)

of recalled words are animals in T1 ($p < 0.01$), where there are objectively fewer animal words. In T3, where non-animals are heterogeneous words rather than people's names, this share is 66%, significantly higher than in T2 ($p < 0.01$). Finally, in T4, where men' and women's names are separated out, significantly fewer recalled words are animals (50%) than in T2 ($p = 0.02$).[24]

4.3 Experiment 2: Interference from Irrelevant Data

Participants are shown 40 images, which consist of words and numbers, 20 of which are orange, and 20 are blue.[25] In all the treatments, participants are asked the following question:

> "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%.[26]

In this experiment, a subset of experiences – blue words and blue numbers – are irrelevant for assessing the hypotheses, which concern 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. This is interference from irrelevant data. 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.

[24]While 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.
[25] All words in this experiment are related to time (e.g., "Second", "Week", "Duration", etc.), although we do not ask about word categories so this does not matter for the analysis.
[26] The slider begins with no default, so that participants have to click somewhere on the slider and then move the drag-able icon (that then appears where they first click) to indicate their answer.

As in Proposition 5, when assessing $\hat{\pi}(w|o)$, interference from the irrelevant blue words increases in the true probability that a blue image is a word, $\pi(w|b)$, while smearing towards 50:50 increases as the true probability that the orange image is a word, $\pi(w|o)$, gets further from 0.5.

To identify interference from irrelevant data, in what we call a *Neutral* treatment, we fix at 50% the share of orange images that are words, $\pi(w|o) = 0.5$. We then vary the distribution of irrelevant blue images across three sub-treatments. In sub-treatment NL, no blue image is a word, $\pi(w|b) = 0$, so all blue images are numbers. Here the hypothesis "word|orange" faces no interference from irrelevant data, while its alternative "number|orange" faces maximal interference. In sub-treatment NM, 50% of blue images are words, $\pi(w|b) = 0.5$. Here the two hypotheses face balanced interference from irrelevant data. In sub-treatment NH, all blue images are words, $\pi(w|b) = 1$, so "word|orange" faces maximal interference from irrelevant data while "number|orange" faces no such interference. Proposition 5 then predicts that $\hat{\pi}(w|o)$ should be overestimated in NL, correctly estimated in NM, and underestimated in NH.

In the next set of treatments, we increase the correct answer $\pi(w|o)$ above 50%. Interference from the alternative hypothesis should then become stronger, promoting underestimation of $\hat{\pi}(w|o)$ for any given strength of intrusion from irrelevant data.

In the *Intermediate* treatment, 55% of orange images are words, $\pi(w|o) = 0.55$. We then modulate interference from irrelevant data across two extreme sub-treatments: in sub-treatment IL, no blue image is a word ($\pi(w|b) = 0$), so interference from irrelevant data may cause overestimation of $\hat{\pi}(w|o)$ despite it being the likely hypothesis. In sub-treatment IH, all blue images are words ($\pi(w|b) = 1$), so $\hat{\pi}(w|o)$ should be underestimated.

Finally, in the *Common* treatment, 70% of orange images are words, creating even stronger smearing toward 50:50. We again vary interference from irrelevant data: in CL no blue image is a

word ($\pi(w|b) = 0$), in CM, 30% of blue images are words ($\pi(w|b) = 0.3$), and in CH, all blue images are words ($\pi(w|b) = 1$).

Table 2 summarizes these treatments.[27]

Table 2: Treatments in Experiment 2

Treatment	Distribution	Distribution of Irrelevant Data	Sample Sizes	Elicited Belief
Neutral	50% Orange Words, 50% Orange Numbers	NL: 0% Blue Words	*N* = 147	P(Word | Orange)
		NM: 50% Blue Words	*N* = 146	
		NH: 100% Blue Words	*N* = 151	
Intermediate	55% Orange Words, 45% Orange Numbers	IL: 0% Blue Words	*N* = 158	P(Word | Orange)
		IH: 100% Blue Words	*N* = 154	
Common	70% Orange Words, 30% Orange Numbers	CL: 0% Blue Words	*N* = 154	P(Word | Orange)
		CM: 30% Blue Words	*N* = 149	
		CH: 100% Blue Words	*N* = 144	

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

4.4 Experiment 2 Results

Panel A of Figure 7 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.

[27] We did not include an *IM* treatment due to sample size limitations.

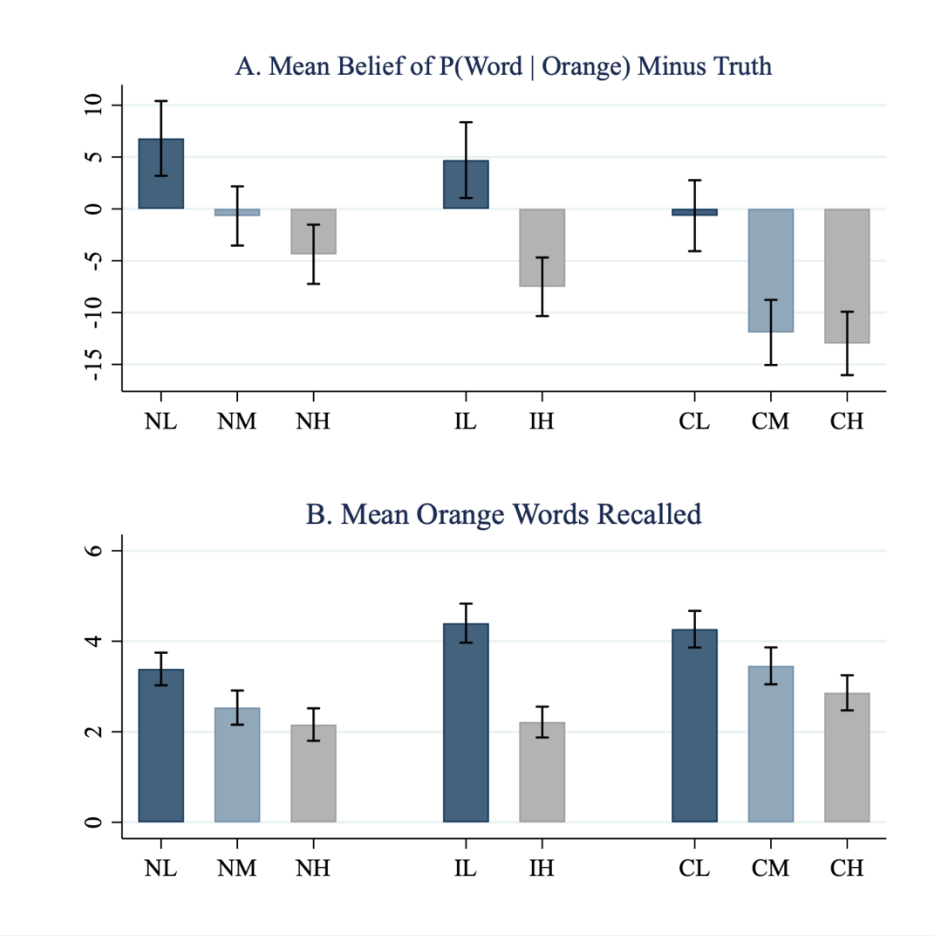


Figure 7: Testing Prediction 4

Notes: 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 words. In the *H* Treatments, all blue images are numbers. 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.

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 ($p < 0.01$ in each case). When normatively irrelevant blue words are more numerous, interference in recall of

orange words is stronger. Recall data in Panel B supports this mechanism: subjects recall fewer correct orange words when $\pi(w|b)$ is higher.[28]

Again in line with our predictions, overestimation of $\hat{\pi}(w|o)$ arises only if orange words are sufficiently rare. 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 $\pi(w|b) = 0$. Only in the *Neutral* and *Common* treatments interference from the alternative hypothesis is sufficiently weak that $\hat{\pi}(w|o)$ is overestimated if it faces no interference from irrelevant data, $\pi(w|b) = 0$.[29]

In sum, Experiment 2 shows that, consistent with Proposition 5, the 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 in the database. However, as the hypothesis becomes too likely (see the *Common* treatments), it is again underestimated because interference from the alternative hypothesis becomes very strong.

These results also shed light on over/underreaction to data. The data $D =$ "orange" is weakly informative about H_1= "words", formally, $\pi(w|o) \geq \pi(w|b)$, in treatments NL, NM, IL, CL and CM. Because in these treatments $\pi(w|o) \geq 0.5$, we are in the upper part of Region A of Figure 5. Consistent with our model, the experimental results indicate that overreaction prevails when "word|orange" faces no interference from irrelevant data (in NL, IL, CL), underreaction prevails when the hypothesis is very likely and the signal is weaker (in CM), while beliefs are close to Bayesian in NM. In the treatments NH, IH, CH, by contrast, the data $D =$ "orange" is

[28] 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 sub-treatments, 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 here we focus on categories for which it is difficult to incorrectly list a word as being in the wrong category.

[29] 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.

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, in these treatments we see overreaction for the rare H_2= "number", namely $\hat{\pi}(n|o)$ is overestimated or equivalently $\hat{\pi}(w|o)$ is underestimated.

Overall, and consistent with Corollary 3, Experiment 2 shows that the varying strengths of interference from the alternative hypothesis and from irrelevant data cause a switch from over to underreaction to informative data, and these effects show up in actual recall patterns. Regularities in selective memory unify different biases in probability judgments.

5. Noise

The model also yields novel predictions about "noise", that is, variability in beliefs given the same experiences and the same question (Kahneman et al. 2021). As shown in Proposition 2, noise naturally arises from sampling variation in recall. We now study noise in the experimental data and compare it with our model's predictions. The results are for the most part correlational but provide further, if suggestive, evidence that memory underlies the causal effects of Section 4.

Figure 8 pools the T2 and T3 treatments of Experiment 1, where the true distribution of words includes 40% animals and 60% non-animals.[30] 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 recall are both dependent on retrieval fluency (Panel C).[31] Our recall measurement makes

[30] Results look similar if we include the other treatments. We exclude the other treatments here because either the true frequency or the question asked differs from the T2 treatment.

[31] The simplest interpretation of Figure 8 is that the experiences randomly recalled when making probability judgments later cue recall of the same experiences when asked to list exemplars. If so, probability assessments and free recall

it unlikely that this correlation is mechanical: subjects are (separately) incentivized to correctly recall up to 15 exemplars of each hypothesis, and anchoring to previously stated probability estimates would require subjects to jointly adjust their specific recollections of every hypothesis.

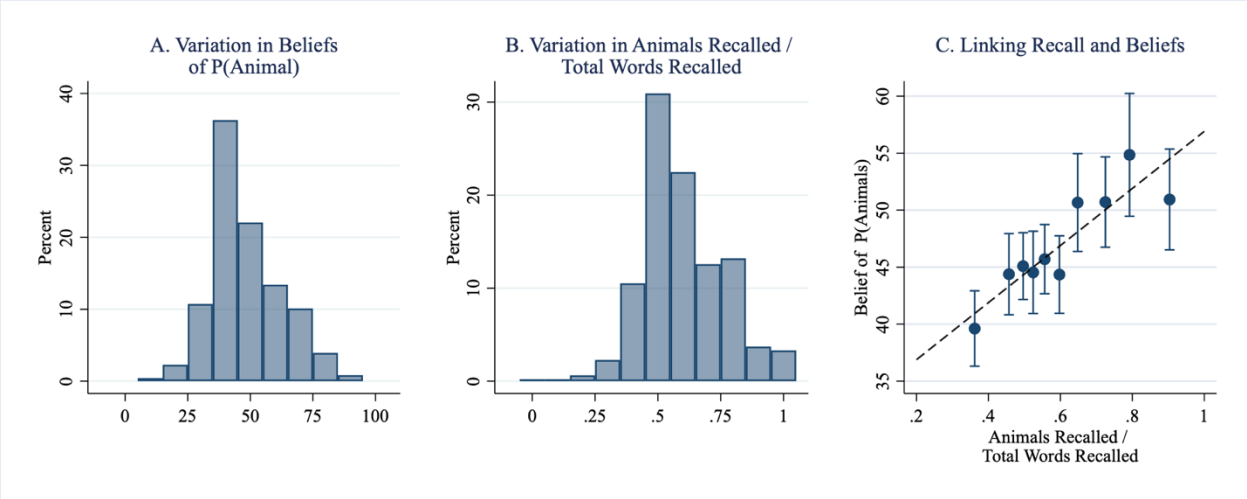


Figure 8: The relationship between recall of examples and beliefs

Notes: 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.

Figure 8 is consistent with the connection between variability in recall and noise in judgments. Our model makes further and finer predictions about this relationship.

Proposition 6:

1. *The variance of* $\hat{\pi}(H_i)$ *decreases in* T*, the total recall attempts for each hypothesis.*
2. *If* $\hat{\pi}(H_i) \in \left[\sqrt{5} - 2{,}3 - \sqrt{5}\right]$*, the variance of* $\hat{\pi}(H_i)$ *decreases in* $S(H_i, H_i)$ *and* $S\left(H_j, H_j\right)$.

are correlated because the distribution of recalled experiences are similar in the two cases. Alternatively, there may be individual-level differences in the subjective similarity of objects to a given cue, making it easier for some subjects to recall certain exemplars for a given category than for others, which will also be reflected in probability assessments.

First, the level of noise is negatively correlated with recall: as the number of recall attempts T gets larger, the share of successes for each hypothesis converges to its expected value, resulting in less heterogeneity. Second, if beliefs are not too far from 50:50, making hypotheses more self-similar should reduce noise in assessment. This is intuitive: when hypotheses are more self-similar, their recall is more successful, which increases sample size, in turn reducing the variability of beliefs.[32] Conversely, noise increases if hypotheses are less self-similar.

To assess the first prediction, Figure 9 shows the cross-sectional relationship between the total number of words recalled and the conditional variance of the fraction of recalled words that are animals (Panel A) and of beliefs (Panel B), pooling the four treatments of Experiment 1. We view the total number of words recalled by a subject as a proxy for T (so subjects are allowed to have different sample lengths). Consistent with Proposition 6, we see a negative and statistically significant relationship in both cases ($p < 0.01$ for both). More sampling yields less variability in relative recall of hypotheses and, correspondingly, in beliefs.[33,34] In Appendix B, we show that another proxy for T—the time subjects spend answering the beliefs question—is also negatively correlated with the variance both of the fraction of recalled animals and of beliefs.

[32] The fact that average odds need to be sufficiently close to 50:50 highlights an important non-monotonicity in the relationship between heterogeneity and noise: if the heterogeneity of a particular hypothesis is sufficiently high, its likelihood will converge to zero, eventually reducing the level of noise. This is however not the case in our experiment, in which probability judgments are far from corners.

[33] To test the statistical significance for both dependent variables in Figure 9, we estimate by maximum likelihood a model in which the mean changes linearly and the conditional variance changes multiplicatively in total number of recalled words. The dashed curves in Figure 9 show predicted conditional variances from these estimates.

[34] The negative relationship shown in Panel A of Figure 9 could in principle be partly mechanical—if a subject correctly recalls all 30 words, it must be the case that 40% of them were animals. Note, however, that the recall task occurs *after* the beliefs question, so subjects cannot consult their list of recalled words when forming their probabilistic judgments. The negative relationship in Panel B is therefore not mechanical, and in any case we see a negative relationship in Panel A even for participants who recall only a small fraction of the words.

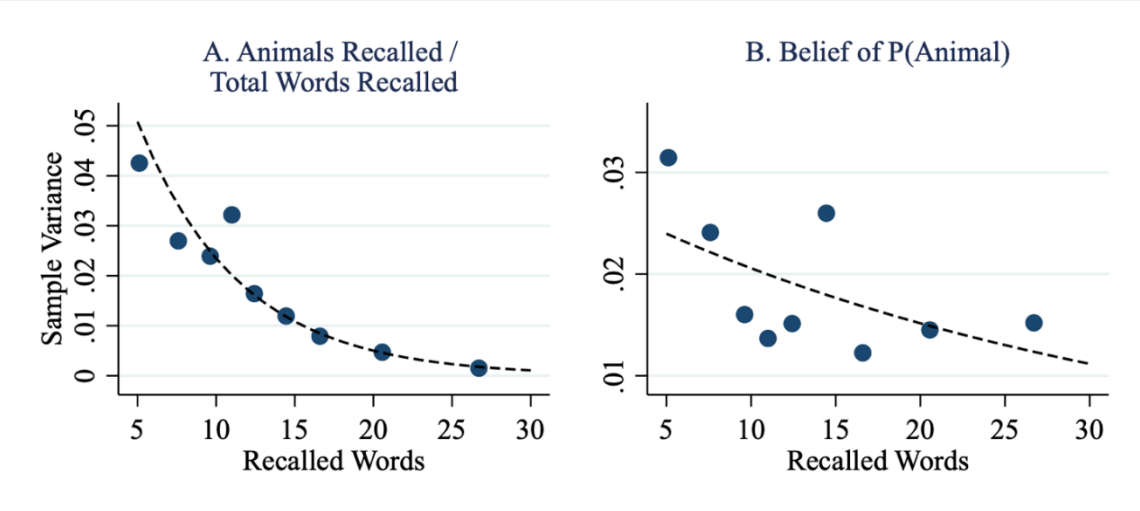


Figure 9: The Relationship between Recall and Noise

Notes: Panel A shows the sample variance of the fraction of recalled words that are animals (y-axis), conditional on decile of total number of recalled words (x-axis). Panel B shows the sample variance of beliefs about the probability animals, also conditional on decile of total number of recalled words (x-axis). The dotted lines show predicted values from maximum likelihood estimates of a model where the mean of the dependent variable varies linearly and the conditional variance multiplicatively in the total number of recalled words. Both panels restrict the data to *T1* and *T3*, where the true frequency of animals is 40%.

We can also exploit cross-treatment differences in similarity to test the second prediction in Proposition 6. Recall that the hypothesis alternative to animals is much less self-similar in T3 than in T2 (where non-animals are all people's names). Proposition 6 holds that decreasing retrieval fluency for non-animals should increase variance in the relative recall of animals (i.e., recall of animals divided by recall of all words) and, therefore, raise the variance in beliefs. Figure 10 shows that, indeed, there is greater variance in the fraction of recalled words that are animals in T3 than in T2 ($p < 0.01$) and correspondingly, greater variance in beliefs ($p < 0.01$).

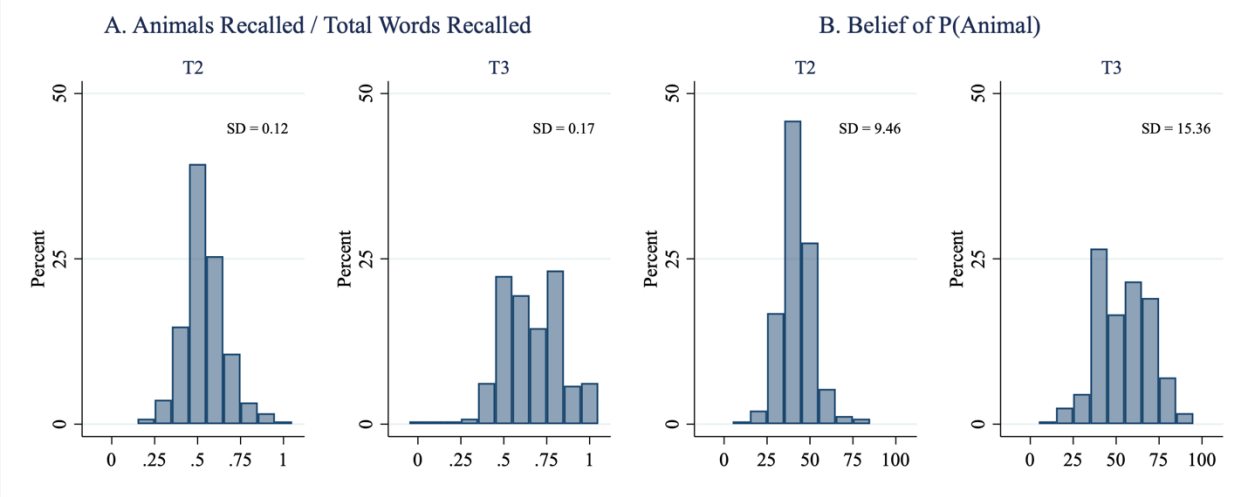


Figure 10: Treatment effects on Noise

Notes: Panel A shows the distribution of the fraction of recalled words that are animals in T2 (leftmost graph) and in T3 (second graph to the left). Panel B shows the distribution of beliefs about the probability of animals in the same treatments.

In sum, selective recall seems to be a promising avenue to think about systematic biases in probability judgments and belief heterogeneity.

6. Conclusion

We have presented a model of memory-based probability judgments, whose two principal ingredients are 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 both hypotheses and data. Similarity helps retrieve relevant experiences, but also invites interference by allowing for experiences that are inconsistent with the hypothesis at hand (but sufficiently similar to it) to come to mind. The central new insight is that a hypothesis is underestimated when, compared to its alternative, it is more vulnerable to interference –because it is either 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 Kahneman and Tversky (1974), overestimation of the probabilities of unlikely hypotheses, conjunction and disjunction fallacies in experimental data, as well as under- and over-reaction 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 found strong supportive evidence.

The analysis in this paper 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 decision maker, 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 et al. 1997; Brown et al. 2000). People combine past experiences with simulated ones (Kahneman and Miller 1986; Schachter 2009), with the ease of simulation also driven by perceived similarity (Woltz and Gardner 2015). In this way, memory shapes forecasts. In addition, individuals often have both statistical and experiential information, such as in the literature on the description-experience gap in risky choice (Erev and Hertwig 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 also 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 decision maker's perceived similarity depends on the range of experiences encountered in the past, 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 paper can be applied to many problems, including consumer choice, advertising, persuasion, political positioning, product branding, and many others.

REFERENCES

Anderson, M. C., and Spellman, B. A. (1995). On the status of inhibitory mechanisms in cognition: memory retrieval as a model case. *Psychological Review*, *102*(1), 68-100.

Anderson, J. R., and Reder, L. M. (1999). The fan effect: new results and new theories. *Journal of Experimental Psychology: General*, *128*(2), 186-197.

Azeredo da Silveira, R., and Woodford, M. (2019). Noisy memory and over-reaction to news. In *American Economic Review: Papers and Proceedings*, *109*, 557-561.

Barberis, N., Shleifer, A., and Vishny, R. (1998). A model of investor sentiment. *Journal of Financial Economics*, *49*(3), 307-343.

Barseghyan, L., Molinari, F., O'Donoghue, T., and Teitelbaum, J. C. (2013). The nature of risk preferences: evidence from insurance choices. *American Economic Review*, *103*(6), 2499-2529.

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

Billot, A., Gilboa, I., Samet, D., and Schmeidler, D. (2005). Probabilities as similarity-weighted frequencies. *Econometrica, 73(4), 1125-1136.*

Bordalo, P., Coffman, K., Gennaioli, N., Schwerter, F., and Shleifer, A. (2020). Memory and representativeness. *Psychological Review*, *128*(1), 71-85.

Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2016). Stereotypes. *Quarterly Journal of Economics*, *131*(4), 1753-1794.

Bordalo, P., Gennaioli, N., Ma, Y., and Shleifer, A. (2020). Overreaction in macroeconomic expectations. *American Economic Review*, *110*(9), 2748-2782.

Bordalo, P., Gennaioli, N., LaPorta, R., and Shleifer, A. (2019). Diagnostic expectations and stock returns. *Journal of Finance*, *74*(6), 2839-2874.

Bordalo, P., Gennaioli, N., LaPorta, R., and Shleifer, A. (2020). Expectations of fundamentals and stock market puzzles, w27283. National Bureau of Economic Research.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2012). Salience theory of choice under risk. *Quarterly Journal of Economics*, *127*(3), 1243-1285.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2018). Diagnostic expectations and credit cycles. *Journal of Finance*, *73*(1), 199-227.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2020). Memory, attention, and choice. *Quarterly Journal of Economics*, *135*(3), 1399-1442.

Bouchaud, J. P., Krueger, P., Landier, A., and Thesmar, D. (2019). Sticky expectations and the profitability anomaly. *Journal of Finance*, *74*(2), 639-674.

Brown, N. R., Buchanan, L., and Cabeza, R. (2000). Estimating the frequency of nonevents: the role of recollection failure in false recognition. *Psychonomic Bulletin and Review*, *7*(4), 684-691.

Chan, L. K., Jegadeesh, N., and Lakonishok, J. (1996). Momentum strategies. *Journal of Finance*, *51*(5), 1681-1713.

Chiappori, P. A., Salanié, B., Salanié, F., and Gandhi, A. (2019). From aggregate betting data to individual risk preferences. *Econometrica*, *87*(1), 1-36.

Clancy, K., Bartolomeo, J., Richardson, D., and Wellford, C. (1981). Sentence decision-making: the logic of sentence decisions and the extent and sources of sentence disparity. *Journal of Criminal Law and Criminology*, *72*(2), 524-554.

Coibion, O., and Gorodnichenko, Y. (2012). What can survey forecasts tell us about information rigidities? *Journal of Political Economy*, *120*(1), 116-159.

Danz, D., Vesterlund, L., and Wilson, A. J. (2020). Belief elicitation: Limiting truth telling with information on incentives, w27327. National Bureau of Economic Research.

Dasgupta, I., Schulz, E., Tenenbaum, J. B., and Gershman, S. J. (2020). A theory of learning to infer. *Psychological Review*, *127*(3), 412-441.

Dasgupta, I., and Gershman, S. J. (2021). Memory as a computational resource. *Trends in Cognitive Sciences. 25(3*), 240-251.

Deese, J. (1959). Influence of inter-item associative strength upon immediate free recall. *Psychological Reports*, *5*(3), 305-312.

Dougherty, M. R., Gettys, C. F., and Thomas, R. P. (1997). The role of mental simulation in judgments of likelihood. *Organizational Behavior and Human Decision Processes*, *70*(2), 135-148.

Dougherty, M. R., Gettys, C. F., and Ogden, E. E. (1999). MINERVA-DM: a memory processes model for judgments of likelihood. *Psychological Review*, *106*(1), 180-209.

Edwards, W. (1982). Conservatism in human information processing. In D. Kahneman, P. Slovic, and A. Tversky (Eds.), *Judgment under Uncertainty: Heuristics and Biases* (pp. 359-369). Cambridge: Cambridge University Press.

Enke, B., and Graeber, T. (2019). Cognitive uncertainty, w26518. National Bureau of Economic Research.

Enke, B., Schwerter, F., and Zimmermann, F. (2020). Associative memory and belief formation, w26664. National Bureau of Economic Research.

Hertwig, R., and Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences*, *13*(12), 517-523.

Fischhoff, B., Slovic, P., and Lichtenstein, S. (1978). Fault trees: sensitivity of estimated failure probabilities to problem representation. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(2), 330-344.

Frydman, C., and Jin, L. J. (2020). Efficient coding and risky choice. *Quarterly Journal of Economics.* Forthcoming.

Gabaix, X. (2019). Behavioral inattention. *Handbook of Behavioral Economics: Applications and Foundations 1*, 261-343.

Gennaioli, N., and Shleifer, A. (2010). What comes to mind. *Quarterly Journal of Economics*, *125*(4), 1399-1433.

Gennaioli, N., Shleifer, A., and Vishny, R. (2012). Neglected risks, financial innovation, and financial fragility. *Journal of Financial Economics*, *104*(3), 452-468.

Grether, D. M. (1980). Bayes rule as a descriptive model: the representativeness heuristic. *Quarterly Journal of Economics*, *95*(3), 537-557.

Griffin, D., and Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, *24*(3), 411-435.

Jenkins, J. G., and Dallenbach, K. M. (1924). Obliviscence during sleep and waking. *American Journal of Psychology*, *35*(4), 605-612.

Kahana, M. J. (2012). *Foundations of human memory*. OUP USA.

Kahneman, D., and Fredrick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and biases: The psychology of intuitive judgment, 49,* 81.

Kahneman, D., Fredrickson, B. L., Schreiber, C. A., and Redelmeier, D. A. (1993). When more pain is preferred to less: adding a better end. *Psychological Science*, *4*(6), 401-405.

Kahneman, D., and Miller, D. T. (1986). Norm theory: comparing reality to its alternatives. *Psychological Review*, *93*(2), 136-153.

Kahneman, D., Sibony, O., and Sunstein, C. R. (2021). *Noise: a flaw in human judgment*. Little, Brown.

Kahneman, D., and Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237-251.

Kahneman, D., and Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, *47*(2), 263-292.

Keppel, G. (1968). Verbal learning and memory. *Annual Review of Psychology*, *19*(1), 169-202.

Khaw, M. W., Li, Z., and Woodford, M. (2020). Cognitive imprecision and small-stakes risk aversion. *The Review of Economic Studies*. Retrieved from https://doi-org.ezp-prod1.hul.harvard.edu/10.1093/restud/rdaa044

Kwon, S. Y., and Tang, J. (2020). Reactions to news and reasoning by exemplars. *Available at SSRN*.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., and Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(6), 551-578.

Lohnas, L. J., Polyn, S. M., and Kahana, M. J. (2015). Expanding the scope of memory search: modeling intralist and interlist effects in free recall. *Psychological Review*, *122*(2), 337-363.

Malmendier, U., and Nagel, S. (2011). Depression babies: do macroeconomic experiences affect risk taking?. *Quarterly Journal of Economics*, *126*(1), 373-416.

McGeoch, J. A. (1932). Forgetting and the law of disuse. *Psychological Review*, *39*(4), 352-370.

Mullainathan, S. (2002). A memory-based model of bounded rationality. *Quarterly Journal of Economics*, *117*(3), 735-774.

Mullainathan, S., Schwartzstein, J., and Shleifer, A. (2008). Coarse thinking and persuasion. *Quarterly Journal of Economics*, *123*(2), 577-619.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175-220.

Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual review of Psychology*, *43*(1), 25-53.

Pantelis, P. C., Van Vugt, M. K., Sekuler, R., Wilson, H. R., and Kahana, M. J. (2008). 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*(6), 1182-1195.

Prelec, D. (1998). The probability weighting function. *Econometrica*, *66*(3), 497-527.

Roediger, H. L., and McDermott, K. B. (1995). Creating false memories: remembering words not presented in lists. *Journal of experimental psychology: Learning, Memory, and Cognition*, *21*(4), 803-814.

Sanborn, A. N., and Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*(12), 883-893.

Schwartzstein, J. (2014). Selective attention and learning. *Journal of the European Economic Association*, *12*(6), 1423-1452.

Schwartzstein, J., and Sunderam, A. (2021). Using models to persuade. *American Economic Review*, *111*(1), 276-323.

Shiffrin, R. M. (1970). Memory search. In *Models of Human Memory*, 375-447.

Slamecka, N. J. (1968). An examination of trace storage in free recall. *Journal of Experimental Psychology*, *76*(4), 504-513.

Sloman, S., Rottenstreich, Y., Wisniewski, E., Hadjichristidis, C., and Fox, C. R. (2004). Typical versus atypical unpacking and superadditive probability judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(3), 573-582.

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, *50*(3), 665-690.

Sydnor, J. (2010). (Over) insuring modest risks. *American Economic Journal: Applied Economics*, *2*(4), 177-199.

Tenenbaum, J. B., and Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629-640.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327-352.

Tversky, A., and Kahneman, D. (1973). Availability: a heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207-232.

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, *185*(4157), 1124-1131.

Tversky, A., and Kahneman, D. (1983). Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293-315.

Tversky, A., and Koehler, D. J. (1994). Support theory: a nonextensional representation of subjective probability. *Psychological Review*, *101*(4), 547-567.

Underwood, B. J. (1957). Interference and forgetting. *Psychological Review*, *64*(1), 49-60.

Wachter, J. A., and Kahana, M. J. (2019). A retrieved-context theory of financial decisions, w26200. National Bureau of Economic Research.

Whitely, P. L. (1927). The dependence of learning and recall upon prior intellectual activities. *Journal of Experimental Psychology*, *10*(6), 489-508.

Woltz, D. J., and Gardner, M. K. (2015). Semantic priming increases word frequency judgments: evidence for the role of memory strength in frequency estimation. *Acta Psychologica*, *160*, 152-160.

Wu, G., and Gonzalez, R. (1996). Curvature of the probability weighting function. *Management Science*, *42*(12), 1676-1690.

Proofs

Proposition 1 Let $r(H_i) = \frac{S(H_i,H_i)\cdot|H_i|}{S(H_i,H_i)\cdot|H_i|+S(H_i,H_j)\cdot|H_j|+S(H_i,\overline{D})\cdot|\overline{D}|}$ be the recall fluency for hypothesis $i = 1,2$. Then:

$$\omega(H_i) \equiv r(H_i)\cdot\frac{N}{|H_i|} = \frac{1}{\left[\pi(H_i)+\frac{S(H_i,H_j)}{S(H_i,H_i)}\cdot\pi(H_j)\right]\pi(D)+\frac{S(H_i,\overline{D})}{S(H_i,H_i)}\pi(\overline{D})},$$

where $\pi(H_i) = 1-\pi(H_j)$ is the true probability of hypothesis i in D, $\pi(D)$ is the relevant share of the database and $\pi(\overline{D}) = 1-\pi(D)$ the irrelevant one. Property 1 immediately follows from the fact that $S(H_i,H_j) < S(H_i,H_i)$. Property 2 immediately follows by inspection.

Proposition 2 Let $r(H_i)$ be the recall fluency of hypothesis $i = 1,2$. The sampling process implies that, with samples of size T, the number of successes N_i in recalling hypothesis $i = 1,2$ follows a binomial $Bin(T, r(H_i))$. Then, by the central limit theorem, one can take the following asymptotic approximations $\frac{N_i - Tr(H_i)}{\sqrt{T}} = z_i \sim N\left(0, r(H_i)(1-r(H_i))\right)$. Then, noting that the odds in favour of H_1 are equal to N_1/N_2 we see that these follow:

$$\frac{N_1}{N_2} \sim \frac{r(H_1)+\frac{z_1}{\sqrt{T}}}{r(H_2)+\frac{z_2}{\sqrt{T}}} \approx \frac{r(H_1)}{r(H_2)} - \frac{r(H_1)}{r(H_2)^2}\cdot\sqrt{\frac{1}{T}}z_2 + \frac{1}{r(H_2)}\cdot\sqrt{\frac{1}{T}}z_1 + O\left(\frac{1}{T}\right).$$

If T is large enough, given the distribution of z_i, the following normal approximation holds:

$$\frac{N_1}{N_2} \to N\left(\frac{r(H_1)}{r(H_2)}, \frac{1}{T}\cdot\left(\frac{r(H_1)^2}{r(H_2)^4}\cdot r(H_2)(1-r(H_2)) + \frac{1}{r(H_2)^2}\cdot r(H_1)(1-r(H_1))\right)\right)$$

which can be written in the form of Proposition 2.

Proposition 3 First, we prove that $\frac{\hat{\pi}(H_1)}{\hat{\pi}(H_2)}$ is monotonically increasing in $\pi(H_1)$: note that $\frac{\hat{\pi}(H_1)}{\hat{\pi}(H_2)} = \frac{\pi(H_1)}{\pi(H_2)}\cdot\frac{\pi(H_2)+\frac{S(H_1,H_2)}{S(H_2,H_2)}\pi(H_1)}{\pi(H_1)+\frac{S(H_1,H_2)}{S(H_1,H_1)}\pi(H_2)}$, with $\pi(H_2) = 1-\pi(H_1)$. Taking the derivative of $\frac{\hat{\pi}(H_1)}{\hat{\pi}(H_2)}$ with respect to $\pi(H_1)$ simply yields:

$$\frac{d\frac{\hat{\pi}(H_1)}{\hat{\pi}(H_2)}}{d\pi(H_1)} \propto \frac{S(H_1,H_2)}{S(H_2,H_2)}\pi(H_1)^2 + \frac{S(H_1,H_2)}{S(H_1,H_1)}\cdot(1-\pi(H_1))\cdot\left(\left(2\cdot\frac{S(H_1,H_2)}{S(H_2,H_2)}-1\right)\cdot\pi(H_1)+1\right)$$

The first term is clearly positive. As $0 \le \pi(H_1) \le 1$, it suffices to show:

$$\left(2\cdot\frac{S(H_1,H_2)}{S(H_2,H_2)}-1\right)\cdot\pi(H_1)+1 \ge 0$$

This is trivially true as $\frac{S(H_1,H_2)}{S(H_2,H_2)} \ge 0$: $\left(2\cdot\frac{S(H_1,H_2)}{S(H_2,H_2)}-1\right)\cdot\pi(H_1)+1 \ge -\pi(H_1)+1 \ge 0$.

Second, the above expression shows that for $0 < \pi(H_1) < 1$ and $S(H_1,H_2) > 0$, $\frac{d\frac{\hat{\pi}(H_1)}{\hat{\pi}(H_2)}}{d\pi(H_1)} > 0$, and thus we have strict monotonicity. Furthermore, for $\frac{\hat{\pi}(H_1)}{\hat{\pi}(H_2)} = \frac{\pi(H_1)}{\pi(H_2)}$ out of $\pi(H_1) \neq 0, 1$, then we must have:

$$\pi(\mathrm{H}_1) + \frac{\mathrm{S}(\mathrm{H}_1,\ \mathrm{H}_2)}{\mathrm{S}(\mathrm{H}_1,\mathrm{H}_1)}\ \pi(\mathrm{H}_2) = \pi(\mathrm{H}_2) + \frac{\mathrm{S}(\mathrm{H}_1,\mathrm{H}_2)}{\mathrm{S}(\mathrm{H}_2,\mathrm{H}_2)}\ \pi(\mathrm{H}_1)$$
$$\leftrightarrow \left(1 - \frac{S(H_1,H_2)}{S(H_2,H_2)}\right)\pi(\mathrm{H}_1) = \left(1 - \frac{S(H_1,H_2)}{S(H_1,H_1)}\right)\pi(H_2),$$

Which implicitly defines π^*.

Finally, note:

$$\hat{\pi}(\mathrm{H}_1) > \pi(H_1) \leftrightarrow \frac{\pi(\mathrm{H}_2)+\frac{\mathrm{S}(\mathrm{H}_1,\mathrm{H}_2)}{\mathrm{S}(\mathrm{H}_2,\mathrm{H}_2)}\pi(\mathrm{H}_1)}{\pi(\mathrm{H}_1)+\frac{\mathrm{S}(\mathrm{H}_1,\mathrm{H}_2)}{\mathrm{S}(\mathrm{H}_1,\mathrm{H}_1)}\pi(\mathrm{H}_2)} > 1 \leftrightarrow \left(1 - \frac{S(H_1,H_2)}{S(H_2,H_2)}\right)\pi(\mathrm{H}_1) < \left(1 - \frac{S(H_1,H_2)}{S(H_1,H_1)}\right)\pi(H_2).$$

Assuming that $S(H_1,H_2) < S(H_2,H_2)$, the above is satisfied iff $\pi(H_1) < \pi^*$, as desired.

Corollary 1 The proof follows from the inspection of the expression:

$$\frac{\hat{\pi}(\mathrm{H}_1)}{\hat{\pi}(\mathrm{H}_2)} = \frac{\pi(\mathrm{H}_1)}{\pi(\mathrm{H}_2)} \cdot \frac{\pi(\mathrm{H}_2) + \frac{\mathrm{S}(\mathrm{H}_1,\mathrm{H}_2)}{\mathrm{S}(\mathrm{H}_2,\mathrm{H}_2)}\ \pi(\mathrm{H}_1)}{\pi(\mathrm{H}_1) + \frac{\mathrm{S}(\mathrm{H}_1,\ \mathrm{H}_2)}{\mathrm{S}(\mathrm{H}_1,\mathrm{H}_1)}\ \pi(\mathrm{H}_2)}$$

Corollary 2 Let $r(A;H_2)$ denote the probability of sampling event A when thinking about H_2. Then, the share of successful recall events for H_2 that fall in subset $H_{21} \subset H_2$ are:

$$\frac{r(\,H_{21};H_2)}{r(H_2)} = \frac{S(\,H_{21},H_2)\pi(\,H_{21})}{S(H_2,H_2)\pi(H_2)}.$$

H_2' is undersampled relative to its true frequency if and only if $S(H_2',H_2) < S(H_2,H_2)$.

Given that: $S(H_2,H_2) = S(\,H_{21},H_2)\cdot\pi(\,H_{21}|H_2) + S(H_2\setminus H_{21},H_2)\pi\cdot(H_2\setminus H_{21}|H_2)$, this condition can be written as $S(\,H_{21},H_2) < S(H_2\setminus H_{21},H_2)$.

In turn, note:

$$S(\,H_{21},H_2) = S(\,H_{21},H_2')\pi(\,H_{21}|H_2) + S(\,H_{21},H_2\setminus H_{21})\pi(H_2\setminus H_{21}|H_2)$$
$$S(H_2\setminus H_{21},H_2) = S(\,H_{21},H_2\setminus H_2')\pi(\,H_{21}|H_2) + S(H_2\setminus H_{21},H_2\setminus H_{21})\pi(H_2\setminus H_{21}|H_2)$$

Thus, noting that $S(A,B) = S(B,A)$, we have that $S(\,H_{21},H_2) < S(H_2\setminus H_{21},H_2)$ if and only if:

$$[S(\,H_{21},H_{21}) - S(\,H_{21},H_2\setminus H_{21})]\cdot\pi(\,H_{21}|H_2) < [S(H_2\setminus H_{21},H_2\setminus H_{21}) - S(\,H_{21},H_2\setminus H_{21})]\cdot\pi(H_2\setminus H_{21}|H_2)$$

If the events are equally self-similar, $S(\,H_{21},H_{21}) = S(H_2\setminus H_{21},H_2\setminus H_{21})$, given that self similarity is larger than cross similarity, the condition holds if and only if $\pi(\,H_{21}|H_2) < \pi(H_2\setminus H_{21}|H_2)$.

Proposition 4 Given two subsets H_{21} and H_{22}, with three hypotheses (H_1,H_{21},H_{22}) the recall fluency of H_1, $r(H_1)$, is the same as when there are only hypotheses H_1 and H_2. We then have:

$$r(H_{2i}) = \frac{S(H_{2i},H_{2i})\pi(H_{2i})}{S(H_{2i},H_{2i})\pi(H_{2i}) + S\big(H_{2i},H_{2j}\big)\pi\big(H_{2j}\big) + S(H_{2i},H_1)\pi(H_1)}.$$

Denote $S(H_{2i}, H_{2i}) = S^*$ and $S(H_{2i}, H_1) = S_*$ which are by assumption independent from $i = 1,2$. We can then write:

$$r(H_{2i}) = \frac{1}{2}\frac{S^*\pi(H_2)}{\frac{[S^* + S(H_{2i}, H_{2j})]}{2}\pi(H_2) + S_*\pi(H_1)}.$$

Given the symmetry of similarity, $r(H_{21}) = r(H_{22})$. As a result, the probability of successes in sampling H_2, namely the sum of successes in sampling H_{21} and H_{22} follows a binomial distribution $Bin(2T, r(H_{21}))$. As a result, when the hypotheses is split into the two subsets, the average number of successes in recalling H_2 per attempt T is:

$$r'(H_2) = \frac{S^*\pi(H_2)}{\frac{[S^* + S(H_{2i}, H_{2j})]}{2}\pi(H_2) + S_*\pi(H_1)}.$$

When the hypothesis is not split, we have:

$$r(H_2) = \frac{S(H_2, H_2)\pi(H_2)}{S(H_2, H_2)\pi(H_2) + S(H_2, H_1)\pi(H_1)},$$

which, given the assumption, $S(H_{21}, H_1) = S(H_{22}, H_1) = S_*$ is equal to:

$$r(H_2) = \frac{S(H_2, H_2)\pi(H_2)}{S(H_2, H_2)\pi(H_2) + S_*\pi(H_1)}.$$

Because the two subsets are equally likely, it is immediate to find that $\frac{[S^* + S(H_{2i}, H_{2j})]}{2} = S(H_2, H_2)$. It then immediately follows that $r'(H_2) > r(H_2)$ if and only if $S^* > S(H_2, H_2)$, which is equivalent to $S(H_{2i}, H_{2i}) > S(H_{2i}, H_{2j})$.

Proposition 5 See Appendix A2 for a proof of a more general result.

Proposition 6 We can work out the distribution of $\hat{\pi}(H_i)$ starting from the characterization of the odds in Proposition 2. Using the same logic of the proof of Proposition 2, for large enough T the odds of H_j relative to H_i are distributed as:

$$\frac{N_j}{N_i} \to N\left(\frac{r(H_j)}{r(H_i)}, \frac{1}{T}\cdot\left(\frac{r(H_j)^2}{r(H_i)^4}\cdot r(H_i)(1 - r(H_i)) + \frac{1}{r(H_i)^2}\cdot r(H_j)\left(1 - r(H_j)\right)\right)\right).$$

By applying the delta-rule on $f(\epsilon) = \frac{1}{1+x+\epsilon} = \frac{1}{1+x} - \frac{1}{(1+x)^2}\,\epsilon + O(\epsilon^2)$, where $x = \frac{r(H_j)}{r(H_i)}$ is the mean above and ϵ is the Gaussian discrepancy from it, we find that $\hat{\pi}(H_i)$ follows the asymptotic distribution:

$$\hat{\pi}(H_i) = \frac{1}{1 + \frac{N_j}{N_i}} \sim N\left(\frac{r(H_i)}{r(H_i) + r(H_j)}, \frac{1}{T}\cdot\frac{r(H_i)^2 r(H_j)^2}{[r(H_i) + r(H_j)]^4}\left[\frac{1 - r(H_i)}{r(H_i)} + \frac{1 - r(H_j)}{r(H_j)}\right]\right).$$

The model thus predicts noise in $\hat{\pi}_i = \hat{\pi}(H_i)$ to be equal to be a function of recall fluencies:

$$v(\hat{\pi}_i) = \frac{1}{T}\cdot\frac{r_i^2 r_j^2}{(r_i + r_j)^4}\left(\frac{1 - r_i}{r_i} + \frac{1 - r_j}{r_j}\right),$$

where $r_i = r(H_i)$. This can be written as:

$$\nu(\hat{\pi}_i) = \frac{1}{T} \cdot \frac{r_i r_j \left(r_i + r_j - 2 r_i r_j\right)}{\left(r_i + r_j\right)^4}.$$

Clearly the expression fulfils $\nu(\hat{\pi}_i) = \nu\left(\hat{\pi}_j\right) = \nu(1 - \hat{\pi}_i)$.

To prove the first property, it is evident that $\frac{\partial \nu(\hat{\pi}_i)}{\partial T} < 0$. For the second property, after some algebra one can find that:

$$\frac{\partial \nu(\hat{\pi}_i)}{\partial r_i} \propto r_j^2 (1 - 4 r_i) + r_i r_j (4 r_i - 3) - 4 r_i^2,$$

Define $R = r_i + r_j$. After some algebra one can find:

$$\frac{\partial \nu(\hat{\pi}_i)}{\partial r_i} \propto 1 - 4\hat{\pi}_i - \hat{\pi}_i^2 - 4\hat{\pi}_i (1 - \hat{\pi}_i)^2 R.$$

A sufficient condition for the above derivative to be negative is obtained by imposing $R \mapsto 0$, which yields:

$$\frac{\partial \nu(\hat{\pi}_i)}{\partial r_i} < 0 \quad if \quad \hat{\pi}_i > \sqrt{5} - 2,$$

which also yields:

$$\frac{\partial \nu(\hat{\pi}_i)}{\partial r_j} < 0 \quad if \quad \hat{\pi}_i < 3 - \sqrt{5}.$$

Appendix A2: Generalization of the conditional assessments result

In this section, we prove the result of the under-and-overestimation of conditional beliefs for general similarity specification. In other words, the individual is assessing $\pi(H_1|D)$. For notational simplicity, we shall denote the four sub-populations as follows: $H_1D, H_2D, H_1\overline{D}, H_2\overline{D}$, and we shall use as a loose-hand $S(A,B)$, where $A, B \in \{H_1D, H_2D, H_1\overline{D}, H_2\overline{D}\}$, the similarity between any two given sub-populations.

We assume for simplicity that the self-similarity of all subgroups is equal to 1, and $\pi(D) = \pi(\overline{D})$ Furthermore, we assume the following inequalities regarding the similarity between subgroups:

$$S(H_iD,\ H_i\overline{D}) > S(H_iD, H_{-i}\overline{D}),$$

Where $i = 1, 2$ and $-i = 2, 1$. To give an intuition behind this assumption, let us use the example in the paper $H_1 =$ "accident" and $H_2 =$ "sickness" for $D =$ "young" and $\overline{D} =$ "older". The above assumption is saying that the events that belong to the same cause of death among different populations are more similar to each other than events that do not share the same cause of death nor the population (e.g. a young death by accident is more similar to an older death by accident than an older death by sickness). This assumption is not only satisfied in the set-up of Proposition 5, but also for general similarity functions that increase in the number of overlapping features.

Applying Proposition 2 yields the following:

$$\frac{\hat{\pi}(H_1|D)}{\hat{\pi}(H_2|D)}$$

$$= \frac{\pi(H_1D)}{\pi(H_2D)}\left[\frac{\pi(H_2D) + S(H_1D, H_2D)\pi(H_1D) + S(H_2D,\ H_2\overline{D})\pi(H_2\overline{D}) + S(H_2D,\ H_1\overline{D})\pi(H_1\overline{D})}{\pi(H_1D) + S(H_1D, H_2D)\pi(H_2D) + S(H_1D,\ H_1\overline{D})\pi(H_1\overline{D}) + S(H_1D,\ H_2\overline{D})\pi(H_2\overline{D})}\right]$$

Setting $\psi = \frac{\hat{\pi}(H_1|D)}{\hat{\pi}(H_2|D)} \Big/ \frac{\pi(H_1 D)}{\pi(H_2 D)}$ as the distortion in the likelihood ratio, note that individuals overestimate $\pi(H_1|D)$ if and only if $\psi > 1$. This occurs if and only if:

$$\begin{aligned}(S(H_2D,\, H_2\overline{D}) + S(H_1D,\, H_1\overline{D}) - S(H_2D,\, H_1\overline{D}) - S(H_1D,\, H_2\overline{D}))\pi(H_2|\overline{D}) \\ + \Big(1 + S(H_2D,\, H_1\overline{D}) - \big(S(H_1D, H_2D) + S(H_1D,\, H_1\overline{D})\big)\Big) \\ > 2 \cdot \big(1 - S(H_1D,\, H_2D)\big)\pi(H_1|D)\end{aligned}$$

Now, by our assumption on similarity, note that both coefficients on $\pi(H_2|\overline{D})$ and $\pi(H_1|D)$, $(S(H_2D,\, H_2\overline{D}) + S(H_1D,\, H_1\overline{D}) - S(H_2D,\, H_1\overline{D}) - S(H_1D,\, H_2\overline{D}))$, and $\big(1 - S(H_1D,\, H_2D)\big)$, are positive. Consequently, this implies the following:

Proposition 5(General). For general similarity structure, one has overestimation of $\pi(H_1|D)$ ceteris paribus if:

a) $\pi(H_1|D)$ decreases below a certain threshold

b) $\pi(H_1|\overline{D})$ decreases below a certain threshold

For the similarity function $S(e_k, e_{k'}) = \delta^{\Sigma_i |f_{ki} - f'_{ki}|}$, the above inequality further simplifies to:

$$2\delta \cdot \big(1 - \pi(H_1|\overline{D})\big) + (1 - \delta) > 2\pi(H_1|D).$$

Rearranging gives the formula in Proposition 5.

Appendix A3: Generalization to multiple hypotheses

In this section, we sketch out the extension of our model to the agent assessing multiple hypotheses. In this case, we assume $H_{i=1,\dots J}$ partition D, with $E = D \cup \overline{D}$. We assume now that the agent goes through the train of thought for each of the $J > 2$ hypotheses. Then, the ease of retrieval for hypothesis j is given by:

$$r(H_i) = \frac{\pi(H_i)}{\pi(H_i) + \sum_{\{j \neq i\}} \frac{S(H_i, H_j)}{S(H_i, H_i)}\ \pi(H_j) + \frac{S(H_i, \overline{D})}{S(H_i, H_i)}\ \frac{\pi(\overline{D})}{\pi(D)}}.$$

Then, as before, the agent's assessment of hypothesis is given by

$$\hat{\pi}(H_j) = \frac{R_j}{\sum_k R_k}.$$

Finally, regarding the generalization of Proposition 2 for $J > 2$ hypotheses, the result regarding the mean follows from an argument regarding law of large numbers. Again, a small detail is to assume that $T \mapsto \infty$ such that the probability of $\sum_k R_k = 0$ becomes vanishingly small, holding J fixed. Then, note that we have:

$$\hat{\pi}(H_j) = \frac{R_j}{R_j + \sum_{k \neq j} R_k} \mapsto_p \frac{r(H_j) + z_j/\sqrt{T}}{r(H_j) + \sum_{k \neq j} r(H_k) + z_{-j}/\sqrt{T}},$$

Where $z_j \sim N(0,\ r(H_j)\left(1 - r(H_j)\right),\ z_{-j} \sim N\left(0,\ \sum_{k \neq j} r(H_k)(1 - r(H_k))\right)$.

The rest of the derivation proceeds similarly.

Appendix B: Details of Experimental Design

In this section, we describe the design of the two experiments in greater detail. Both experiments were pre-registered on the AEA RCT Registry, with ID AEARCTR-0006676.

Design of Experiment 1

Experiment 1 was conducted in March of 2021 among Bocconi undergraduates. We recruited participants via email from the experimental economics listserv. Participants completed the experiment online from home (due to Covid restrictions). They earned a 4 euro Amazon gift card as an incentive to take the survey, and had the chance to earn an addition 2 euro bonus for correct responses. The median respondent took less than 10 minutes to complete the survey. In total, 1,200 respondents participated in the survey (this sample size was preregistered), roughly 240 for each of the five treatments.

Participants could choose to take the survey in either Italian or English, of which 19% chose the latter. Both the text of the questions and the words in the memory task were translated according to participants' choice.

After a consent form, participants were told that the survey would include several questions for which they could earn a 2 euro bonus for answering correctly. At the end of the experiment, the computer would randomly choose one of these questions as the "Bonus Question" to base their bonus on.

Participants were then told they would be shown a series words, one by one in a random order, which would take about a minute. They were told that the Bonus Question would be about these images, and were therefore encouraged to pay attention. They then answered three simple comprehension questions, with corrections for errors.

After the words, participants had to wait 10 seconds before proceeding to the questions. They were encouraged to "Take a moment to reflect on the words you saw."

They were then asked the probabilistic question about the percent chance that it fell into various categories, as described in the main text. They were told that, if this question was chosen as the Bonus Question, they would earn the 2 euro bonus if their answer was within 5 percentage points of the right answer.

Participants then saw the confidence question asking they how certain they were that their previous answers were within 5 percentage points of the right answer.

Participants then answered the free recall questions, which asked them to "please list up to 15 [category] that you remember seeing in the list of words we showed you. You do **not** have to fill in all 15 lines." If a recall question was chosen as the Bonus question, participants' were told that their "chance of earning the 2 euro bonus will increase by 10 percentage points for every correct answer, and it will decrease by 10 percentage point for every incorrect answer you're your chance cannot go below zero or above 100%). Don't worry about typos or misspellings."

Participants were then asked a series of follow-up questions. The first asked about whether they felt they paid more attention to the early, middle, or late part of the sequence of words (the large majority say the early part). The next asked about whether they noticed that the words fell into any categories and prompted them to write what categories they remembered seeing.

Next, they were asked if, when watching the words, they wrote any of them down or took a video to refer to later. 13% report doing so, though excluding them from the analysis does not change any qualitative results (though their overall performance on the recall task is, unsurprisingly, better than average).

Finally the survey asked about native language, age, gender, and whether they found the survey questions difficult or easy to answer and whether they found the instructions confusing or clear. The majority found it difficult (by far the most common reason given in a free response is that it was difficult to remember all the words) and 99% found the instructions very or moderately clear.

T5 treatment in Experiment 1

In addition to the treatments described in the main text, in Experiment 1 we also ran a treatment called T5. T5 was identical to the T1 except we replaced women's names in T1 with "ocean animals" (e.g., "Shark", "Whale", etc.). Note that, in T1, all the animals were "land animals" (e.g., "Lion", "Dog", etc). In addition, instead of asking about the probability of the randomly chosen word being an "animal" as in T1, in T5 the question asked for the probability that it was a "land animal." The free recall task in T5 correspondingly asked respondents to recall up to 15 "land

animals." The purpose of this treatment was to increase the cross-similarity between the hypotheses under consideration. In T5, "Land Animals" and "Other" are arguably more similar than "Animals" and "Other" in T1. According to our theory, this should reduce the ease of recall of land animals in T5 compared to T1. Panel B of Figure B1 shows that, indeed, respondents correctly recall fewer land animals in T5 than in T1.[35]

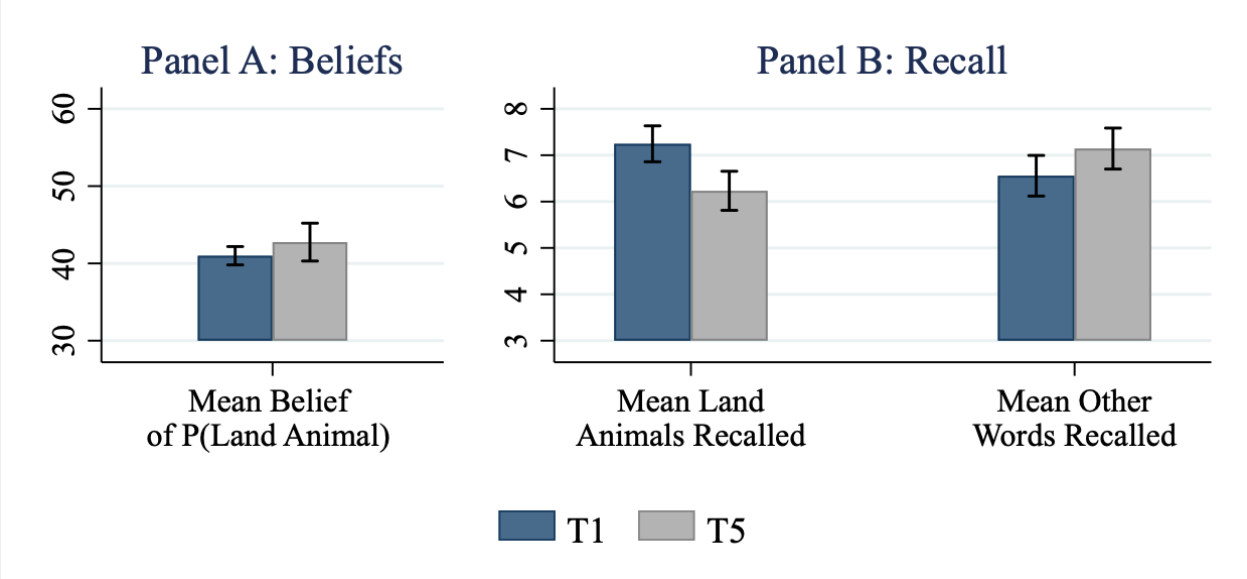


Figure B1: Beliefs and Recall in T1 vs T5

Notes: Panel A shows the average belief of the probability of (land) animals minus the true proportion in T1 and T5. Panel B shows the number of land animals and non-land-animals recalled in the free recall task in these treatments. Bands show 95% confidence intervals.

However, there appears to have been significant confusion about what qualified as an "ocean animal." 27% of respondents list an ocean animal when asked to recall a land animal (in contrast,

[35] Whether there should be greater ease of recall for the other category in T1 than in T5 depends on whether men's names and women's names are more self-similar than men's names and ocean animals. In fact we find somewhat higher recall of non-land-animals in T5 than in T1 ($p = 0.07$), suggesting that if anything the opposite is true. An alternative explanation, outside the model, is that ocean animals are simply more memorable or salient than women's names. Consistent with the effects on recall, beliefs about the probability of land animals are slightly higher in T5 than in T1, though the difference is not statistically significant (point estimate = 1.8pp, $p = 0.20$).

in T1, no respondents list a name when asked to recall animals), suggesting that this distinction was less natural than we anticipated.

Additional Analyses of Experiment 1

Time as a Proxy for *T*

Figure B2 plots the relationship between the time subjects spent on the beliefs question (winsorized at 100 seconds) and the sample variance of the fraction of recalled words that were animals (Panel A) and of beliefs about P(Animal) (Panel B). We see a statistically significant ($p < 0.01$) negative relationship in both cases.

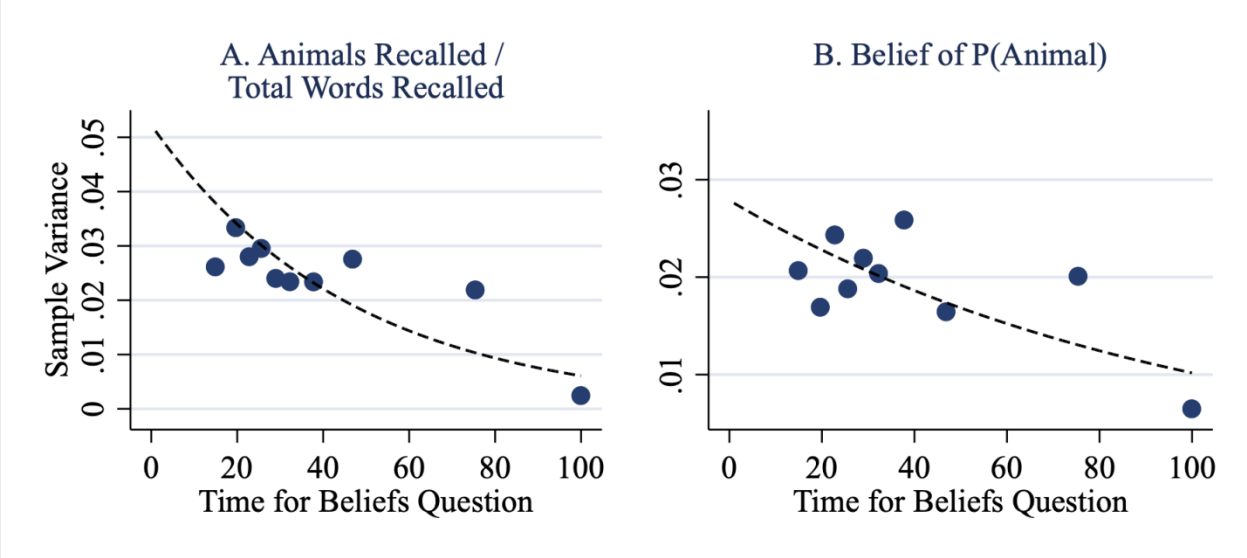


Figure B2: The Relationship between Recall Time and Noise

Notes: Panel A shows the sample variance of the fraction of recalled words that are animals (y-axis), conditional on decile of time spent on the beliefs question (x-axis). Panel B shows the sample variance of beliefs about the probability animals, also conditional on decile of time spent on the beliefs question (x-axis). The dotted lines show predicted values from maximum likelihood estimates of a model where the mean of the dependent variable varies linearly and the conditional variance multiplicatively in time spent on the beliefs question. Both panels restrict the data to *T1* and *T3*, where the true frequency of animals is 40%.

Recency Effects

Figure B3 plots the probability of recalling an exemplar as a function of its chronological position in the sequence.

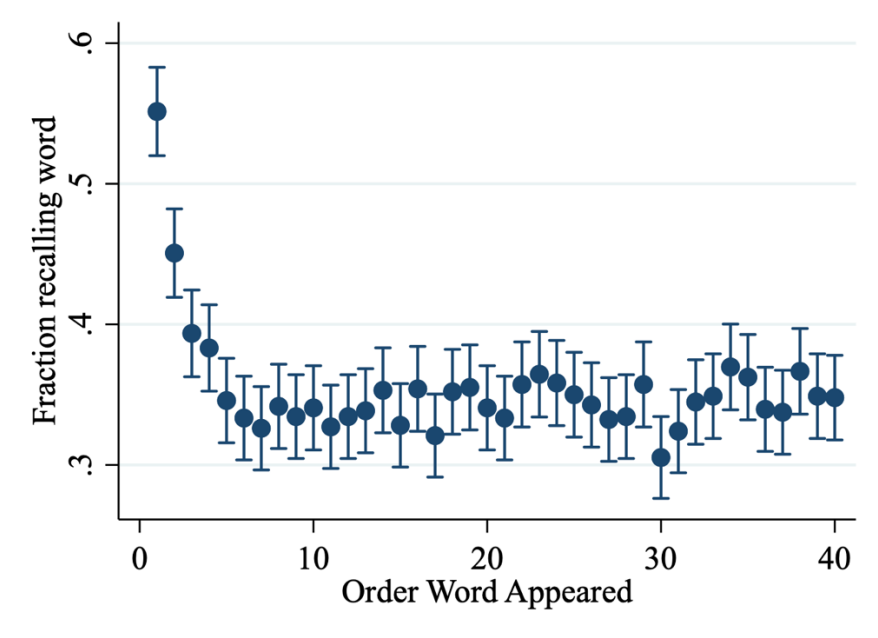


Figure B3: Primacy effect in Recall

Notes: Figure shows the proportion of respondents who listed a word in one of the free recall tasks (y-axis) depending on the chronological order in which the word appeared to them (x-axis). The order of words was randomized. Bands show 95% confidence intervals.

Consistent with existing work on memory (Kahana, 2012), there is a strong primacy effect in recall, in which early experiences are likelier to be recalled than late ones. If retrieval from memory reflects probability beliefs, then subjects who see many Animals early should have higher subjective probability of Animals, and vice versa. Contrary to the literature's finding, however, we do not find a strong recency effect, or the tendency to recall the latest entries. Indeed, regressing beliefs on the number of animals shown in the first 5 words shows that indeed those who happened to be shown more animals first, controlling for treatment, report a higher probability for animals:

$$P^{belief}(Animal) = Constant + .90\ (.28) \cdot \#\ Animals\ in\ First\ 5\ Words$$

We take this to be further evidence of the role of recall in shaping beliefs.[36]

Relationship to cognitive uncertainty and beliefs

By linking probability beliefs to measurable recall, our model also speaks to research connecting cognitive imprecision and uncertainty to probability judgments. Enke and Graeber (2019) find greater attenuation to 50-50 for agents who report a higher level of subjective, or cognitive uncertainty. The authors show that their results can arise from a Bayesian processing of noisy cognitive signals – the greater the noise, the greater the subjective uncertainty and shrinkage towards the prior mean. However, beyond its indirect impact on subjective uncertainty and probability judgments, measuring cognitive noise remains an important challenge.

Our measurement of free recall in conjunction with probability elicitation provides a complementary insight into the underlying determinants of cognitive uncertainty. If people's probabilistic beliefs are informed by sampling from memory, an important source of noise and uncertainty can come from the inability to recall relevant data. To evaluate this approach, after participants give their probabilistic beliefs (but before the recall tasks), we ask them how confident they are that their answers are within five percentage points of the true answer (the threshold for earning a bonus for accuracy). They respond by dragging a slider that ranged from "Very Uncertain" to "Very Certain," which we convert to a 0-100 scale.

[36] We take these results to be somewhat suggestive. For example, we do not find the analogous impact on recall: people who see a lot of animals at the beginning do not disproportionately remember animals. Second, our effect is concentrated in participants who get 4 or 5 animals in the first 5 words.

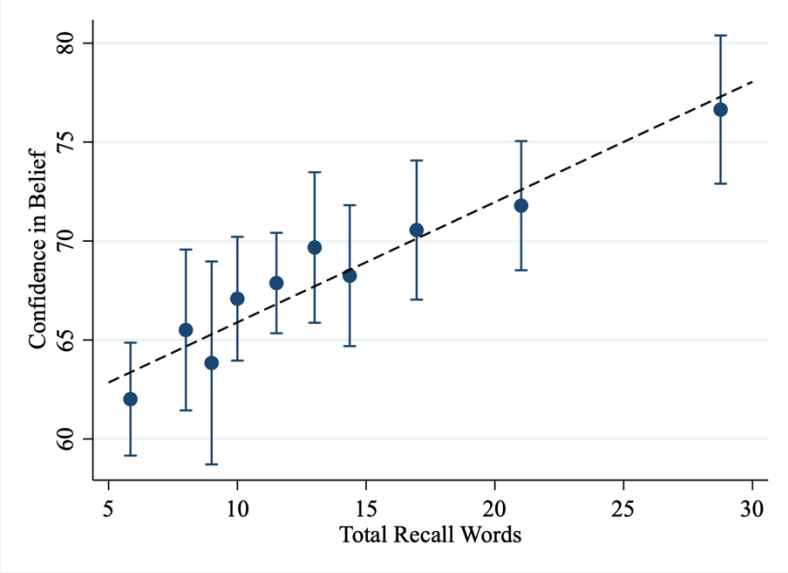


Figure B4: Confidence and Recall

Notes: This figure shows the average confidence in respondents' belief of the probability of animals, conditional on the total number of recalled words in the free recall tasks (binned by decile). Bands show 95% confidence intervals. The dashed line shows the OLS line of best fit.

Figure B4 shows a strong positive relationship between the total number of recalled entries and this measure of participants' confidence. The red line shows the OLS estimate, which is highly significantly positive ($p < 0.001$). To summarize, our exploratory analysis suggests a strong link between recall and subjective uncertainty. Our suggestive findings are consistent with the insight that failure to retrieve relevant information leads to higher subjective uncertainty: the limitation of memory is an important driver of subjective uncertainty.

Design of Experiment 2

Experiment 2 was conducted in November 2020, also online with Bocconi undergraduates recruited in the same was as Experiment 1. Participants earned a 4 euro Amazon gift card for completing the survey, plus a possible 2 euro bonus for accuracy. As The median respond took

about 9 minutes to take the survey. In total 1,203 respondents participated in the experiment, approximately 150 per treatment. The pre-registered sample size was 1,200, though three additional respondents completed the survey before we closed it.

The survey had the same design as Experiment 1, with a few exceptions. First, there were no comprehension questions following the instructions. Second, respondents were told that there would be 40 images, each of which would be either a word or a number and would be blue or orange. After participants saw the words, there was not an enforced 10 second delay as in Experiment 1 (in practice the median respondent spent 4 seconds on the transition page after the images ended and before the questions about them).

After this, respondents were asked "Suppose the computer randomly chose an image from the images you just saw. It is orange. What is the percent chance it is a word? Please indicate your answer by clicking on the scale below and then dragging the slider." A slider below this text went from 0 to 100 (with no default starting value). The confidence question after this was the same as in Experiment 1.

After this question (which is the primary question we focus on), respondents were also asked an analogous question but assuming the randomly chosen image was blue. Finally they were asked the same question but supposing they did not know the color. Both of these questions also included a confidence question asking how confident they were that their answer was within 5 percentage points of the right answer.

Before the free recall question, respondents answered 4 questions of the following type: "Was X [this word appeared in either blue or orange] (in blue or orange) among the images you saw?" We chose 4 random words from the same list of time-related words but that were *not* among the images the respondent had seen. Two of these words appeared in blue and two appeared in orange, in a random order. If any of these questions was chosen to be paid on, the respondents earned the 2 euro bonus if they answered "no" (since this was always the correct answer).

The free recall question asked respondents to list up to 10 orange words that they remembered seeing. They were told that, if this question was chosen for payment, their chance of receiving the 2 euro bonus would increase by 10 percentage points for every correct answer and decrease by 10 percentage points for every incorrect answer (though it could not, of course, go below zero).

Finally the survey ended with some questions including age, native language, how difficult they thought answering question in the survey was, whether they found the instructions confusing, and whether they experienced any technical issues. The survey also presented a multiple choice questions asking respondents what, when they were viewing the images, they expected us to ask of them. 58% responded that they expected to be asked to list specific words, 39% expected to be asked to say how many images of different types (words, etc.) they saw, and only 17% expected to be asked about the colors of words. 38% said they "did not have anything in particular in mind".

Additional Analyses of Experiment 2

In Experiment 2, all respondents are first asked the probability that a randomly drawn image was a word conditional on it being orange. This question is the primary object of interest for our analysis. However, after this question, we also ask the probability that a randomly drawn blue image is a word and then the probability that a randomly drawn word unconditional on color is a word. For completeness, we report the average beliefs for these questions here. We did not ask corresponding recall questions for either blue words or for words unconditional on color.

Figure B5 shows the mean beliefs about the probability of words both conditional on blue (Panel A) and unconditional on color (Panel B). Panel A shows that while mean beliefs about the probability of words conditional on blue change dramatically between the H and L treatments, they do not go all the way to 100% and 0%, respectively. This may be because respondents inferred that we would not ask the question if the answer were trivial. When orange images are 50% words in the NM treatment (meaning that blue images are also 50% words), the average belief of P(Word | Blue) is 46.3 percent, lower than the truth ($p = 0.01$). For all other questions, the average belief is attenuated toward 50%, consistent with our theory's prediction that rare hypotheses should be exaggerated.

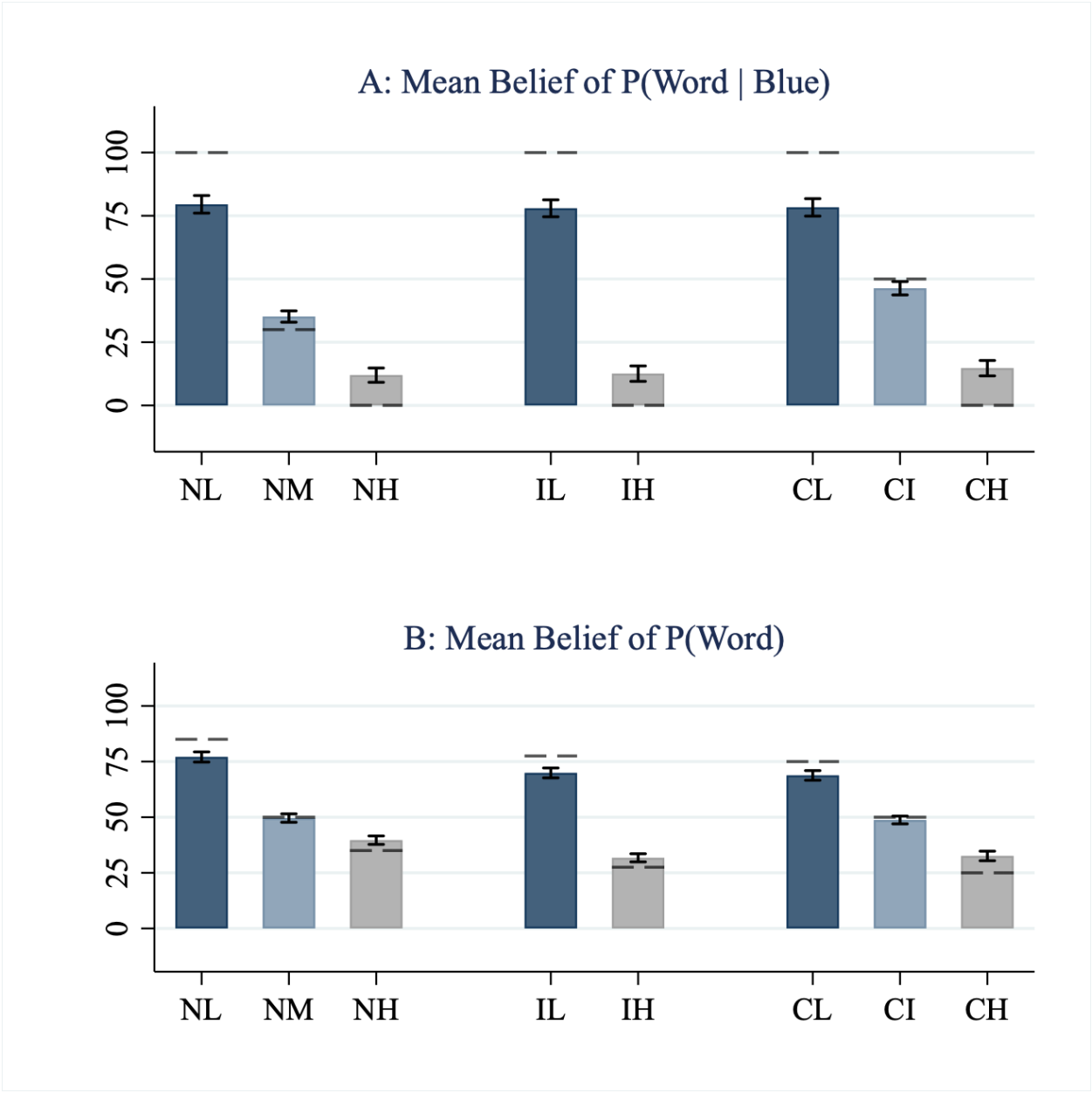


Figure B5: Other Beliefs in Experiment 2

Notes: Panel A shows the average belief of the probability of the randomly drawn image being a word conditional on it being blue. Panel B shows the average belief of the probability of the randomly drawn image being a word unconditional on its color. In the *L* treatments, all blue images were words. In the *H* Treatments, all blue images were numbers. 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. Dashed lines show the correct answer for each treatment.

In addition to the recall task described in the main text, in which respondents were asked to list orange words that they recalled seeing, the survey also included a "misrecognition" task. In particular, respondents were asked four questions, each asking them whether a specific word in a specific color was among the images they saw. Two of the named words were blue, and two were orange, and they appeared in a random order. In each case, the named word was *not* among the images they were previously shown.

Figure B6 below shows the average number of blue and orange words that respondents erroneously reported recognizing from the images they were shown. Overall, we do not see large or systematic differences in the number of orange words mis-recognized across treatments. We do see large differences across treatment in the number of blue words mis-recognized. This is perhaps less surprising, as the frequency of blue words changed much more dramatically across treatments than the frequency of orange words.

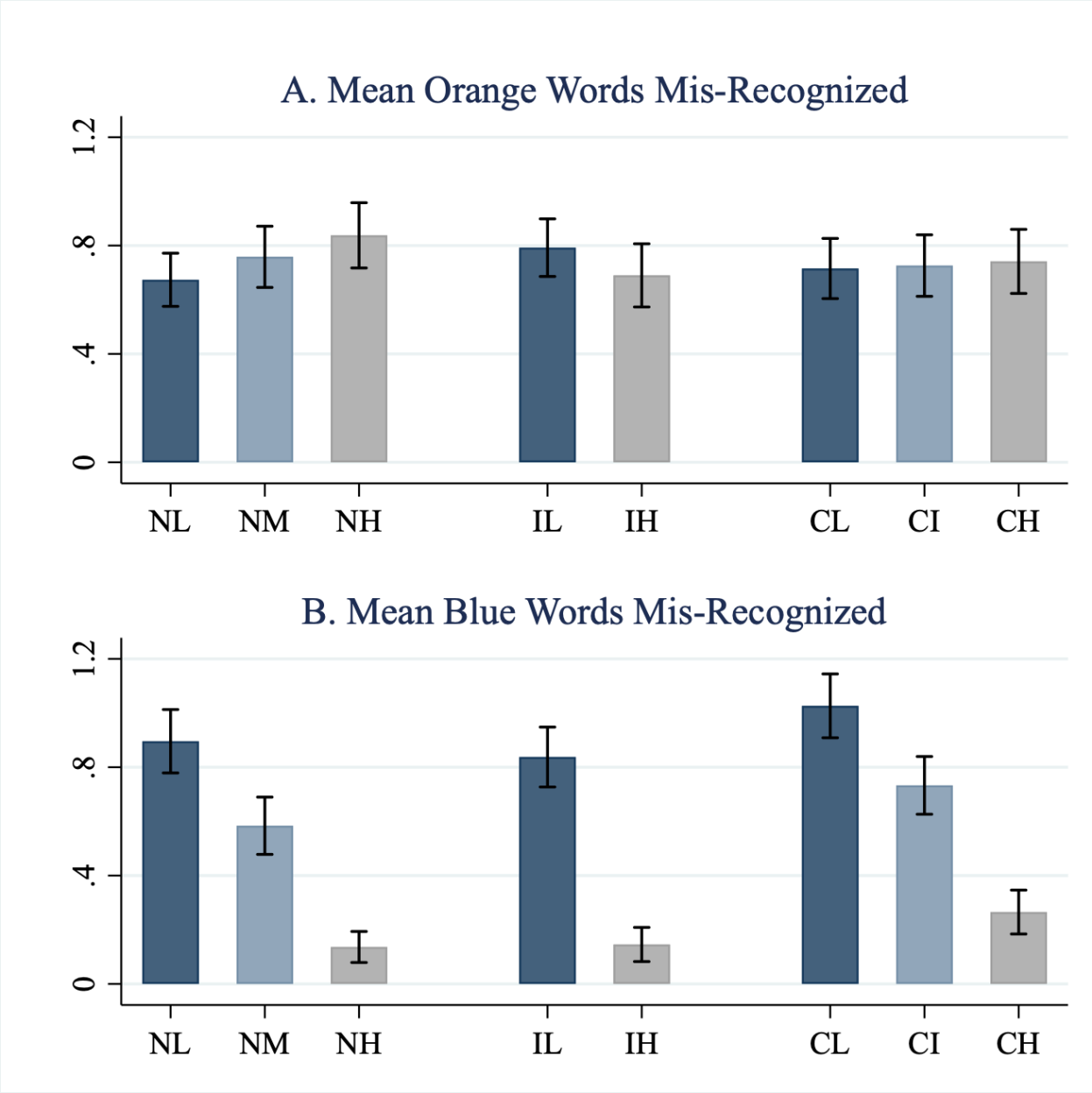


Figure B6: Mis-Recognition in Experiment 2

Notes: Panels A and B respectively show the average number of blue and orange words that respondents erroneously reported recognizing from the images they were shown. Bands show 95% confidence intervals.