Associative learning and representativeness*

Michael Jacob Kahana† James D. Paron‡ Jessica A. Wachter§

March 20, 2025

Abstract

The representativeness heuristic constitutes a striking departure from Bayesian updating. According to a strong form of the heuristic, agents reverse a conditioning arguments: for example inferring that a patient is more likely than not to have a rare disease, conditional on a positive test result. The correct inference is that a positive test result is more likely than not, conditional on disease. Recent research implicates representativeness in a wide range of financial market anomalies, with potential consequences for the real economy. However, the cognitive foundations of the representativeness heuristic (RH) remain unknown. Here, we show that the RH emerges from a theory of associative memory and recognition, leading to a cognitive foundation for the RH, and a means of integrating the RH into economic models involving decision-making under uncertainty.

Preliminary, Comments welcome

JEL codes: E03, G02
keywords:interference; base-rate neglect; context; recognition memory

*We thank Sudeep Bhatia, Pedro Bordalo, Spencer Kwon, Alice Healy, Peter Maxted, Nikolai Roussanov, John Sakon and participants at the 2020 Society of Financial Studies Cavalcade and at the Spring 2024 Memory Beliefs and Choice Meeting for helpful comments. We have no conflicts to disclose. We are grateful for financial support from NIH grant MH55687.

†Department of Psychology, University of Pennsylvania; Email: kahana@psych.upenn.edu

‡Department of Finance, The Wharton School, University of Pennsylvania and NBER; Email: jparon@wharton.upenn.edu

§Department of Finance, The Wharton School, University of Pennsylvania.

1 Introduction

Our memories of past experiences and their associated contexts guide not only our thoughts but also our choices. We choose to save when we think of the need for future resources or when we fear a negative shock to our wealth, and we choose to consume a specific product when positive experiences associated with that product come to mind. Whereas theories of economic choice have traditionally assumed that rational agents recall the correct probability distributions associated with different states of the world, some economic models contemplate the role of memory in shaping how we construct these distributions from our record of past experiences (Bordalo et al., 2020, 2018, 2019a; Nagel and Xu, 2018; Mullainathan, 2002; Wachter and Kahana, 2024).

Tversky and Kahneman (1973) follow a venerable tradition of psychological research in suggesting that subjective probability reflects memory for the frequency of occurrence. Further, this line of work has shown that memory for frequency may be biased by other factors that influence either the probability of remembering specific instances or the mnemonic strength of the retrieved information (Estes, 1976). Tversky and Kahneman show, in a series of experiments, that memories of instances that closely match the prototype of a given category (i.e., they are representative of that category) come to mind more easily than less representative instances (Kahneman and Tversky, 1972; Tversky and Kahneman, 1974, 1983). The enhanced retrievability of these instances (termed availability by Tversky and Kahneman (1973)) boosts probability judgments for these categories.

This idea has led to a fruitful area of research in economics and, in particular financial decision-making. Certain events in the environment may bring the beliefs about future events, and even a focus on the future itself, to mind. Consider the standard savings-consumption problem where ones desire to save is based on one's expected future consumption stream. Here, fear of a long recession might lead agents to save. Such fear presumably arises from memories of past experiences or knowledge regarding recessions and their relation to features of the environment. A series of influential

papers proposes that the representativeness heuristic (RH) lies behind such diverse phenomena as overvaluation of growth stocks (Barberis et al., 1998; Bordalo et al., 2019a), predictability of bond returns (Bordalo et al., 2018), securitization prior to the crisis (Gennaioli and Shleifer, 2018), and boom/bust cycles in aggregate investment (Bordalo et al., 2019b).

Here we consider whether the RH may arise from a set of core psychological principles governing the encoding and retrieval of associations in memory. We specifically show how a model that embodies these principles can explain the bias in individuals' beliefs that has come to be known as the representativeness heuristic. Our work is based in the retrieved-context theory of Howard and Kahana (2002). Kahneman and Tversky (1972) relate representativeness to "what comes to mind," i.e. memory retrieval. Indeed, Tversky and Kahneman (1973) focus on the notion of availability in explaining representativeness. The idea is that representative memories come to mind easily, but why? Our answer to this question lies in notions of associations, a latent context, and similarities that have long been part of the toolkit of associative memory (Kahana et al., 2005). But whereas theories of memory grapple with data from studies that control the conditions of the encoding, retention, and retrieval of items and associations, Kahneman and Tversky asked naive subjects to make probability judgments in response to survey questions involving hypothetical scenarios. A seminal experiment by Bordalo et al. (2021), however, provides the foundation for building a memory-based theory of representativeness. Following the classic tradition of probability learning experiments, which often supported rational choice models (Estes, 1972), Bordalo et al. (2021) had subjects learn sets of novel associations, varying the degree to which a particular response was representative of a given stimulus. Their findings, which reify Kahneman's representativeness heuristic in a rigorous setting, form the basis of our analysis.

As emphasized by Bordalo et al. (2023), grounding the representativeness heuristic in basic cognitive theory offers the potential to unify seemingly disparate ideas in behavioral economics. Representativeness may be important in its own right, but a

question remains: Does it have a connection to ideas other behavioral mechanisms researchers have proposed and that have found support in the data? Such ideas include extrapolative expectations (Barberis et al., 2015), imperfect common knowledge (Angeletos and La'O, 2009), social influences on beliefs (Burnside et al., 2016), and difficulty in unraveling correlations (Enke and Zimmermann, 2017). These ideas each have an apparent tie to an evolving mental database that is based on a series of imperfectly retrieved perceptual data.

2 What is the representativeness heuristic?

Before developing a formal definition of the heuristic, it is useful to consider some motivating examples.

Example 1 (Kahneman and Tversky (1973))**.** *Subjects are given a short description of an intelligent, organized, introverted student and are asked to rank, in order of likelihood, several different fields of graduate study. Subjects tend to state that the student is more likely to study computer science than humanities, even though there are many more students in humanities than in computer science.*

More precisely, one group of study participants were asked to estimate base rates of academic subjects across nine fields. A second group was given a detailed description of an introverted man (Tom W.), and asked to rank the nine fields in terms of similarity, while the third group was asked to rank the nine fields "in order of the likelihood that Tom W. is now a graduate student in each of these fields." They found a negative correlation between base rate estimates and likelihood (the opposite of what Bayesian probability computations would predict), and a near-perfect correlation between the similarity measures and the likelihood. Specifically, participants gave the base rate of humanities as far higher than computer science, but both the similarity and the likelihood of computer science as higher than the humanities,

Example 2 (Casscells et al. (1978))**.** *Doctors are told that a test for a condition has*

come back positive. Doctors tend to believe that the patient has the condition, even though most patients who test positive do not have the condition.

Casscells et al. (1978) asked a group of medical professionals the following question: "If a test to detect a disease whose prevalence is 1/1000 has a false positive rate of 5 percent, what is the chance that a person found to have a positive result actually has the disease, assuming that you know nothing about the person's symptoms or signs?" The correct answer is around 1 in 51 (less than 2%): 1 of the 1000 will have an accurate positive test and 5% of the remaining 999 (about 50 people) will have false positives. The mean response was 55.9% and the modal response, given by 45% of participants, was 95%. 18% of participants gave the correct answer.

Bordalo et al. (2021) offer a different type of example.

Example 3 (Bordalo et al. (2021))**.** *Study participants view a sequence of images: 10 orange numbers, 15 blue numbers, and 25 blue words. When asked the likely color of an object, given that it is a number, 50% of participants say orange is the likely color.*

As Bordalo et al. (2016) emphasize, examples such as these display a *kernel of truth*: computer science students are more likely to be introverted than those in the general population. Those that have a condition are more likely to test positive for that condition. And, in the Bordalo et al. (2021) experiment, if an item is orange, it must be a number. However, inference by study participants goes further, to the incorrect conclusion that introverted students are more likely to study computer science, ignoring the base rate of computer science students in the population (as of the 1970s), and concluding that numbers are more likely to be orange, ignoring the base rate of blue images. However, while subjects exhibit evidence of non-rational expectations (the elicited subjective and the physical distributions clearly fail to match), the inference does not seem *that* irrational – we can see ourselves making a similar mistake. One senses that these examples are getting at a very real cognitive deficit to which humans are subject, analogous to an optical illusion.

To formalize these notions, assume there is a population of interest, denoted Ω. The examples above generally relate involve pairs, so we think of each member of the population having a pair of features (for example, being a number and being blue, or being an introvert and being a salesman). An advantage of this framing is that it allows us to connect to the large literature in psychology on associative recognition. As in a standard paired associates task (either cued recall or associative recognition), one element of the pair will form the cue. Following Bordalo et al. (2021), we refer to the set of possible cues as $\mathcal{D} = \{d_1, \ldots, d_K\}$. By analogy with Bayesian updating, the cue forms the "data" on which the agent conditions. The other element of the pair is subject to retrieval based on the cue. Again following Bordalo et al., we refer to this as the *hypothesis* and denote the set of possible hypotheses by $\mathcal{H} = \{h_1, \ldots, h_J\}$.Let P be the physical probability measure P over $\Omega = \mathcal{H} \times \mathcal{D}$.

The experiment or survey elicits probability-type statements, which we capture in an elicited probability measure $\hat{P}$. It is worth contrasting this elicited measure with a subjective probability measure as commonly understood. The psychological theory implicit in many models of economic decision-making is that an agent carries a probability measure and updates it in an manner that either approximates the laws of probability, or departs from these laws in certain defined ways. Our view is different: the agent responds to cues and it is the associations that the agent carries in their head and updates. The laws of associations, not the mathematical laws of probability, govern the updating process.

We now formulate a general statement of the kernel of truth and the representativeness heuristic.

Definition. *The physical probability measure P and the hypothesis-data pair (h, d) exhibits the* kernel of truth *if*

$$P(h|d) > P(h), \tag{1}$$

or equivalently

$$P(h|d) > P(h|{-d}), \tag{2}$$

where $-d$ represents $D \setminus \{d\}$, the elements of D that are not equal to d.

The definitions (1) and (2) are equivalent because

$$P(h) = P(h|d)P(d) + P(h|-d)P(-d).$$

Namely $P(h)$ is a weighted average of $P(h|d)$ and $P(h|-d)$. $P(h|d)$ exceeds the average, $P(h)$, if and only if it exceeds the item with which it is being averaged, $P(h|-d)$.

Definition. *Elicited probabilities $\hat{P}$ for the hypothesis-data pair (h, d) exhibit the* representativeness heuristic *if (h, d) exhibits the kernel of truth, and if*

$$\hat{P}(h|d) > \hat{P}(-h|d), \tag{3}$$

while

$$P(h|d) \leq P(-h|d). \tag{4}$$

An observation d raises the probability of h (condition (1)); however, such an observation does not go so far as to imply that h is more probable than the alternative $-h$ (condition (4)). The subject, however, states that it does (condition (3)). Condition (4) is to rule out the less interesting case where elicited probabilities and physical probabilities agree on the relative likelihood of h and its alternative.

In Example 1, the cue d is the statement that Tom is "intelligent, organized, and introverted." The hypotheses h and $-h$ are that he majors in computer science or the humanities respectively. The fact that he is introverted makes him more likely to be a computer science major than if we knew nothing about him; equivalently, computer science majors are more likely to be introverted than humanities majors. However, it does not follow that he is more likely to major in computer science than in the humanities simply because he is introverted; this ignores the base rate of humanities majors. In Example 2, the cue is the fact that the patient tests positive. This makes it more likely that the patient has the disease than if we knew nothing; it is also the case

that having the disease implies a postive test. However, because of the much higher base rate of no disease, having a positive test doesn't imply the presence of disease.[1] Finally, in Example 3, the cue is 'Number' (we will analyze this example in detail in what follows). The hypotheses are orange and blue. Knowledge that the image is a number implies that orange is more likely than if one knew nothing. It is also the case that knowing and image is orange implies it must be a number. However, knowing that an image is a number doesn't imply that it must be orange.

As the examples above illustrate, the kernel of truth is closely related to the idea that the study participants take the correct conditional probabilities and reverse them. Participants state that $P(h|d) > P(-h|d)$, whereas in fact $P(d|h) > P(d|-h)$. They believe a positive test result (d) implies a disease (h), whereas it is only the disease that implies a positive test result. To see the relation to the kernel of truth, note that $P(h|d) = P(h)\frac{P(d|h)}{P(d)}$. The kernel of truth states that $P(h|d)$ exceeds $P(h)$. This occurs if and only if $P(d|h) > P(d)$, which in turns occurs if and only if $P(d|h) > P(d|-h)$ by the logic above. Thus if $P(d|h) > P(d|-h)$, then there is a kernel of truth and vica versa.[2]

Interestingly, Tversky and Kahneman (1974) argue against the view that the agent reverses the conditioning argument. They do so by hypothesizing a Bayesian agent who has a specific confusion regarding conditioning and then showing that this form of beliefs does not explain the range of their findings. They do not, however, proposal an alternative cognitive mechanism.

[1]Formally, the example implies the kernel of truth because $P(h|d) = P(h,d)/P(d) \approx 0.001/.05 = 1/50$, whereas $P(h) = 1/1000$. Note that we have used $P(h,d) = P(h)$ under the assumption that the false negative rate $P(-d|h) = P(h,-d)/P(h) = 0$; further note $P(d) = P(h,d) + P(-h,d) = 0.001 + 0.05 \times 0.999 \approx 0.05$.

[2]Participants in the Casscells et al. (1978) give the answer of 0.95 when asked about the probability of a disease given a positive test result. This value is neither $P(h|d)$, nor is it $P(d|h)$, but rather $P(-d|-h)$, or one minus the false positive rate of $P(d|-h)$.

3 A cognitive theory of the representativeness heuristic

Here we offer a cognitive theory of the representativeness heuristic, based on *associative recognition*. The psychology literature has long studied the task of how pairs become associated. The associative recognition task is ultimately asking whether two items belong together. At first glance, the surveys above do not seem to resemble an associative recognition task: they seem to call for a probabilistic judgement, which is why prior work has emphasized concepts such as availability (Tversky and Kahneman, 1973). However, we argue that this difference is merely superficial: at a deeper level, probabilistic judgement as measured by these tasks is a matter of recognizing an association.

3.1 Model

We use the same underlying memory model as in Wachter and Kahana (2024) to capture associations, which is ultimately based in the temporal context model of Howard and Kahana (2002). Recently, the Howard and Kahana (2002) model has been generalized to associative recognition (Kahana et al., 2024). Here, we apply that generalization to modeling the representativeness heuristic.

Agents view features: basis vectors in a high dimensional vector space. These features become embedded in memory together with a temporal context, meaning an internal mental state is persistent and evolves through time. This memory matrix governs the recall process. Specifically, given an initial condition M_0, memory is comprised of outer products of contexts $\{x_i\}_{i=1}^t$ and features $\{f_j\}_{j=1}^t$:

$$M_t \equiv M_{i-1} + x_i f_i^\top \tag{5}$$

$$= M_0 + \sum_{i=1}^{t} x_i f_i^\top \tag{6}$$

Context evolves endogenously. The core assumption of the temporal context model (TCM) is that context is autoregressive (Wachter and Kahana, 2024). Here, we adopt the process:

$$x_{i+1} = x_i \text{ if } i+1 \text{ is simultaneous to } i \tag{7}$$

$$x_{i+1} \in \{x_1, \ldots, x_i\}^{\perp} \text{ otherwise} \tag{8}$$

While context is autoregressive, the agent orthogonalizes context between paired associates, and again, when receiving the cue. There are two aspects of this worth remarking on: first the identical contexts for features experienced simultaneously, and, second, the orthogonal contexts otherwise.

To address the first, note that classically, the Howard and Kahana (2002) model assumes that features of the environment (typically words in a list) are basis vectors. Others, however, have found it useful to model features in the physical environment to be norm-1 vectors more generally. The second approach allows for features to have similarities outside outside of context, itself a point of controversy, as the purest form of TCM states that all similarity should be mediated by context.

In this paper, the question of what to do with, say, a blue number, is front and center. Our main contention is that blue and number form a pair which participants learn, in a similar manner to learning pairs of words in the associative recognition task. There are subtle differences: pairs for example may be presented one directly after the other, and study participants seem to recall the specific order, whereas one cannot directly control whether the blue or number aspect of a feature is experienced first (it is not known whether it is possible to experience features truly simultaneously or whether the question is even well-defined). In this paper, we use the mechanism of autoregressive context to solve this problem. Features that are observed very close in time are experienced under contexts that have differences that are vanishingly small.[3]

[3]More precisely, in a continuous-time limit of an autoregressive context evolution equation, one might allow for Poisson arrival of features, and hence retrieved context. However, for features that are part of the same object, one could say that it is not only probability zero, but actually impossible

Because the contexts are indistinguishable, the features are tied together in context. The process defined by (7) formalizes this by saying that features experienced simultaneously are experienced under identical contexts. This is functionally equivalent to the approach taken by Kahana et al. (2024).

The second aspect of the context evolution equation, shown in (8) is that context is orthogonalized between items, so tha teach element of the list has a unique, orthogonal context. This stands in contrast to the standard model of autoregressive context in the temporal context model. Assuming orthogonal contexts greatly simplifies our analysis, allowing us to compute closed-form expressions for contextual similarity. However, it means that, for inter-item associations, the main mechanism of the temporal context model is turned off. From the perspective of explaining the representativeness heuristic, this turns out to not be problematic. As long as (1) associations can form for items that are experienced simultaneously and (2) context drives decisions in the model, it is not necessary for the paired items themselves to be connected through context. Clearly this is a convenient assumption, but is it realistic? Osth and Fox (2019) indeed show that recency and temporal contiguity, which require autoregressive context between pairs, appears to be suppressed in the associative recognition task. Kahana et al. (2024) show that these effects are still present, but they are indeed more subtle. It would be optimal for an agent who is trying to remember a pair, to orthogonalize context between a pair. From several perspectives, then, we can dispense with between-pair temporal contiguity and assume contexts are orthogonal.

Given the assumptions above, associative recognition proceeds as in Kahana et al. (2024). We use the notation x_{Cue} to denote the context retrieved by feature f_{Cue} using the memory matrix. Constants are scaled to be norm 1, and we suppress the t subscript. When confronted with a feature, the memory system retrieves a context:

$$x_{\text{Cue}} \propto M f_{\text{Cue}}.$$

for a feature to intervene; that is what it means to be simultaneous.

To recognize an association, the agent considers the context retrieved by target feature f_{TARGET}:

$$x_{\text{TARGET}} \propto M f_{\text{TARGET}}$$

and compares these retrieved contexts based on their inner products to some threshold $c_{recog} \in (0, 1)$:

$$x_{\text{CUE}} \cdot x_{\text{TARGET}} > c_{recog}$$

If the threshold is exceeded, then the Cue and Target are recognized as a pair. If there are two targets, as is the case with the representativeness heuristic, the one with the higher inner product is the one the agent chooses as the match. Retrieved contexts are scaled so as to be norm 1 according to the standard Euclidian (L^2) norm, implying that the inner product is equivalent to cosine similarity (we discuss the choice of norm below).

In this reconition-based model, as in the free recall model of Wachter and Kahana (2024), decisions are based on context x_t. Understanding how cueing works is therefore key to understanding the model. Consider what happens when an agent is cued with a feature, such as test result, a personality trait, or as in Bordalo et al. (2021), an object type. The cueing recovers all the contexts previously associated with that feature:

$$\begin{aligned} x_{\text{CUE}} = M_{t-1} f_{\text{CUE}} &\propto M_0 f_{\text{CUE}} + \sum_{s=1}^{t-1} (x_s f_s^\top) f_{\text{CUE}} \\ &\propto M_0 f_{\text{CUE}} + \sum_{s=1}^{t-1} x_s (f_s^\top f_{\text{CUE}}). \end{aligned} \tag{9}$$

Because features that are orthonormal basis vectors, $f_s^\top f_{\text{CUE}}$ equal either to zero or one. Equation 9 shows that features evoke the past contexts under which they are experienced. A context x appears in the sum in (9) if the corresponding f_s equals the current feature; otherwise it does not. Thus context is a weighted average of past contexts under which that feature was experienced. This basic contextual retrieval enabled the jump-back-in-time that drove decision making in Wachter and Kahana

(2024), and it will drive probabilistic assessments in this paper.

Before stating and proving a general implication of this model, we show how it works in an explicit calculation motivated by Example 3 above. Consider a list of three elements, a blue number 5, an orange number 3, and a blue word 'CAT.' Each becomes embedded in its own context:

$$\begin{array}{rcll} '5' & - & \text{Blue} & \leftrightarrow x_1 \\ '3' & - & \text{Orange} & \leftrightarrow x_2 \\ '\text{CAT}' & - & \text{Blue} & \leftrightarrow x_3 \end{array}$$

We model these items in the features space as:

$$\begin{array}{cccc} \textsc{Number} & \textsc{Word} & \textsc{Blue} & \textsc{Orange} \\ \updownarrow & \updownarrow & \updownarrow & \updownarrow \\ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \end{array}$$

The following represents how features are bound together with context in a partic-

ipant's mental representation of the list:

$$\underbrace{\begin{matrix}\text{Blue } '5'\\ \begin{bmatrix}0\\0\\1\\0\end{bmatrix}, \begin{bmatrix}1\\0\\0\\0\end{bmatrix}\end{matrix}}_{x_1} \quad // \quad \underbrace{\begin{matrix}\text{Orange } '3'\\ \begin{bmatrix}0\\0\\0\\1\end{bmatrix}, \begin{bmatrix}1\\0\\0\\0\end{bmatrix}\end{matrix}}_{x_2} \quad // \quad \underbrace{\begin{matrix}\text{Blue } '\text{CAT}'\\ \begin{bmatrix}0\\0\\1\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\\0\end{bmatrix}\end{matrix}}_{x_3} \tag{10}$$

The mental representation (10) implies that the context-to-features matrix equals

$$\begin{aligned} M_T^\top &= M_0^\top + \sum_{t=1}^{T} f_t x_t^\top \\ &= M_0^\top + \left(\begin{bmatrix}0\\0\\1\\0\end{bmatrix} + \begin{bmatrix}1\\0\\0\\0\end{bmatrix}\right)\begin{bmatrix}1 & 0 & 0\end{bmatrix} + \left(\begin{bmatrix}0\\0\\0\\1\end{bmatrix} + \begin{bmatrix}1\\0\\0\\0\end{bmatrix}\right)\begin{bmatrix}0 & 1 & 0\end{bmatrix} \\ &\quad + \left(\begin{bmatrix}0\\0\\1\\0\end{bmatrix} + \begin{bmatrix}0\\1\\0\\0\end{bmatrix}\right)\begin{bmatrix}0 & 0 & 1\end{bmatrix}. \end{aligned} \tag{11}$$

We abstract away from pre-experimental associations by setting all elements of M_0 to

zero. Then, suppressing time subscripts,

$$M = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} \tag{12}$$

When asked, "what is the likely color of the number?" the cue is the feature 'Number,' which retrieves context

$$x_{\text{NUMBER}} \equiv \frac{Mf_{\text{NUMBER}}}{||Mf_{\text{NUMBER}}||} \propto M \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \propto \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{bmatrix}. \tag{13}$$

Targets are 'Blue' and 'Orange':

$$x_{\text{BLUE}} \equiv \frac{Mf_{\text{BLUE}}}{||Mf_{\text{BLUE}}||} \propto \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \propto \begin{bmatrix} 1/\sqrt{2} \\ 0 \\ 1/\sqrt{2} \end{bmatrix} \tag{14}$$

and

$$x_{\text{ORANGE}} \equiv \frac{Mf_{\text{ORANGE}}}{||Mf_{\text{ORANGE}}||} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \tag{15}$$

Note how:

$$\begin{array}{ccc} x_{\text{NUMBER}} \cdot x_{\text{ORANGE}} & > & x_{\text{NUMBER}} \cdot x_{\text{BLUE}} \\ \| & & \| \\ 1/\sqrt{2} & & 1/2 \end{array}$$

Because the context of 'Number' more closely resembles that of 'Orange', the participant reports 'Orange' as the most likely color. The presence of the blue words in effect drives out the memory of the blue numbers.

Retrieved context theory, combined with the tools of recognition memory, offers an entirely novel hypothesis. Likelihoods arise from associative strengths in memories retrieved in response to a cue. The agent looks directly at the similarity with the latent context evoked by the cue. Probabilities emerge not from recall but rather from a moment of recognition.

We can use Example 3 to gain intuition for how elicited probabilities change with an increase in the sample size or the proportion of one type of item versus another as part of the observation set. When adding a list item, we add a context, and therefore a row to M:

$$M = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix}$$

Note that the last row implies that the 4th list item is a blue number. Now the following retrieved contexts are

$$x_{\text{BLUE}} \propto \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} \propto \begin{bmatrix} 1/\sqrt{3} \\ 0 \\ 1/\sqrt{3} \\ 1/\sqrt{3} \end{bmatrix}, \quad x_{\text{ORANGE}} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

whereas

$$x_{\text{NUMBER}} \propto \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \end{bmatrix} \propto \begin{bmatrix} 1/\sqrt{3} \\ 1/\sqrt{3} \\ 0 \\ 1/\sqrt{3} \end{bmatrix}$$

We see that

$$\begin{array}{ccc} x_{\text{NUMBER}} \cdot x_{\text{BLUE}} & > & x_{\text{NUMBER}} \cdot x_{\text{ORANGE}} \\ \| & & \| \\ 2/3 & & 1/\sqrt{3} \end{array}$$

Doubling the number of blue numbers tilts the balance in favor of blue numbers, though note that probabilities remain distorted.

More generally, let N be the sample size and $\mathcal{L}$ for the empirical likelihood density so that, for example, $\mathcal{L}(\textsc{Orange})N$ is the count of orange items on the list, and $\mathcal{L}(\textsc{Number}, \textsc{Orange})N$ is the count of orange numbers. Note that feature ORANGE will retrieve a context vector with $\mathcal{L}(\textsc{Orange})N$ nonzero elements. Because the context vector must have norm 1, each of these elements equals $1/\sqrt{\mathcal{L}(\textsc{Orange})N}$. Analogous results hold true for feature NUMBER. Now consider the inner product of these two context vectors. Because the inner product takes the sum of the nonzero elements, the sum is taken over at most $\min(\mathcal{L}(\textsc{Number}), \mathcal{L}(\textsc{Orange}))N$ items. In this example, it is taken over exactly $\mathcal{L}(\textsc{Orange})N$ of them, not only because there are fewer orange items, but because all orange items are numbers. If there were some orange items that were not numbers, the context overlap would be less than perfect, and there would be fewer than $\mathcal{L}(\textsc{Orange})N$ nonzero elements. That, is not all orange items occur in the number context. In both cases, however, the number of nonzero elements scales with N. Thus the inner product does not depend on N. In the specific case of orange numbers, the inner product equals $\mathcal{L}(\textsc{Number}, \textsc{Orange})/\sqrt{\mathcal{L}(\textsc{Number})\mathcal{L}(\textsc{Orange})}$, where $\mathcal{L}(\textsc{Number}, \textsc{Orange})$ is the joint likelihood of an object being both a number and the color orange.

This reasoning highlights the importance of using L^2 scaling, namely the standard Euclidean norm. For if we were to use L^1 scaling, the inner product would actually fall with the sample size, a counterintuitive outcome. Each of the elements in the context vector associated with feature ORANGE would equal $1/\mathcal{L}(\textsc{Orange})N$, whereas each item in the context vector associated with feature NUMBER equals $1/\mathcal{L}(\textsc{Number})N$. The inner product would equal $\mathcal{L}(\textsc{Number}, \textsc{Orange})/(N\sqrt{\mathcal{L}(\textsc{Number})\mathcal{L}(\textsc{Orange})})$, a decreasing function of N. While one could redefine cognitive processes through cosine similarity (namely, use not the inner product but the scaled inner product), and thus recover the same result under context vectors scaled under the L^1 norm, such an approach introduces an extra step without adding explanatory power. Note that the

original motivation for L^1 scaling, which is that the elements of the context vectors sum to 1, does not apply here.

Theorem 1. *Suppose the agent's sample matches the population such that items appear in their correct proportion. Then the agent ranks the relative probability of h given cue d according to:*

$$x_h \cdot x_d = \frac{P(h,d)}{\sqrt{P(h)P(d)}} \tag{16}$$

$$\propto \frac{P(h,d)}{\sqrt{P(h,d)+P(h,-d)}}. \tag{17}$$

The second line follows from

$$P(h)P(d) = (P(h,d)+P(h,-d))P(d),$$

and dropping the constant of proportionality $1/\sqrt{P(d)}$.

While (16) doesn't sum to 1 when summed over h, we can still use it to gain intuition for elicited probabilities $\hat{P}$ and how they differ from physical probabilities P. Note that

$$\hat{P}(h|d) \approx \frac{P(h,d)}{\sqrt{P(h)P(d)}} = P(h|d)\sqrt{\frac{P(d)}{P(h)}}$$

Because $P(d)$ falls out as an item of interest, probabilities will be distorted when the two hypotheses have notably different base rates $P(h)$.

A second consequence of the theorem is hypothesis h_1 is chosen over h_2

$$x_{h_1} \cdot d > x_{h_2} \cdot d$$

if and only if

$$P(h_1|d)\frac{1}{\sqrt{P(h_1)}} > P(h_2|d)\frac{1}{\sqrt{P(h_2)}}. \tag{18}$$

Suppose h_1 and h_2 are the only two possible hypotheses ($h_2 = -h_1$). Then we can

rewrite (18) as

$$P(h_1|d) > P(h_2|d)\sqrt{\frac{1-P(h_2)}{P(h_2)}}$$

We apply this to Example 3, because in that case the researchers have some degree over the database and the cue. Moreover, Bordalo et al. (2021) provide a control condition (discussed further below) to adjust for any prior differences in memory that may be contributing to the results.

We apply (17), first using the values from the blue treatment: 10 orange numbers, 15 blue numbers, and 25 blue words. Recall that $h = \textsc{Blue}$ and $d = \textsc{Number}$. In particular:

$$x_{\textsc{Blue}} \cdot x_{\textsc{Number}} = \frac{P(\textsc{Blue}, \textsc{Number})}{\sqrt{P(\textsc{Blue})P(\textsc{Number})}} = \frac{15}{\sqrt{15+25}\sqrt{15+10}} = 0.47 \tag{19}$$

$$x_{\textsc{Orange}} \cdot x_{\textsc{Number}} = \frac{P(\textsc{Orange}, \textsc{Number})}{\sqrt{P(\textsc{Orange})P(\textsc{Number})}} = \frac{10}{\sqrt{10}\sqrt{15+10}} = 0.63 \tag{20}$$

There are more blue numbers, but their probability is penalized by the large number of blue words.

3.2 Questions regarding probability and number

The above discussion maps statements about relative likelihood into the paired associates task. However, for the range of data we are interested in matching, it will be useful to show how the model can produce explicitly probabilistic statements.

One view is that agents remember instances and sample from memory. This view, which invokes free recall, seems unnecessarily complicated. Some tasks, moreover, seem to make this difficult, as study participants, would, for example, have to store individual instances of blue or orange shapes. Perhaps these study participants count these shapes as part of an experiment, but if this were systematically the case, they would not exhibit the observed bias both in their probability judgements and in their estimated counts of objects.

More likely is that participants use the similarities as discussed above, along with either the information that probabilities must sum to one, or a specific probability if that is part of a cue. An example of the latter is in the Casscells et al. (1978) survey where participants answer that those who test positive have a disease with 95% probability; 5% being given to them as the false-positive rate. In the former case, the psychology literature has a long-standing method, known as the Luce choice rule.

Let $F(\cdot)$ be an increasing continuous function defined on non-negative real numbers. Then the Luce choice rule gives

$$\hat{P}(h|d) = \frac{F(x_h \cdot x_d)}{\sum_{h' \in \mathcal{H}} F(x_{h'} \cdot x_d)} \tag{21}$$

If the question pertains to the number of instances, the agents estimate a quantity $\hat{N}$ and then apply $\hat{P}(h|d)\hat{N}$. To summarize:

1. Judgements regarding whether which pair is most likely, or rankings according to pairs are made using cosine similarity of retrieved context vectors.

2. Direct probability assessments are made using cosine similarity, combined with a Luce choice rule (unless the agent is specifically cued with a probability).

3. Numerical assessments are made using direct probability assessments, combined with knowledge about how many items were on the list.

Note that this theory binds questions 2 and 3 with question 1. In practice, these questions may be asked sequentially, in which case the participant's own response to question 1 will create a data point, itself influencing questions 2 and 3 along the liens described above.

Given these assumptions, probabilistic judgement tasks with well-defined alternatives become the classic paired associates task.

4 Discussion

4.1 Comparison with Bhatia (2017)

For many years, the view of representativeness heuristic and related phenomena were shaped by the survey results of the Kahneman and Tversky studies mentioned above. These results measured cognitive erros in settings that aimed to mimic day-to-day decisions. They also form a striking and memorable narrative of their own. However, they come with limitations. For one, it is not possible to separate out whether the confusion is due to a faulty database or the recall process itself. This point emerges forcefully in the work of Bhatia (2017), who shows that the RH can emerge from social interactions, reinforced by repeated exposures to certain images in the media (actively promoting a simplified view of reality).

The advantage of the experiments in Bordalo et al. (2021) is the control they have over the conditions, including the agent's database. These results point to a cognitive theory. In the next subsection, we show how our model can account for the full range of these results.

4.2 Application to Bordalo et al. (2021)

Bordalo et al. (2021) overcome the limitations of these prior studies by building an experimental framework based on abstract images with no prior associations. Study 1 forms the basis of Example 3. For ease of exposition, and important detail was left out. It is possible that any difference between the reporting of blue and orange might be due to something other than the experiment. While the set-up of Bordalo et al. (2021) allows for some control of the database, of course it is not possible to fully control what database participants enter the experiment with. Thus Bordalo et al. randomly assign participants to one of two treatments. Each participant sees a sequence of 50 of these abstract images. In the first condition, they see 10 orange numbers, 15 blue numbers, and 25 gray shapes (the *gray* treatment). In the second condition, they see 10 orange

numbers, 15 blue numbers, and 25 blue words (the *blue* treatment). Participants are then asked several questions:

Q1. "An image was randomly drawn from the images that were just shown to you. The chosen image showed a number. What is the likely color of the chosen image?"

Q2. "An image was randomly drawn from the images that were just shown to you. The chosen image showed a number. What is the probability the number is orange?"

Q3. "How many orange numbers were shown to you?"

Q4. "'How many blue numbers were shown to you?"

The model in Bordalo et al. (2021) makes the following predictions:

Prediction 1. *The blue treatment reduces the assessed likelihood that the randomly drawn number is blue in Q1 and Q2 relative to the gray treatment.*

Prediction 2. *The blue treatment reduces the share of blue numbers recalled in Q3 and Q4 relative to the gray treatment.*

Note that a Bayesian model would predict no difference between the gray and the blue treatments. The above discussion shows how the model accommodates the results of the first study, which shows that participants

4.2.1 Study 1

In Study 1, Bordalo et al. (2021) found that participants in the blue treatment were significantly more likely to state that the likely color of a randomly-drawn number was orange (50% versus 30%). They also reported a higher probability of the number being orange.

We first show a simplified example of the gray treatment, in which all 25 blue words are replaced by gray shapes. We map these items in the features space by replacing words with shapes and adding a basis vector for gray:

$$\begin{array}{ccccc} \textsc{Number} & \textsc{Shape} & \textsc{Blue} & \textsc{Orange} & \textsc{Gray} \\ \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow \\ \begin{bmatrix}1\\0\\0\\0\\0\end{bmatrix} & \begin{bmatrix}0\\1\\0\\0\\0\end{bmatrix} & \begin{bmatrix}0\\0\\1\\0\\0\end{bmatrix} & \begin{bmatrix}0\\0\\0\\1\\0\end{bmatrix} & \begin{bmatrix}0\\0\\0\\0\\1\end{bmatrix} \end{array}$$

with corresponding mental representation:

$$\begin{array}{ccccc} \text{Blue } '5' & // & \text{Orange } '3' & // & \text{Gray } \square \\ \underbrace{\begin{bmatrix}0\\0\\1\\0\\0\end{bmatrix}, \begin{bmatrix}1\\0\\0\\0\\0\end{bmatrix}} & & \underbrace{\begin{bmatrix}0\\0\\0\\1\\0\end{bmatrix}, \begin{bmatrix}1\\0\\0\\0\\0\end{bmatrix}} & & \underbrace{\begin{bmatrix}0\\0\\0\\0\\1\end{bmatrix}, \begin{bmatrix}0\\1\\0\\0\\0\end{bmatrix}} \\ x_1 & & x_2 & & x_3 \\ \begin{bmatrix}1\\0\\0\end{bmatrix} & & \begin{bmatrix}0\\1\\0\end{bmatrix} & & \begin{bmatrix}0\\0\\1\end{bmatrix} \end{array} \tag{22}$$

Applying the same reasoning as in (11) implies that the memory matrix equals

$$M^\top = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

with

$$M = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix} \tag{23}$$

Participants are asked: "what color is the number?" The cue is number. Retrieved number context is identical to (13). The target is color. Each color retrieves a basis vector context:

$$x_{\text{Blue}} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad x_{\text{Orange}} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad x_{\text{Gray}} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

The inner product with the context vector now gives the correct ranking – namely equal likelihood for blue and for orange.

Substituting the actual values of the experiment, in which the 25 blue words are replaced by 25 gray shapes, we find:

$$x_{\text{Blue}} \cdot x_{\text{Number}} = \frac{P(\text{Blue}, \text{Number})}{\sqrt{P(\text{Blue})P(\text{Number})}} = \frac{15}{\sqrt{15}\sqrt{15+10}} = 0.77$$

with $x_{\text{Orange}} \cdot x_{\text{Number}}$ unchanged from (20). Under the gray treatment, the agent correctly recalls a greater likelihood of the number being blue.

4.2.2 Study 2

Bordalo et al. (2021) then examine what happens to probability assessment when they introduce orange words to the list.

Prediction 3. *The assessed probability that a random number is orange in Q1 and Q2 and the share of recalled orange numbers in Q3 and Q4 decrease with the number of orange words k in the decoy distribution.*

As with Predictions 1 and 2, a Bayesian model produces no effect.

To build intuition regarding what our model has to say about this manipulation, we add one orange word:

$$\begin{aligned} '5' &- \text{Blue} \leftrightarrow x_1 \\ '3' &- \text{Orange} \leftrightarrow x_2 \\ '\text{CAT}' &- \text{Blue} \leftrightarrow x_3 \\ '\text{DOG}' &- \text{Orange} \leftrightarrow x_4 \end{aligned}$$

with corresponding representation:

$$\underbrace{\begin{matrix}\text{Blue } '5'\end{matrix}}_{\begin{bmatrix}0\\0\\1\\0\end{bmatrix},\begin{bmatrix}1\\0\\0\\0\end{bmatrix}} // \underbrace{\begin{matrix}\text{Orange } '3'\end{matrix}}_{\begin{bmatrix}0\\0\\0\\1\end{bmatrix},\begin{bmatrix}1\\0\\0\\0\end{bmatrix}} // \underbrace{\begin{matrix}\text{Blue } '\text{CAT}'\end{matrix}}_{\begin{bmatrix}0\\0\\1\\0\end{bmatrix},\begin{bmatrix}0\\1\\0\\0\end{bmatrix}} // \underbrace{\begin{matrix}\text{Orange } '\text{DOG}'\end{matrix}}_{\begin{bmatrix}0\\0\\0\\1\end{bmatrix},\begin{bmatrix}0\\1\\0\\0\end{bmatrix}} \tag{24}$$

$$\begin{matrix}x_1\\\begin{bmatrix}1\\0\\0\\0\end{bmatrix}\end{matrix} \quad \begin{matrix}x_2\\\begin{bmatrix}0\\1\\0\\0\end{bmatrix}\end{matrix} \quad \begin{matrix}x_3\\\begin{bmatrix}0\\0\\1\\0\end{bmatrix}\end{matrix} \quad \begin{matrix}x_4\\\begin{bmatrix}0\\0\\0\\1\end{bmatrix}\end{matrix}$$

Following the reasoning of (11), we have:

$$M^\top = \left(\begin{bmatrix}0\\0\\1\\0\end{bmatrix} + \begin{bmatrix}1\\0\\0\\0\end{bmatrix}\right)\begin{bmatrix}1&0&0&0\end{bmatrix} + \left(\begin{bmatrix}0\\0\\0\\1\end{bmatrix} + \begin{bmatrix}1\\0\\0\\0\end{bmatrix}\right)\begin{bmatrix}0&1&0&0\end{bmatrix}$$
$$+ \left(\begin{bmatrix}0\\0\\1\\0\end{bmatrix} + \begin{bmatrix}0\\1\\0\\0\end{bmatrix}\right)\begin{bmatrix}0&0&1&0\end{bmatrix} + . \left(\begin{bmatrix}0\\0\\0\\1\end{bmatrix} + \begin{bmatrix}0\\1\\0\\0\end{bmatrix}\right)\begin{bmatrix}0&0&0&1\end{bmatrix}$$

so that

$$M^\top = \begin{bmatrix}1&1&0&0\\0&0&1&1\\1&0&1&0\\0&1&0&1\end{bmatrix}$$

Note that the first three columns of $M^\top$ are the same as in Study 1, as are the first 3 rows of M:

$$M = \begin{bmatrix}1&0&1&0\\1&0&0&1\\0&1&1&0\\0&1&0&1\end{bmatrix}. \tag{25}$$

As in Study 1, participants are asked "What is the likely color of the number?" The cue, feature 'Number,' retrieves context

$$x_{\textsc{Number}} \propto M\begin{bmatrix}1\\0\\0\\0\end{bmatrix} = \begin{bmatrix}1\\1\\0\\0\end{bmatrix} \propto \begin{bmatrix}1/\sqrt{2}\\1/\sqrt{2}\\0\\0\end{bmatrix}. \tag{26}$$

The answer can only be blue or orange. We find the following retrieved contexts:

$$x_{\text{Blue}} = \begin{bmatrix} 1/\sqrt{2} \\ 0 \\ 1/\sqrt{2} \\ 0 \end{bmatrix}, \quad x_{\text{Orange}} = \begin{bmatrix} 0 \\ 1/\sqrt{2} \\ 0 \\ 1/\sqrt{2} \end{bmatrix}$$

Now, orange and blue are equally good matches for the cue 'Number', consistent with the findings of Bordalo et al. (2021):

$$x_{\text{Number}} \cdot x_{\text{Orange}} \quad = \quad x_{\text{Number}} \cdot x_{\text{Orange}}$$

That is, the bias reduced (in the case of this numerical example, all the way to zero).

4.2.3 Study 3

Study 3 focuses in on the importance of the cue. The idea is to add an attribute – in this case size – that orange numbers have and blue numbers don't. If the agent were thinking probabilistically, then asking about an item's size versus color would yield the same bias – after all it is the same underlying items that are being asked about. Bordalo et al. (2021) make the following prediction:

Prediction 4. *The share or recalled orange/large numbers and the assessment of the probability that a random number is orange/large is higher in the color treatment than in the size treatment.*

To implement this study, we first introduce features to represent size.

$$\text{Number} \quad \text{Word} \quad \text{Blue} \quad \text{Orange} \quad \text{Small} \quad \text{Large}$$

$$\updownarrow \qquad \updownarrow \qquad \updownarrow \qquad \updownarrow \qquad \updownarrow \qquad \updownarrow$$

$$\begin{bmatrix}1\\0\\0\\0\\0\\0\end{bmatrix} \quad \begin{bmatrix}0\\1\\0\\0\\0\\0\end{bmatrix} \quad \begin{bmatrix}0\\0\\1\\0\\0\\0\end{bmatrix} \quad \begin{bmatrix}0\\0\\0\\1\\0\\0\end{bmatrix} \quad \begin{bmatrix}0\\0\\0\\0\\1\\0\end{bmatrix} \quad \begin{bmatrix}0\\0\\0\\0\\0\\1\end{bmatrix}$$

In Bordalo et al. (2021), the blue numbers are small, but everything else is large. Therefore, consider the following list and corresponding mental representation:

$$\underbrace{\begin{bmatrix}0\\0\\0\\0\\1\\0\end{bmatrix}, \begin{bmatrix}0\\0\\1\\0\\0\\0\end{bmatrix}, \begin{bmatrix}1\\0\\0\\0\\0\\0\end{bmatrix}}_{\text{Small, Blue } '5'} \;//\; \underbrace{\begin{bmatrix}0\\0\\0\\0\\0\\1\end{bmatrix}, \begin{bmatrix}0\\0\\0\\1\\0\\0\end{bmatrix}, \begin{bmatrix}1\\0\\0\\0\\0\\0\end{bmatrix}}_{\text{Large, Orange } '3'} \;//\; \underbrace{\begin{bmatrix}0\\0\\0\\0\\0\\1\end{bmatrix}, \begin{bmatrix}0\\0\\1\\0\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\\0\\0\\0\end{bmatrix}}_{\text{Large, Blue } '\text{CAT}'} \; 1 \qquad (27)$$

$$\underbrace{\begin{bmatrix}1\\0\\0\end{bmatrix}}_{x_1} \quad \underbrace{\begin{bmatrix}0\\1\\0\end{bmatrix}}_{x_2} \quad \underbrace{\begin{bmatrix}0\\0\\1\end{bmatrix}}_{x_3}$$

and the context-to-features matrix adds rows corresponding to the new features

$$M^\top = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

whereas M adds columns

$$M = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix} \tag{28}$$

Because the first column is unaffected, context retrieved by feature 'Number' is the same as (13):

$$x_{\textsc{Number}} \propto M f_{\textsc{Number}} \propto \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{bmatrix}. \tag{29}$$

When asked, "what is the likely color of the number," the participant once again compares similarities of retrieved contexts of blue and orange to (29), obtaining the same answers as in Study 1:

$$\begin{array}{ccc} x_{\textsc{Number}} \cdot x_{\textsc{Orange}} & > & x_{\textsc{Number}} \cdot x_{\textsc{Blue}} \\ \| & & \| \\ 1/\sqrt{2} & & 1/2 \end{array}$$

and thus the percent of participants answering that orange is unchanged from Study 1.

If the participants were, however, remembering more orange numbers, one might think they might also remember more large numbers, since the orange numbers are large

whereas the blue (forgotten) numbers are small. However, this is not what occurs.

When asked "what is the likely size of the number?" the participant retrieves contexts associated with small and with large

$$x_{\text{SMALL}} \propto M f_{\text{SMALL}} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

and

$$x_{\text{LARGE}} \propto M f_{\text{LARGE}} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \propto \begin{bmatrix} 0 \\ 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$

and therefore

$$x_{\text{SMALL}} \cdot x_{\text{NUMBER}} > x_{\text{LARGE}} \cdot x_{\text{NUMBER}}$$

implying that the participant is more likely to recall a small than a large number, whereas the participant is less likely to recall a blue than an orange number, in line with the Bordalo et al. (2021) results.

4.3 Comparison with the Bordalo et al. (2021) model

Bordalo et al. (2021) posit that the agent assesses the probability of h using a Luce choice rule:[4]

$$\tilde{P}(h \mid d) = \frac{S(h,d)}{\sum_{h' \in \mathcal{H}} S(h',d)}, \tag{30}$$

where S is a similarity function that Bordalo et al. directly specify based on the true probability density:

$$S(h,d) = P(h,d)^{\alpha} P(-h,d)^{-\beta} P(h,-d)^{-\gamma}, \tag{31}$$

[4]We slightly deviate from their notation here for convenience. The original paper specifies that $\tilde{P}(h \mid d) \propto e^{S(h,d)}$ and then defines $S(h,d)$ as the log of (31).

where Bordalo et al. (2021) note that when h is takes on a total of two values, it is not possible to separately identify β and γ, and for this reason set $\beta = 0$.

Bordalo et al. (2021) further specialize to the case of $\alpha = 1 + \gamma$, implying that

$$\tilde{P}(h \mid d) \propto P(h \mid d) \left[\frac{P(h \mid d)}{P(h \mid -d)} \right]^{\gamma} \tag{32}$$

which reduces to the case of Bayesian updating for $\gamma = 0$. In contrast, interference by alternative data $-d$ occurs for $\gamma > 0$. Bordalo et al. (2021) note that this model makes 4 predictions, as described above, each of which is born out in the data. We can readily verify that our model makes the same four predictions because it shares key characteristics with (32): $\tilde{P}(h|d)$ is increasing in $P(h|d)$ and decreasing in $P(h|-d)$. These emerge naturally from (17).

Technically, this statement is true for statements for which comparing the inner products is sufficient: namely simple recognition statements regarding whether one hypothesis is more likely than another. When eliciting an actual probability or number, the Luce choice rule is required. Then probabilities will be functions of (16), not (17), for which the comparative static is less obvious. However, for power specifications of $F(\cdot)$, the term $P(d)$ is separable and drops out of the analysis, implying the same analysis as when determining when one hypothesis is more likely than another. Even when $F(\cdot)$ does not take the power form, but is otherwise increasing, the result continues to hold because the comparative static holds $P(d)$ constant.

In the Bordalo et al. (2021) model, γ is a flexible parameter, introduced for the purpose of fitting the representativeness heuristic. By contrast, our model provides a cognitive foundation for γ, and, in so doing, eliminates γ entirely as a free parameter.[5] It also connects two previously disparate fields of study: the study of recognition, and the study of probability assessment. By drawing this connection, our hope is to stimulate further study in both areas.

[5]More precisely, the model offers a cognitive foundation for why $P(h, -d)$ would enter into the probability assessment of $P(h|d)$. Besides offering this cognitive foundation with a more specific functional form, it also implies a *different* functional form, which could potentially be verified experimentally.

Table 1: **Lists of possibilities in the Tversky and Kahneman (1983) conjunction experiment**

This table lists the possibilities given to participants in the conjunction experiment of Tversky and Kahneman (1983). See the main text for details about the experiment.

Possibilities for Linda	Possibilities for Bill
Linda is a teacher in elementary school.	Bill is a physician who plays poker for a hobby.
Linda works in a bookstore and takes Yoga classes.	Bill is an architect.
Linda is active in the feminist movement. (F)	Bill is an accountant. (A)
Linda is a psychiatric social worker.	Bill plays jazz for a hobby. (J)
Linda is a member of the League of Women Voters.	Bill surfs for a hobby.
Linda is a bank teller. (T)	Bill is a reporter.
Linda is an insurance salesperson.	Bill is an accountant who plays jazz for a hobby. (A&J)
Linda is a bank teller and is active in the feminist movement. (T&F)	Bill climbs mountains for a hobby.

4.4 Application to the Conjunctive Fallacy

Example 4 (Tversky and Kahneman (1983))**.** *Subjects are given a description of an outspoken, single woman who is passionate about social-justice issues. They report that it is less likely that she is a bank teller than that she is both a bank teller and a feminist, even though this is logically impossible. Similarly, when given a description of a mathematical but unimaginative man, subjects report that it is less likely that he is a jazz player than that he is an accountant and a jazz player.*

In this experiment, participants read a description of a woman named Linda and a man named Bill. About Linda, they read: "Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations." About Bill, they read: "Bill is 34 years old. He is intelligent, but unimaginative, compulsive, and generally lifeless. In school, he was strong in mathematics but weak in social studies and humanities." Participants were then asked to rank, in order of likelihood, the possibilities given in Table 1. Over 80% of participants rank, from more to less likely, F>T&F>T for Linda and A>A&J>J for Bill. Regard-

less of beliefs, this likelihood ranking is a logical impossibility. Tversky and Kahneman (1983) refer to the tendency of subjects to rank the conjunction as more likely than one of the constituents as the *conjunctive fallacy*.

Here we show how the model can capture the conjunctive fallacy of Tversky and Kahneman (1983). Consider subjects asked to rank possibilities for Bill (see Table 1). We model the relative rankings of Accountant (A), Jazz player (J), and both (A&J).

While Bill as presented in the Tversky and Kahneman (1983) experiment may be a composite feature, we keep the notation simply by assuming it is a basis vector, denoted f_{Bill}. We assume f_{Bill} tends to be seen along with f_{A} and only rarely with f_{J}. Meanwhile, f_{J} is seen frequently in different contexts. As in Section 4.2, we consider the very simplest list that illustrates these ideas:

$$
\begin{array}{ccc}
\textsc{Bill} & \textsc{Accountant (A)} & \textsc{Jazz player (J)} \\
\updownarrow & \updownarrow & \updownarrow \\
\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}
\end{array}
$$

$$
\begin{array}{ccccc}
\text{Bill who is an accountant} & // & \text{Jazz player} & // & \text{Accountant} \\
\underbrace{\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}} & & \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} & & \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \\
x_1 & & x_2 & & x_3 \\
\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} & & \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} & & \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}
\end{array}
\tag{33}
$$

The mental representation (33) implies that the context-to-features matrix equals:

$$\begin{aligned} M_T^\top &= M_0^\top + \sum_{t=1}^{T} f_t x_t^\top \\ &= M_0^\top + \left(\left[\begin{array}{c} 1 \\ 0 \\ 0 \end{array} \right] + \left[\begin{array}{c} 0 \\ 1 \\ 0 \end{array} \right] \right) \left[\begin{array}{ccc} 1 & 0 & 0 \end{array} \right] + \left(\left[\begin{array}{c} 0 \\ 0 \\ 1 \end{array} \right] \right) \left[\begin{array}{ccc} 0 & 1 & 0 \end{array} \right] \\ &\quad + \left(\left[\begin{array}{c} 0 \\ 1 \\ 0 \end{array} \right] \right) \left[\begin{array}{ccc} 0 & 0 & 1 \end{array} \right] . \end{aligned} \tag{34}$$

Abstracting from M_0 and suppressing time subscripts gives us:

$$M^\top = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

with

$$M = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

The cue is the feature 'Bill':

$$x_{\text{Bill}} \equiv \frac{M f_{\text{Bill}}}{||M f_{\text{Bill}}||} \propto \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix},$$

which picks up the single context in which Bill is seen. The targets are Accountant

together with Jazz, Accountant, and Jazz):

$$x_{\text{A\&J}} \equiv \frac{M(f_{\text{A}} + f_{\text{J}})}{\|M(f_{\text{A}} + f_{\text{J}})\|} = \begin{bmatrix} 1/\sqrt{3} \\ 1/\sqrt{3} \\ 1/\sqrt{3} \end{bmatrix} \tag{35}$$

$$x_{\text{J}} \equiv \frac{Mf_{\text{J}}}{\|Mf_{\text{J}}\|} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \tag{36}$$

$$x_{\text{A}} \equiv \frac{Mf_{\text{A}}}{\|Mf_{\text{A}}\|} = \begin{bmatrix} 1/\sqrt{2} \\ 0 \\ 1/\sqrt{2} \end{bmatrix} \tag{37}$$

From which we see that

$$x_{\text{Bill}} \cdot x_{\text{A}} > x_{\text{Bill}} \cdot x_{\text{A\&J}} > x_{\text{Bill}} \cdot x_{\text{J}}$$

Rather than assessing probabilities through traditional inference (which would lead, through the laws of probability, to the chance of Accountant and Jazz player being less than or equal to Jazz player, participants assess probabilities through context matches with a target. Because Accountant and Jazz player contains within it the word "Accountant," it brings up the context associated with Accountant. While not as good a match as Accountant on its own, it is still better than Jazz on its own.

4.5 Comparison with Wachter and Kahana (2024)

Here, we offer a cognitive mechanism to explain the RH, building on our earlier work (Wachter and Kahana, 2024). Wachter and Kahana take the view that lifetime portfolio decisions were best modeled as free recall, in which the agent simulates future states by freely recalling prior experiences, with weights determined by present context and by the associations between features and context, In contrast, this paper is based on

recognition. Indeed, a main insight of this paper is that probabilistic judgement is best thought of as *recognition*, a subject in which there is a long literature (Kahana et al., 2005).[6] Unlike a free recall or serial recall task that involves the production of words (and, secondarily, a decision as to whether to speak the word), the probabilistic judgement task involves no production – the cue provides the possibilities– and only a judgement as to which is more likely. While probabilistic judgement is not modeled in the memory literature directly, it shares a fundamental similarity with recognition memory tasks, and most specifically, associative recognition (Kahana et al., 2005).

Ultimately, as one turns back to real-world decisions, the question of which of our decisions are best thought of as free recall and which of recognition still lies open. Regardless, we take the view that these two paradigms can be united under the common theme of associative memory.

Key to our modeling assumptions is that free recall is *not* required – the agent need not generate any features. This has an important implication for interpreting these experiments. Cases in which agents are asked to rank the likelihood of outcomes, as is standard in the Kahneman and Tversky methodology, require no production – one can simply evaluate the best match with the target. Cases in which one is producing probabilities from a given set of possibilities (or the possibilities are obvious because there is a small number and they are repeated many times) can be handled with the Luce choice rule (see Equation 21). However, as Bordalo et al. (2023) emphasize, if asked to produce a probability and one is not given the possibilities, then free recall inevitably comes into play, and the probability of any given item may be biased upward or downward based on what the presence of the features does to cue or suppress items during free recall.[7]

[6] As discussed above, a key question in a probabilistic judgement task whether the difficulty in recalling the alternatives, as Bordalo et al. (2023) emphasize, plays a role.

[7] For example, if one were to apply the Luce choice rule to examine probabilities related to the conjunctive fallacy (which Tversky and Kahneman (1983) do not do), it would seem most natural to divide by the elements in the relevant column of Table 1, rather than the implied set of possibilities (e.g. accountant but not a jazz player, accountant and a surfer but not a jazz player).

4.6 General results

Consider for simplicity the special case in which the agent observes the population distribution, so that $\mathcal{L}(\cdot) = P(\cdot)$. Recall that$P(h, d)$ refers to the density of individuals sharing features d and h, $P(-h, d)$ refers to the density of individuals sharing d and not sharing h, and analogously for $P(h, -d)$.

$$\hat{P}(h|d) \quad \propto \quad F\left(\frac{P(h,d)}{\sqrt{P(h,d)+P(h,-d)}}\right) \tag{38}$$

Note that F is strictly increasing. It follows that $\hat{P}(h|d)$ is strictly increasing in $P(h, d)$ and decreasing in $P(h, -d)$.[8] It then follows from arguments in (Bordalo et al., 2021, Appendix A) that the four predictions hold.

We now turn to the predictions of Bordalo et al. (2021). Prediction 1 states that the subject is less likely to state that the color of the number is blue (Q1), and that the probability of the number is blue (Q2) in the blue treatment as compared to the gray treatment.

This stylized model does not capture many realistic aspects of the recall process. For example, it assumes prior associations in M that may influence answers. It also assumes that the agent fully resets context between observations, which is also not realistic. However, while these assumptions may affect the level of the response, they are much less likely to affect the differential between the blue and gray treatments. Here, the model is clear, the difference between the inner products is significant under the blue treatment, whereas under the gray treatment they are equal.

While comparing inner products determines which color serves as a better match, and hence which is recognized as appearing with number, the inner products do not sum to 1. Thus to determine elicited probabilities, we assume the agent uses some version of the Luce choice rule. It suffices that $F(\cdot)$ be increasing to obtain the result

[8]The same result holds true even without separability, assuming that this comparative static holds $P(d)$ constant.

that the likelihood of blue numbers is lower under the blue treatment as compared with the gray treatment. This shows the model satisfies Prediction 1.

The explanation for Prediction 2 is similar. Rather than free-recalling each individual object, it is more likely that participants estimate the number of objects (in this case 50), and apply the probability estimate from the Luce choice rule in Prediction 1. The finding that the share of numbers that are blue or orange correlates strongly with the corresponding likelihood further supports this interpretation.

Note that the stylized version of the model does not contain a mechanism for individual heterogeneity, and is not calibrated to match levels of recall in either the gray or blue treatments. With this caveat in mind, we can nonetheless perform a back-of-the-envelope calibration of $F(\cdot)$, assuming it takes the form of a power function:

$$\hat{P}(h \,|\, d) \propto \left(\frac{P(h,d)}{\sqrt{P(h)P(d)}} \right)^{\eta}, \tag{39}$$

for some parameter $\eta > 0$. We can calibrate η using Study 1 of Bordalo et al. (2021), who find that participants in the blue treatment estimated a 7-percentage-point higher probability of orange than did those in the gray treatment. In both the blue and gray treatments, there are 10 orange numbers and no orange words, so (39) implies

$$\hat{P}(\textsc{Orange} \,|\, \textsc{Number}) \propto \left(\frac{P(\textsc{Orange}, \textsc{Number})}{\sqrt{P(\textsc{Orange})P(\textsc{Number})}} \right)^{\eta} = \left(\frac{10}{\sqrt{10}\sqrt{15+10}} \right)^{\eta}. \tag{40}$$

In the gray treatment, there are 15 blue numbers and no blue words, so

$$\hat{P}_{\text{gray}}(\textsc{Blue} \,|\, \textsc{Number}) \propto \left(\frac{P(\textsc{Blue}, \textsc{Number})}{\sqrt{P(\textsc{Blue})P(\textsc{Number})}} \right)^{\eta} = \left(\frac{15}{\sqrt{15}\sqrt{15+10}} \right)^{\eta}. \tag{41}$$

And in the blue treatment, there are 15 blue numbers and 25 blue words, so

$$\hat{P}_{\text{blue}}(\textsc{Blue} \,|\, \textsc{Number}) \propto \left(\frac{P(\textsc{Blue}, \textsc{Number})}{\sqrt{P(\textsc{Blue})P(\textsc{Number})}} \right)^{\eta} = \left(\frac{15}{\sqrt{15+25}\sqrt{15+10}} \right)^{\eta}. \tag{42}$$

To calibrate η, we match $\hat{P}_{\text{blue}}(\textsc{Orange} \mid \textsc{Number}) - \hat{P}_{\text{gray}}(\textsc{Orange} \mid \textsc{Number}) = 0.07$, or

$$\frac{(10/\sqrt{10})^\eta}{(10/\sqrt{10})^\eta + (15/\sqrt{40})^\eta} - \frac{(10/\sqrt{10})^\eta}{(10/\sqrt{10})^\eta + (15/\sqrt{15})^\eta} = 0.07. \tag{43}$$

Solving for η implies the estimate

$$\eta \approx 0.57.$$

We now turn to Prediction 3, which states the the assessed probability that a random number is orange decreases with the number of orange words k in the decoy distribution (see also Section 4.2.2 which discusses Bordalo et al. (2021) Study 2, designed to address this prediction). We can easily see the model satisfies this prediction by using the formula (32). In particular,

$$x_{\textsc{Blue}} \cdot x_{\textsc{Number}} = \frac{P(\textsc{Blue}, \textsc{Number})}{\sqrt{P(\textsc{Blue})P(\textsc{Number})}} = \frac{15}{\sqrt{15+25-k}\sqrt{15+10}}$$

and

$$x_{\textsc{Orange}} \cdot x_{\textsc{Number}} = \frac{P(\textsc{Orange}, \textsc{Number})}{\sqrt{P(\textsc{Orange})P(\textsc{Number})}} = \frac{10}{\sqrt{10+k}\sqrt{15+10}}$$

At around $k = 5.4$, the model no longer predicts that participants will recall that orange is more likely. Indeed, Bordalo et al., Figure 2 shows that once $k = 6$, a majority no longer states that orange is likely, and once $k = 22$, only 30% of participants do. Again, one should use caution in interpreting the absolute levels of recall. The model clearly satisfies Prediction 3 in that the inner products converge and then pass each other as orange words replace blue words in the distribution.

Prediction 4 contrasts the *color* treatment with a *size* treatment, and relies on the results of Bordalo et al. (2021) Study 3 i (see Section 4.2.3). Study 3 shows a sequence of 50 images as before. Except in this case, the images vary according to a third attribute, name size. All orange numbers are large, and all blue numbers are small, while all words are blue and large.

The color treatment consists of the Q1 and Q2 defined above. The size treatment

replaces Q1 with "What is the likely font size of a randomly drawn number?" and "What is the probability of a randomly drawn number being large?" This question presents a direct contrast with Bayesian inference, in that all the large numbers are orange and all the small numbers are blue. Thus, under unbiased recall, the actual objects elicited by the questions in the size and the color treatments are identical. However, as Section 4.2.3 shows, the cue determines which items the agent recalls.

Specifically, the analysis for the color treatment is the same as above, whereas the analysis for the size treatment has $d = \textsc{Large}$ as opposed to $d = \textsc{Blue}$. Thus:

$$x_{\textsc{Large}} \cdot x_{\textsc{Number}} = \frac{P(\textsc{Large}, \textsc{Number})}{\sqrt{P(\textsc{Large})P(\textsc{Number})}} = \frac{10}{\sqrt{10+25}\sqrt{15+10}}$$

and

$$x_{\textsc{Small}} \cdot x_{\textsc{Number}} = \frac{P(\textsc{Smal}, \textsc{Number})}{\sqrt{P(\textsc{Smal})P(\textsc{Number})}} = \frac{15}{\sqrt{15}\sqrt{15+10}}$$

Now the large blue words prevent recall of the large orange numbers (which are fewer in any case). According to the model, the agent concludes that Small is a better match for Number, and thus (accurately) reports numbers as more likely to be small. The data support this prediction. Participants are significantly more likely to assess "orange is likely" in the color treatment (40%), than "large is likely" in the size treatment (17%). This analysis highlights the importance of the cue. It is a result that is difficult to reconcile with any kind of Bayesian analysis, but straightforward to understand with the tools of cued recall.

4.7 Finance application: Long-term earnings expectations

There is a mass of firms, indexed i. Each firm has a long-term growth rate, or "type," $\lambda_i \in \{\lambda_\ell, \lambda_h\}$, where $0 < \lambda_\ell < \lambda_h < 1$. From period t to period $t+1$, a firm's earnings will either experience high growth or low growth; without loss of generality, we say that this growth rate $g_{i,t+1}$ is either 0 or 1. The latent type λ_i determines the probability

of high growth:

$$g_{i,t+1} = \begin{cases} 1 & \text{with probability } \lambda_i, \\ 0 & \text{otherwise.} \end{cases}$$

Suppose that, for a young firm i, an analyst has observed T years of data $\{g_{i,t+1}\}_{t=0}^{T-1}$. The analyst needs to forecast the long-term growth rate $\mathbb{E}_t[g_{i,t+\tau}] = \lambda_i$ using these data in order to value the firm's stock.

To form this forecast, the analyst looks at a historical dataset of long-lived firms' earnings. Specifically, for each firm, the analyst looks at the historical T-period growth $G_{iT} \in \{0, \ldots, T\}$ and the subsequent long-run growth $\sum_{t=T}^{T-1+\tau} g_{i,t+1}$ for some large τ. Because τ is large and growth is i.i.d., $\sum_{t=T}^{T-1+\tau} g_{i,t+1} \approx \lambda_i \tau$, so this is effectively the same as having observed λ_i. In feature space, these short- and long-run growth rates can be represented as

$$\begin{array}{ccccccc} G_{iT} = 0 & \cdots & G_{iT} = T & \lambda_i = \lambda_h & \lambda_i = \lambda_\ell \\ \updownarrow & \cdots & \updownarrow & \updownarrow & \updownarrow \\ \begin{bmatrix} 1 \\ \vdots \\ 0 \\ 0 \\ 0 \end{bmatrix} & \cdots & \begin{bmatrix} 0 \\ \vdots \\ 1 \\ 0 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ 1 \end{bmatrix} \end{array}$$

The pairs $\{G_{iT}, \lambda_i\}$ are observed simultaneously for each firm, so each row of the memory matrix M will encode the pair that was realized for that firm.

Now suppose, given this encoded memory, that the investor observes a firm with data $G_{iT} = \bar{G}$ and must use this information to infer a long-run growth rate λ_i. The cued context is determined using the basis vector corresponding to $G_{it} = \bar{G}$, and the target contexts are determined using the last two basis vectors. We can immediately

apply our general results to see that

$$\hat{P}(\lambda_i|\{g_{i,t+1}\}) \propto F\left(\frac{P(\lambda_i, \{g_{i,t+1}\})}{\sqrt{P(\lambda_i)P(\{g_{i,t+1}\})}}\right).$$

The agent infers that the firm is more likely to be high-growth than low-growth if

$$\frac{P(\lambda_h, \{g_{i,t+1}\})}{\sqrt{P(\lambda_h)}} > \frac{P(\lambda_\ell, \{g_{i,t+1}\})}{\sqrt{P(\lambda_\ell)}},$$

which is more likely to be true if high-growth firms are rare ($P(\lambda_h)$ is small).

Now consider how the investor's expectations evolve upon seeing more data. Specifically, suppose that after τ more periods, the agent has seen growth $G_{i,T+\tau}$. As τ becomes large, the true conditional probability $P(\lambda_h|\{g_{i,t+1}\})$ will tend to 0 or 1, and so too will the subjective probability $\hat{P}(\lambda_i|\{g_{i,t+1}\})$. Hence, the model predicts that expectation errors will tend to correct, and, in turn, realized stock returns will tend to be predictably lower for firms which ex ante were expected to have higher long-term growth rates.

Note that this logic also applies to the time-series of growth for the aggregate market (or any single stock or portfolio). One can think of λ_i as a time-varying latent state, and G_{iT} as the most recent earnings data from which the analyst can infer λ_i. If λ_i is persistent, then the analyst may associate high recent growth with high future long-term growth, and hence have excessively high expectations of future growth.

References

Angeletos, G.-M. and La'O, J. (2009). Incomplete information, higher-order beliefs and price inertia. *Journal of Monetary Economics*, 56:S19–S37.

Barberis, N., Greenwood, R., Jin, L., and Shleifer, A. (2015). X-CAPM: An extrapolative capital asset pricing model. *Journal of Financial Economics*, 115(1):1–24.

Barberis, N., Shleifer, A., and Vishny, R. (1998). A model of investor sentiment. *Journal of Financial Economics*, page 37.

Bhatia, S. (2017). Associative judgement and vector space semantics. *Psychological Review*, 124(1):1–20.

Bordalo, P., Coffman, K., Gennaioli, N., Schwerter, F., and Shleifer, A. (2021). Memory and representativeness. *Psychological Review*, 128(1):71–85.

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

Bordalo, P., Conlon, J. J., Gennaioli, N., Kwon, S. Y., and Shleifer, A. (2023). Memory and probability. *The Quarterly Journal of Economics*, 138(1):265–311.

Bordalo, P., Gennaioli, N., Porta, R. L., and Shleifer, A. (2019a). Diagnostic expectations and stock returns. *The Journal of Finance*, 74(6):2839–2874.

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

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

Bordalo, P., Gennaioli, N., Shleifer, A., and Terry, S. J. (2019b). Real credit cycles. Working paper, Boston University, Harvard University, Universita Bocconi, University of Oxford.

Burnside, C., Eichenbaum, M., and Rebelo, S. (2016). Understanding booms and busts in housing markets. *Journal of Political Economy*, 124(4):1088–1147.

Casscells, W., Schoenberger, A., and Graboys, T. B. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine*, 299(18):999–1001.

Enke, B. and Zimmermann, F. (2017). Correlation neglect in belief formation. *The Review of Economic Studies*, 86(1):313–332.

Estes, W. K. (1972). Research and theory on the learning of probabilities. *Journal of the American Statistical Association*, 67(337):81–102.

Estes, W. K. (1976). The cognitive side of probability learning. *Psychological Review*, 83(1):37–64.

Gennaioli, N. and Shleifer, A. (2018). *A Crisis of Beliefs: Investor Psychology and Financial Fragility*. Princeton University Press, Princeton, NJ.

Howard, M. W. and Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, 46(3):269–299.

Kahana, M. J., Jin, B. J., and Halpern, D. J. (2024). A theory of memory for items and associations. Unpublished paper, University of Pennsylvania.

Kahana, M. J., Rizzuto, D. S., and Schneider, A. (2005). Theoretical correlations and measured correlations: Relating recognition and recall in four distributed memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(5):933–953.

Kahneman, D. and Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3(3):430–454.

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

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

Nagel, S. and Xu, Z. (2018). Asset pricing with fading memory. Working paper, University of Chicago and University of Michigan.

Osth, A. F. and Fox, J. (2019). Are associations formed across pairs? a test of learning by temporal contiguity in associative recognition. *Psychonomic Bulletin & Review*, 26(6):1650–1656.

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.

Wachter, J. A. and Kahana, M. J. (2024). A retrieved-context theory of financial decisions*. *The Quarterly Journal of Economics*, 139(2):1095–1147.