Associative learning and representativeness*

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

December 6, 2025

Abstract

Across varied experimental settings, subjects determine the probability of a hypothesis according to the representativeness heuristic, a striking departure from Bayesian updating. Rather than assessing the odds of a hypothesis given data simply by using the likelihood multiplied by the prior, subjects discount the odds based on the probability that the hypothesis might have been generated by some other data, which is irrelevant. We explain these results in a tractable cognitive model grounded in fundamental principles of associative memory and contextual retrieval. The model reproduces the central experimental regularities associated with the representativeness heuristic, including the conjunction fallacy. We then show how the same retrieval mechanism helps account for several important financial-market anomalies, illuminating how distorted probability judgments can propagate into asset prices and ultimately affect the real economy.

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

*We thank Sudeep Bhatia, Pedro Bordalo, Spencer Kwon, Alice Healy, Peter Maxted, Nikolai Roussanov, John Sakon, Yao Zeng, and participants at the Society of Financial Studies Cavalcade, at the Spring 2024 Memory Beliefs and Choice Meeting, at Carnegie Mellon University, London Business School, and the University of Pennsylvania 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. kahana@psych.upenn.edu.

‡Stanford University, Graduate School of Business. jparon@stanford.edu.

§Department of Finance, The Wharton School, University of Pennsylvania, and NBER. jwachter@wharton.upenn.edu

1 Introduction

Our memories of past experiences guide not only our thoughts but also our choices. We choose to save when we think about 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 economic choice theories 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., 2018, 2019a, 2020; Nagel and Xu, 2018; Mullainathan, 2002; Wachter and Kahana, 2024).

Memory takes as its starting place not the rational agent, but the fundamental forces of recency, similarity, and temporal contiguity. Yet economics and psychology for many years traveled together because of a basic fact about cognition: if a pair of words, concepts, or objects frequently occur together, they will form an association, and the more they co-occur the stronger that association will be. Memory should enable agents to learn about reality, which means accurately learning conditional probability distributions. Work such as Estes (1972, 1976) emphasized the ability of memory theories to reproduce Bayesian findings. Among the first to question the resulting consensus was Kahneman and Tversky (1972, 1973) and Tversky and Kahneman (1973), who showed, in a series of experiments, that agents do not follow Bayesian updating. The core aspect of their findings, reproduced in a tighter setting by Bordalo et al. (2021), is that even if subjects observe many pairs $(-h, d)$, they will remember instead (h, d), if they also learn $(-h, -d)$.

We explain this finding with a cognitive theory. Our work is based on the retrieved context theory of Howard and Kahana (2002), and thus follows the basic principles of associative memory and contextual retrieval. However, their model is based on the free recall paradigm. Free recall tends to lead to the basic result observed by Estes (1976), and also derived in Wachter and Kahana (2024). Observations that occur

more frequently will tend to be judged as having higher probability. The point of the representativeness heuristic is that this compelling intuition sometimes fails, and it fails in a predictable way. The key step in our model is treating the probabilistic judgement problem not as a process of recalling instances from memory, but rather as a flash of recognition. Think of recognition as a handshake between a cue and a target. The subject or agent determines, based on the cue, whether the cue and target go together. This determination is not based on the frequency with which they are observed together but whether they were observed under the same context. In other words, it is the similarity of the retrieved contexts that determines the recognition of an association between two objects. We model probabilistic judgement as an associative recognition problem, amenable to the same modeling tools as that well-studied experimental paradigm. We also show how our theory can explain additional findings, such as the conjunction fallacy, and how it can account for financial market anomalies such as excess volatility and the value premium.

Our paper relates to multiple strands of literature. Recent work by Bordalo et al. (2021) casts light back to the work of Kahneman and Tversky (1972), who conceived of non-Bayesian probabilistic judgement as arising from heuristics, one of which they termed "representativeness." Over time, representativeness appeared to merge with another heuristic, which Tversky and Kahneman (1973) term "availability." This merging suggests that Kahneman and Tversky thought of probabilistic judgement as in part a model of memory.[1] Their idea was that enhanced retrievability of certain instances boosts probability judgement for the category to which these instances belong. This idea that retrievability is linked with probability judgement was not unique to Tversky and Kahneman, and does not, by itself, explain the failure of Bayesian updating.

A second strand of literature is in economics and finance. Early work in behavioral finance recognized the importance of the representativeness heuristic (Barberis et al.,

[1]While Tversky and Kahneman (1973) consider availability as a heuristic distinct from representativeness, and later cite these two as separate when developing a taxonomy of heuristics (Tversky and Kahneman, 1974), Tversky and Kahneman (1983) merge the concepts, as do Gennaioli and Shleifer (2010). The advantage of a cognitive theory is that it can offer a unified theory for various heuristics.

1998). Later, building on Gennaioli and Shleifer (2010), Bordalo et al. (2016) develop a model, based on a probability distortion, that can account for the key experimental finding: the mistaken application of the co-occurence of $-h$ and $-d$ (in the notation above). This model became the basis for diagnostic expectations (Bordalo et al., 2018), applied to a wide range of phenomena such as the cross-section of stock returns (Bordalo et al., 2019a), financial crises (Gennaioli and Shleifer, 2018), and why aggregate real investment rates vary over time (Bordalo et al., 2019b). Nonetheless diagnostic expectations is an ad hoc model. Our model produces a cognitive explanation for the same patterns.

Our model is most closely related to that of Bhatia (2017). Building on Kahneman and Tversky (1972)'s basic insight that frequency judgements arise from similarity, Bhatia uses word-embedding models to create a vector representation for each word in the experiment. He then shows that similarity-based judgements match the proximity of the words to each other in a high-dimensional vector space. Frequency of word usage, as captured by the word-embedding model, determines the location of the words and their proximity. His work traces the representativeness heuristic back to information that emerges in language based on common usage. In contrast, we trace the heuristic back to a model of memory, one that is also based on proximity in a vector space. Notably, it can be used in novel situations where researchers already control for any prior language based similarities, as in Bordalo et al. (2021). Bhatia's results emerge as a special case of our model.

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 other behavioral mechanisms? Such mechanisms include extrapolative expectations (Barberis et al., 2015), imperfect common knowledge (Angeletos and La'O, 2009), social influences on beliefs (Burnside et al., 2016), and correlation neglect (Enke and Zimmermann, 2017). These mechanisms may be tied to an evolving mental state that is based on a series of imperfectly retrieved

perceptual data.

The remainder of the paper is organized as follows. Section 2 gives a formal definition of the heuristic and connects our definition with related concepts in the literature. Section 3 describes a cognitive model of the representativeness heuristic, and fits data from recent experimental results. Section 4 extends the model to account for related phenomena such as the conjunction fallacy and puzzling phenomena in finance and macroeconomics. Section 5 concludes.

2 Defining the representativeness heuristic

According to Kahneman and Tversky (1972), "A person who follows [the representativeness] heuristic evaluates the probability of an uncertain event, or a sample, by the degree to which it is: (i) similar in essential properties to its parent population; and (ii) reflects the salient features of the process by which it is generated." Perhaps recognizing the limitations of this definition, the authors also note that representativeness is "easier to assess than to characterize." The experiments in the 1972 paper highlight misperceptions of randomness and that subjects struggle to compute posterior probabilities from sample proportions. These experiments convincingly showed that subjects depart from Bayesian reasoning, in part because they made use of settings where the Bayesian outcome was clear. However, the settings were contrived, and clearly differed from those in which people make decisions relevant to their lives using knowledge obtained from experience.

It was later, in their 1983 study, that Tversky and Kahneman wrote that what makes a hypothesis h representative of data d is both the co-occurence of h and d, and the lack of co-occurence of h and alternative data $-d$. The latter is fundamentally non-Bayesian. In the meantime, in 1973, the authors put forward the experiment that elucidated the concept of representativeness as it is now understood.

Example 1 (Kahneman and Tversky (1973))**.** *Subjects are given a brief description*

of an intelligent, organized, and introverted student and are asked to rank several fields of graduate study in order of likelihood. The subjects tend to state that the student is more likely to study computer science than the humanities, even though there are many more students in the humanities than in computer science.

The authors asked one group of study participants to estimate the base rates of academic subjects in nine fields. A second group received a detailed description of an introverted man (Tom W.) and was 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. Participants judged the base rate of humanities as far higher than that of computer science, but both the similarity and the likelihood of computer science as higher than those of the humanities.

A more recent experiment provides another example of representativeness:

Example 2 (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.*

Crucially, both examples draw on subjects' experience of events. In the prior example, answers from the first group of participants (the group asked to estimate the base case) proxies for that experience. In the second, the experiment itself produces the experience. In neither case are the results reflective of an inability to solve a math problem, but rather capture something about everyday life.

In both cases, subjects' mistaken assessments have a *kernel of truth* (Bordalo et al., 2016). Computer science students are more likely to be introverted than the general population. 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 report incorrect probabilities, their inference does not seem *that* irrational — we can see ourselves making a similar mistake. One senses that these examples are getting at a specific cognitive deficit to which humans are subject, analogous to an optical illusion.

To formalize these notions, define the set Ω to be the population of interest. Each element of Ω consists of a pair of features (for example, being a number and being blue, or being an introvert and being a salesman). In the two experiments above, one of the elements of the pair forms the cue. Following Bordalo et al. (2021), we refer to the set of possible cues as $\mathscr{D} = \{d_1, \ldots, d_K\}$. 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 $\mathscr{H} = \{h_1, \ldots, h_J\}$. Let P be the physical probability measure over $\Omega = \mathscr{H} \times \mathscr{D}$.

The experiment or survey asks questions pertaining to probabilities. Our theory is that subjects' answers need not reflect a probability measure in the traditional sense. The psychological theory implicit in many models of economic decision-making is that an agent carries a probability measure in their head and updates it in a manner that either approximates the laws of probability or departs from these laws in defined ways. Our view is different: the agent responds to cues, and it is the associations the agent carries in their head and updates. The laws of associations, not the mathematical laws of probability, govern the updating process. Nonetheless, it is convenient to use the notation $\hat{P}$ to denote elicited probabilities, keeping in mind that $\hat{P}$ need not have the properties of a probability measure.

We now formulate a general statement of the kernel of truth and the representativeness heuristic. Our definitions are based on those of Bordalo et al. (2016).

Definition. *Given a physical probability measure P,the hypothesis-data pair (h, d) ex-*

hibits 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* $\mathcal{D} \setminus \{d\}$*, the elements of* $\mathcal{D}$ *that are not equal to* d*.*

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

$$P(h) = P(h|d)P(d) + P(h|{-d})P(-d). \tag{3}$$

That is, $P(h)$ 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 averaged, $P(h|{-d})$.[2] Bordalo et al. (2016) refer to the ratio $P(h|d)/P(h|{-d})$ as the *representativeness of data* d *for hypothesis* h.

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*

$$\hat{P}(h|d) > \hat{P}(-h|d), \tag{4}$$

while

$$P(h|d) \leq P(-h|d). \tag{5}$$

An observation d increases the probability of h (condition (1)); such an observation does not go so far as to imply that h is more probable than the alternative $-h$ (condition (5)). However, the subject states that it does (condition (4)). Condition (5) is to rule out the less interesting case where the elicited probabilities and physical probabilities agree on the relative likelihood of h and its alternative.

[2] We use the notation $-h, h$ to denote generic distinct elements of $\mathcal{H}$. In contrast, we use $-d$ to denote $\mathcal{D} \setminus \{d\}$. This removes the need for unnecessary subscripts. In all of the examples we consider, $\mathcal{D}$ consists of two elements, which implies that the notation is consistent across $\mathcal{D}$ and $\mathcal{H}$. However, nothing in the analysis requires that $\mathcal{D}$ consist of only two elements.

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, 'Number' forms the cue. The hypotheses are orange and blue. The 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 that an image is orange implies that it must be a number. However, knowing that an image is a number does not imply that it must be orange.

The puzzle can be stated in both probabilistic and cognitive terms. Probabilistically, suppose one is attempting to measure the likelihood of hypothesis h given data d. What one wants is the posterior probability of h given d, and to rank this probability across various hypotheses. Because the distribution of d does not matter, this is equivalent to the joint likelihood — that is, the joint occurrence — of h and d. Alternatively, it corresponds to the likelihood of observing d conditional on h, multiplied by the prior probability of h. In this Bayesian analysis, the distribution of h conditional on $-d$ plays no role. Once the agent has computed the probability of a joint occurence of h and d, what happens when $-d$ occurs is completely irrelevant.

In cognitive terms, a memory system that stores joint occurrences of events should retrieve the hypothesis most frequently paired with the cue. If d co-occurs more often with $-h$ than with h, then cueing on d should preferentially retrieve $-h$ rather than h; the strength of the d and $-h$ "handshake" ought to exceed that of d and h. Yet the data show this is not what happens: irrelevant memories associated with $-d$ contaminate the connection between d and $-h$.

Kahneman and Tversky (1972) appeal to "representativeness" and "availability": h is (mistakenly) judged more probable than $-h$ because h is representative of d. But

why should h, and not $-h$, be representative of d? Why should h be more available than $-h$? Bordalo et al. (2016) noted that the key missing ingredient in the literature is the explicit role of $P(h|-d)$, which has no normative role in Bayesian inference.

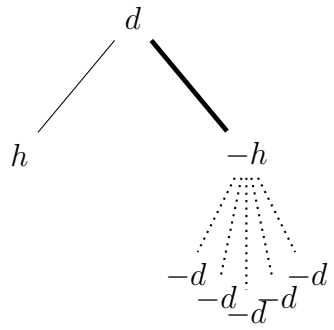


Figure 1: Conditions likely to produce the representativeness heuristic. The agent observes many pairs $(-h, d)$, represented by the dark line, and fewer pairs (h, d). Observations of $-h$ also frequently co-occur with $-d$, but observations of h do not.

3 A cognitive theory of the representativeness heuristic

Here, we offer a cognitive theory of the representativeness heuristic. Our theory is based on *associative recognition*, one of the main experimental paradigms in the literature on human memory (Hockley, 1992; Osth and Dennis, 2024).

In the associative recognition task, the subjects memorize a list of pairs. Following a delay, subjects are presented with a pair and asked if the pair was on the list. The term associative recognition arises from the need for subjects to recognize the association between one member of the pair and the other. The set of pairs presented to subjects is known as *paired associates*. Paired associates are used not only in associative recognition, but also in cued recall, in which one item of the pair cues the subject to try to recall the second item. Note that cued recall, like free recall, requires production.

At first glance, there may seem to be little relation between associative recognition and representativeness. The experiments described in Section 2 require a probabilis-

tic judgment, not recognition of whether two items form a pair. However, we argue that this difference is merely superficial. At a deeper level, probabilistic judgment as measured by these tasks is a matter of answering the question: do two traits belong together or not?

3.1 Model

We start with the Wachter and Kahana (2024) model of associations, which is ultimately based on the temporal context model of Howard and Kahana (2002). The Howard and Kahana model accounts for data from the *free recall* task. In free recall, subjects learn a list of words and then remember the words in any order. Free recall has proven to be a highly useful paradigm in memory theory because it sheds light on how people transition from one thought to another. The literature on human learning and memory highlights three major laws: similarity, recency, and temporal contiguity. Similarity refers to the priority accorded to information that is similar to the presently active features, recency refers to priority given to recently experienced features, and temporal contiguity refers to the priority given to features that share a history of co-occurrence with the presently active features. These laws determine what information get priority in memory. Using free recall, it is possible to isolate the temporal contiguity effect: subjects exhibit better memories for items that occurred close in time to a just-recalled item. This effect, originally conjectured by Aristotle, has broad and deep empirical support.[3]

Wachter and Kahana (2024) model lifetime portfolio decisions. According to this model, 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, the probabilistic judgment task that is the focus of this paper is best thought of as *recognition*. Recognition and recall are distinct: recall requires production on the part of the subject; recognition does not. Recently, Jin et al. (2024)

[3]See Wachter and Kahana (2024, Figure 2).

adapt the Howard and Kahana (2002) model to recognition memory, and show how it can explain a range of experimental phenomena. We combine elements of Jin et al. with Wachter and Kahana.

While similarity, recency, and temporal contiguity are all present both in free recall and in associative recognition, they are not present to the same degree. Specifically, Osth and Fox (2019) show that recency and temporal contiguity effects are muted in associative recognition compared with how they appear in free recall. However, subjects clearly bind a pair together in memory, suggesting a strong form of temporal contiguity intra-pair. These findings inform our modeling decisions.

As in Wachter and Kahana (2024), agents view features f_i: basis vectors in a high-dimensional vector space. These features become embedded in memory together with a temporal context, an internal mental state that evolves through time. Context, through the memory matrix, governs what comes to mind. Specifically, given an initial condition M_0, memory is comprised of outer products of features and contexts $\{f_i, x_i\}_{i=1}^n$:

$$\begin{aligned} M_n &\equiv M_{n-1} + x_i f_i^\top && (6) \\ &= M_0 + \sum_{i=1}^{n} x_i f_i^\top. && (7) \end{aligned}$$

To understand the implications of (7), say that the environment presents the agent with features f_{CUE}. These features cue context $\boldsymbol{x}_{\text{CUE}}$:

$$\boldsymbol{x}_{\text{CUE}} \propto M_n f_{\text{CUE}}, \tag{8}$$

where context is determined up to a scalar multiple. Substituting (8) into (7) yields the following:

$$\boldsymbol{x}_{\text{CUE}} \propto M_0 f_{\text{CUE}} + \sum_{i=1}^{n} x_i (f_i^\top f_{\text{CUE}}). \tag{9}$$

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_i is equal to

the current feature; otherwise it does not. Thus, retrieved 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.

To capture the strength of intra-pair associations and the relative lack of inter-pair associations, we assume that items viewed (nearly) simultaneously are viewed under the same context, and those viewed at different times are viewed under orthogonal contexts. To operationalize this, we assume that context only evolves if something distracts the agent. If there is no distraction, then context remains the same. If there is a distraction, by definition, that introduces orthogonal features, and context jumps to a point that is orthogonal to current context:

$$x_{i+1} \begin{cases} = & x_i & \text{if no distractor,} \\ \in & \{x_1, \ldots, x_i\}^{\perp} & \text{otherwise.} \end{cases}, \tag{10}$$

where $\{x_1, \ldots, x_i\}^{\perp}$ denotes the set of vectors orthogonal to previous contexts. This assumption on contextual evolution significantly simplifies our analysis. It accounts for strong intra-pair associations because pairs of attributes, such as 'Blue' and 'Number,' occur under the same context and are retrieved together. It accounts for the relative lack of inter-pair associations, because it lacks a contextual link between pairs. If our purpose was to advance the modeling of human memory, we would retain this link at the cost of a much more complicated model, because memory data supports inter-pair contextual linkage (just less than in free recall). For this reason, Jin et al. (2024), incorporate autoregressive context into their model.[4] The effect of inter-pair temporal

[4]The data on free recall and associative recognition can be reconciled by allowing contextual persistence to vary over time. The data strongly suggest such time-variation. It has long been known that context can be "reset," through the use of distractor tasks, and the resetting of context has become a standard element of experimental design (Healey et al., 2019). Resetting of contexts implies that the autoregressive coefficient is temporarily equal to zero. To memorize a pair or a set of features experienced together, it may be optimal for a subject to deliberately orthogonalize a context. The idea that a subject can do this for themselves may be surprising, but neural data suggests it is possible (Manning et al., 2016).

contiguity, however, is unlikely to be first-order when accounting for representativeness, which is our focus.

Allowing intra-pair characteristics to be observed under the same context solves a technical problem: how to retain the simplicity of features as basis vectors while accounting for objects having multiple attributes (for example, being both blue and a number). In our framework, 'Blue' and 'Number' form a pair which subjects learn, in a similar manner to learning pairs of words in the associative recognition task. 'Blue' and 'Number' are each basis vectors. Subjects build a *composite representation* (Metcalfe, 1991) using autoregressive context. The only similarity in the model comes through items sharing a context.

In the Jin et al. (2024), extension of the Howard and Kahana (2002) model, whether recognition occurs or not depends on *contextual similarity*. The cue f_{Cue} retrieves context $\boldsymbol{x}_{\text{Cue}}$, (8). To recognize an association, the agent considers the context retrieved by target feature f_{Target}:

$$\boldsymbol{x}_{\text{Target}} \propto M f_{\text{Target}}$$

(suppressing the subscript n). The agent compares these retrieved contexts based on their inner products to some threshold criterion $c_{\text{recog}} \in (0, 1)$:

$$\boldsymbol{x}_{\text{Cue}} \cdot \boldsymbol{x}_{\text{Target}} > c_{\text{recog}} \tag{11}$$

(Stanislaw and Todorov, 1999). If the inner product exceeds the threshold, 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 Euclidean (L^2) norm, implying that the inner product is equivalent to cosine similarity (we discuss the choice of norm below). In Section 3.4, we generalize the model to allow an error term to combine with (11) to determine recognition. We now summarize the model.

Model Summary . *Suppose an agent views n elements of $\mathcal{H} \times \mathcal{D}$ as feature vectors f_i, $i = 1, \ldots, n$. That is, f_i is a composite representation*

$$f_i = f_h + f_d, \quad h \in \mathcal{H}, d \in \mathcal{D}, \tag{12}$$

where f_h and f_d are basis vectors. Feature vectors f_i combine with orthogonal contexts x_i to form memory M as in (7), with M_0 the zero matrix. Let f_j, $j \in \mathcal{H} \cup \mathcal{D}$, represent either a cue ($d \in \mathcal{D}$) or a target ($h \in \mathcal{H}$). When cued with d, the agent recognizes target h as more likely than some other target $-h$ if

$$\boldsymbol{x}_h \cdot \boldsymbol{x}_d > \boldsymbol{x}_{-h} \cdot \boldsymbol{x}_d$$

where $\boldsymbol{x}_j$, $j \in \mathcal{H} \cup \mathcal{D}$, is a norm-1 retrieved context vector such that $\boldsymbol{x}_j \propto M f_j$.

To understand how this model accounts for the representativeness heuristic, consider the following simplified version of Example 2. There is 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 feature space as follows:

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

The following represents how features are bound together with context in a participant's mental representation of the list:

$$\underbrace{\begin{matrix}\text{Blue } '5' \\ \begin{bmatrix}0\\0\\1\\0\end{bmatrix}, \begin{bmatrix}1\\0\\0\\0\end{bmatrix}\end{matrix}}_{x_1} \quad // \quad \underbrace{\begin{matrix}\text{Orange } '3' \\ \begin{bmatrix}0\\0\\0\\1\end{bmatrix}, \begin{bmatrix}1\\0\\0\\0\end{bmatrix}\end{matrix}}_{x_2} \quad // \quad \underbrace{\begin{matrix}\text{Blue } '\text{CAT}' \\ \begin{bmatrix}0\\0\\1\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\\0\end{bmatrix}\end{matrix}}_{x_3} \tag{13}$$

and this results in the memory matrix

$$\begin{aligned} M_n &= M_0 + \sum_{i=1}^{n} x_i f_i^\top \\ &= M_0 + \begin{bmatrix}1\\0\\0\end{bmatrix}\left(\begin{bmatrix}0\\0\\1\\0\end{bmatrix}+\begin{bmatrix}1\\0\\0\\0\end{bmatrix}\right)^\top + \begin{bmatrix}0\\1\\0\end{bmatrix}\left(\begin{bmatrix}0\\0\\0\\1\end{bmatrix}+\begin{bmatrix}1\\0\\0\\0\end{bmatrix}\right)^\top + \begin{bmatrix}0\\0\\1\end{bmatrix}\left(\begin{bmatrix}0\\0\\1\\0\end{bmatrix}+\begin{bmatrix}0\\1\\0\\0\end{bmatrix}\right)^\top \\ &= M_0 + \begin{bmatrix}1\\0\\0\end{bmatrix}\begin{bmatrix}1&0&1&0\end{bmatrix} + \begin{bmatrix}0\\1\\0\end{bmatrix}\begin{bmatrix}1&0&0&1\end{bmatrix} + \begin{bmatrix}0\\0\\1\end{bmatrix}\begin{bmatrix}0&1&1&0\end{bmatrix} \end{aligned} \tag{14}$$

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

zero. Then, suppressing subscripts,[5]

$$M = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}. \tag{15}$$

The cue is the feature 'Number,' which retrieves context

$$\boldsymbol{x}_{\text{NUMBER}} \equiv \frac{M f_{\text{NUMBER}}}{||M f_{\text{NUMBER}}||} \propto M \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \propto \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{bmatrix}. \tag{16}$$

Targets are 'Blue' and 'Orange':

$$\boldsymbol{x}_{\text{BLUE}} \equiv \frac{M f_{\text{BLUE}}}{||M f_{\text{BLUE}}||} \propto \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \propto \begin{bmatrix} 1/\sqrt{2} \\ 0 \\ 1/\sqrt{2} \end{bmatrix} \tag{17}$$

and

$$\boldsymbol{x}_{\text{ORANGE}} \equiv \frac{M f_{\text{ORANGE}}}{||M f_{\text{ORANGE}}||} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}. \tag{18}$$

Note how

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

Because the context of 'Number' more closely resembles that of 'Orange,' the participant reports 'Orange' as the more likely color. The presence of the blue words, in

[5]More realistically, the dimensionality of features and context spaces is probably on the order of 10^7 or more, with 10^7 considered a lower bound on neurons in the brain involved in memory storage. We consider a submatrix of M, and the corresponding subvector of features and context.

effect, drives out the memory of the blue numbers.

Before stating general properties of the model, we develop intuition with two additional examples. First, we consider a list that adds one more blue number. When we lengthen the "experiment" above by one blue number, we add one more row to the block of M that we are considering:

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

Now, the following retrieved contexts are

$$\boldsymbol{x}_{\text{Blue}} \propto \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} \propto \begin{bmatrix} 1/\sqrt{3} \\ 0 \\ 1/\sqrt{3} \\ 1/\sqrt{3} \end{bmatrix}, \quad \boldsymbol{x}_{\text{Orange}} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

whereas

$$\boldsymbol{x}_{\text{Number}} \propto \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \end{bmatrix} \propto \begin{bmatrix} 1/\sqrt{3} \\ 1/\sqrt{3} \\ 0 \\ 1/\sqrt{3} \end{bmatrix}$$

We see that

$$\begin{array}{ccc} \boldsymbol{x}_{\text{Number}} \cdot \boldsymbol{x}_{\text{Blue}} & > & \boldsymbol{x}_{\text{Number}} \cdot \boldsymbol{x}_{\text{Orange}} \\ \| & & \| \\ 2/3 & & 1/\sqrt{3} \end{array}$$

This example suggests that inner products have the reasonable property that, as instances of a pair increase, the likelihood of recognition, as measured by the inner product, also increases.

Second, we replace the blue words with gray shapes. In the Bordalo et al. (2021) experiment, the effect disappears. A simple calculation explains why. We first add a feature vector,

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

and then assume the following simple list:

$$\begin{array}{ccccc} \text{Blue } '5' & // & \text{Orange } '3' & // & \text{Gray } \square \\ \underbrace{\begin{bmatrix}0\\0\\1\\0\\0\end{bmatrix}, \begin{bmatrix}1\\0\\0\\0\\0\end{bmatrix}} & & \underbrace{\begin{bmatrix}0\\0\\0\\1\\0\end{bmatrix}, \begin{bmatrix}1\\0\\0\\0\\0\end{bmatrix}} & & \underbrace{\begin{bmatrix}0\\0\\0\\0\\1\end{bmatrix}, \begin{bmatrix}0\\1\\0\\0\\0\end{bmatrix}} \\ x_1 & & x_2 & & x_3 \\ \begin{bmatrix}1\\0\\0\end{bmatrix} & & \begin{bmatrix}0\\1\\0\end{bmatrix} & & \begin{bmatrix}0\\0\\1\end{bmatrix} \end{array} \tag{19}$$

Applying the same reasoning as in (14) implies that the memory matrix equals

$$M = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix} \tag{20}$$

'Number' again forms the cue. The cue retrieves the context (16). The targets are each of the two possible colors. Each color retrieves a basis vector context:

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

The inner product with the context vector $\boldsymbol{x}_{\text{Number}}$ now gives the correct ranking — namely, equal likelihood for blue and orange.

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 to the latent context evoked by the cue. Probabilities do not emerge from the process of recall, but rather from an instant of recognition.

The assumptions we have made above are not innocuous. There are situations where previous experience, captured in M_0 becomes important. One place is in creating subject-specific variation, which we incorporate below through a noise term. A second case arises when one of the targets is a better match for previous associations than for the list. For example, Bordalo et al. (2023) present subjects with two lists, one containing animals and names and the other containing animals and other heterogeneous words. In both cases, the animals made up 40% of the list. Subjects are then asked, as in the Example 2 above, "Suppose the computer randomly chose a word from the words you just saw. What is the percent chance that it is an animal?" They show that subjects report a significantly higher probability that the word is an animal

when the list contains animals and names versus when the list contains animals and heterogenous words. If the cue is "word on the list," then "animal" and "name" are approximately equally good targets. On the other hand, the target "heterogeneous words" would most likely be a greater match for a prior list context, which in our model would be captured by pre-existing associations in M_0.

3.2 Properties

We now formally prove the properties suggested by the simple examples above.

Theorem 1. *Suppose that the agent views an unbiased sample of elements of $\mathcal{H} \times \mathcal{D}$. Assuming the cognitive process described in the Model Summary, the agent ranks the relative probability of targets $h \in \mathcal{H}$ given cue $d \in \mathcal{D}$ according to:*

$$\boldsymbol{x}_h \cdot \boldsymbol{x}_d = \frac{P(h,d)}{\sqrt{P(h)P(d)}} \tag{21}$$

$$\propto \frac{P(h,d)}{\sqrt{P(h,d)+P(h,-d)}}. \tag{22}$$

Thus, for elements $h, -h \in \mathcal{H}$, inner products satisfy

$$\boldsymbol{x}_h \cdot \boldsymbol{x}_d > \boldsymbol{x}_{-h} \cdot \boldsymbol{x}_d \tag{23}$$

if and only if

$$P(h|d) > P(-h|d)\sqrt{\frac{P(h)}{P(-h)}}. \tag{24}$$

In the case where $\mathcal{H} = \{h, -h\}$, namely $\mathcal{H}$ consists of only two elements, inner products satisfy (23) if and only if

$$P(h|d) > P(-h|d)\sqrt{\frac{P(h)}{1-P(h)}}. \tag{25}$$

Proof. It follows from (7) and the features definition (12) that, when cued by $f_j, j \in$

$\mathcal{H} \cup \mathcal{D}$,

$$\boldsymbol{x}_j \propto M f_j \propto \sum_{i=1}^{n} x_i \mathbb{1}_{f_i^\top f_j = 1},$$

where x_i is context as of time i. For a given j, there are $\#\{i : f_i^\top f_j = 1\}$ nonzero elements of $\boldsymbol{x}_j$, equal to the number of occurrences of primary features vector f_j in the sample. Because context lies on the unit circle, $\boldsymbol{x}_j$ is the unit vector with nonzero entries corresponding to locations in the list or time in the sample when elements of j are observed:

$$\boldsymbol{x}_j(i) = \frac{1}{\sqrt{\#\{i : f_i^\top f_j = 1\}}} \times \begin{cases} 1 & f_i^\top f_j = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Now consider the inner product $\boldsymbol{x}_j \cdot \boldsymbol{x}_k$: the sum of products of entries of $\boldsymbol{x}_j$ and $\boldsymbol{x}_k$. These are entries of ones and zeros scaled by the square root of the number of non-zero entries. Thus

$$\begin{aligned} \boldsymbol{x}_j \cdot \boldsymbol{x}_k &= \#\{i : (f_i^\top f_j \neq 0) \,\&\, (f_i^\top f_k \neq 0)\} \times \frac{1}{\sqrt{\#\{i : f_i^\top f_j = 1\}}} \frac{1}{\sqrt{\#\{i : f_i^\top f_k = 1\}}} \\ &= P(j,k)n \times \frac{1}{\sqrt{P(j)n}} \frac{1}{\sqrt{P(k)n}}. \end{aligned}$$

Dividing both the numerator and the denominator by n gives the relative frequency in the observed sample. Assuming the frequency in the observed sample equals the population density yields (21).

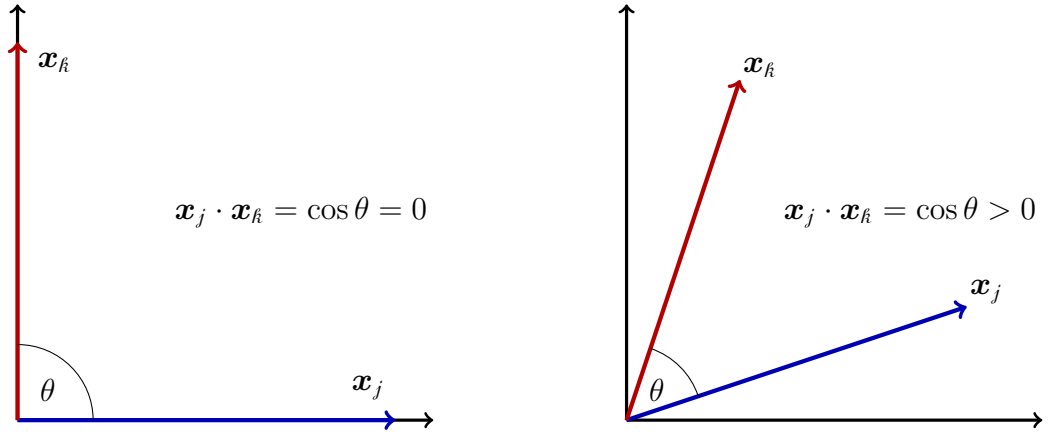
Finally, (22) follows from

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

and dropping the constant of proportionality $1/\sqrt{P(d)}$. Equation (24) follows from applying (21) to both sides of (23) and rearranging. □

Theorem 1 has a geometric interpretation. Because context vectors are unit 1, the inner product is the cosine of the angle between the vectors. If the vectors lie on top of each other, this corresponds to the case in which j and k are always seen together.

When they are never seen together, the vectors are orthogonal, and everything else lies somewhere in between. See, for example, Figure 2. In Panel (a), the context vectors are orthogonal and therefore have zero cosine similarity. In Panel (b), the similarity of the contexts is captured by the fact that the angle between them is smaller, implying a higher cosine similarity.



(a) Orthogonal contexts (b) Associated contexts

Figure 2: Geometric interpretation of cosine similarity between context vectors.

An immediate consequence of this theorem is how likelihood judgments depend on the probability distribution viewed by the agent. We first ask what happens when, conditional on d, h becomes more likely, holding fixed the distribution of d and the conditional probability of h given $-d$, where $-d$ should be understood as any element of $\mathscr{D}$ not equal to d.

The first statement below says that if h is judged to be more likely than $-h$, and more evidence accumulates in favor of h in the sense that $P(h|d)$ rises, then h will be judged as still more likely than $-h$ when cued by d. This is in line with Bayesian reasoning. The second statement says that if h is considered more likely than $-h$, and evidence accumulates that reduces the probability of h conditional on $-d$, then h will be considered as still more likely than $-h$ when cued by d. This is contrary to Bayesian reasoning.[6]

[6]The theorems below highlight the importance of $P(h, -d)$ in likelihood judgments regarding

Theorem 2. *Suppose that pairs of features are observed in proportion to probabilities $P(\cdot)$, resulting in $\boldsymbol{x}_h \cdot \boldsymbol{x}_d > \boldsymbol{x}_{-h} \cdot \boldsymbol{x}_d$ for some $h, -h \in \mathcal{H}$. Assume a new set of feature pairs are observed, updating probabilities to $P'(\cdot)$ such that*

1. *$P'(h|d) > P(h|d)$ but with $P(d)$ and $P(\cdot|-d)$ unchanged, and with $P(-h|d)$ either falling or remaining the same; or*
2. *$P'(h|-d) < P(h|-d)$ but with $P(d)$ and $P(\cdot|d)$ unchanged, and with $P(-h|-d)$ either increasing or remaining the same.*

Then, if $\boldsymbol{x}'$ represents the new retrieved context:

$$\boldsymbol{x}'_h \cdot \boldsymbol{x}'_d > \boldsymbol{x}_h \cdot \boldsymbol{x}_d > \boldsymbol{x}_{-h} \cdot \boldsymbol{x}_d \geq \boldsymbol{x}'_{-h} \cdot \boldsymbol{x}'_d \tag{26}$$

Proof. Because we are only concerned about the relative ranking of $\boldsymbol{x}_h \cdot \boldsymbol{x}_d$ and $\boldsymbol{x}_{-h} \cdot \boldsymbol{x}_d$, it suffices to consider (22), which we can rewrite as:

$$\begin{aligned} \boldsymbol{x}_h \cdot \boldsymbol{x}_d &\propto \frac{P(h|d)P(d)}{\sqrt{P(h|d)P(d) + P(h|-d)P(-d)}} \\ &\propto \sqrt{\frac{P(h|d)}{P(d) + (P(h|-d)/P(h|d))P(-d)}} \end{aligned} \tag{27}$$

which is increasing in $P(h|d)$ and decreasing in of $P(h|-d)$.

Suppose the first condition in the theorem holds. $P'(h|d) > P(h|d)$ implies

$$\begin{aligned} \sqrt{\frac{P'(h|d)}{P'(d) + (P'(h|-d)/P'(h|d))P'(-d)}} &= \sqrt{\frac{P'(h|d)}{P(d) + (P(h|-d)/P'(h|d))P(-d)}} \\ &> \sqrt{\frac{P(h|d)}{P(d) + (P(h|-d)/(P(h|d))P(-d)}}, \end{aligned}$$

where the equality follows from the assumption that $P(d)$ and $P(h|-d)$ remain unchanged, and the inequality follows from the fact that (27) is increasing in $P(h|d)$. This

$P(h, d)$. The importance of $P(h, -d)$ is also recognized by Bordalo et al. (2021) who explicitly incorporate it into a reduced-form model of probabilities. Appendix 3.5 compares our model with theirs.

shows $\boldsymbol{x}'_h \cdot \boldsymbol{x}'_d > \boldsymbol{x}_h \cdot \boldsymbol{x}_d$. Moreover, because we have assumed that $P'(-h|d) \leq P(-h|d)$ for hypothesis $-h \in \mathcal{H}$, similar reasoning implies

$$\begin{aligned}\sqrt{\frac{P'(-h|d)}{P'(d) + (P'(-h|-d)/P'(-h|d))P'(-d)}} &= \sqrt{\frac{P'(-h|d)}{P(d) + (P(-h|-d)/P'(-h|d))P(-d)}} \\ &\leq \sqrt{\frac{P(-h|d)}{P(d) + (P(-h|-d)/(P(-h|d))P(-d)}},\end{aligned}$$

so $\boldsymbol{x}_{-h} \cdot \boldsymbol{x}_d \geq \boldsymbol{x}'_{-h} \cdot \boldsymbol{x}'_d$. Putting these together, we have (26).

Now suppose instead that the second condition in the theorem holds. $P'(h|-d) < P(h|-d)$ implies

$$\begin{aligned}\sqrt{\frac{P'(h|d)}{P'(d) + (P'(h|-d)/P'(h|d))P'(-d)}} &= \sqrt{\frac{P(h|d)}{P(d) + (P'(h|-d)/P(h|d))P(-d)}} \\ &> \sqrt{\frac{P(h|d)}{P(d) + (P(h|-d)/(P(h|d))P(-d)}},\end{aligned}$$

where the first equality follows from the assumption that $P(d)$ and $P(h|d)$ remain unchanged, and the second follows from the fact that (27) is decreasing in $P(h|-d)$. This shows $\boldsymbol{x}'_h \cdot \boldsymbol{x}'_d > \boldsymbol{x}_h \cdot \boldsymbol{x}_d$. Moreover, because we have assumed that $P'(-h|-d) \geq P(-h|-d)$ for hypothesis $-h \in \mathcal{H}$, similar reasoning implies

$$\begin{aligned}\sqrt{\frac{P'(-h|d)}{P'(d) + (P'(-h|-d)/P'(-h|d))P'(-d)}} &= \sqrt{\frac{P(-h|d)}{P(d) + (P'(-h|-d)/P(-h|d))P(-d)}} \\ &\leq \sqrt{\frac{P(-h|d)}{P(d) + (P(-h|-d)/(P(-h|d))P(-d)}},\end{aligned}$$

so $\boldsymbol{x}_{-h} \cdot \boldsymbol{x}_d \geq \boldsymbol{x}'_{-h} \cdot \boldsymbol{x}'_d$. Putting these together, we have (26). □

We can also show that, given sufficient evidence of h conditional on d, the agent will rank the likelihood of h above any alternatives when cued with d. This is consistent with Bayesian reasoning.

Theorem 3. *Keeping $P(d)$ and $P(h|{-}d)$ fixed, for $P(h|d)$ sufficiently close to 1, $\boldsymbol{x}_h \cdot \boldsymbol{x}_d > \boldsymbol{x}_{-h} \cdot \boldsymbol{x}_d$.*

Proof. Suppose that the statement were false. Then there is a sequence of samples, indexed by n, such that as $P^{(n)}(h|d)$ approaches 1, the inequality is violated. Consider a subsequence such that $P^{(n)}(h|d)$ monotonically approaches 1, and simultaneously $P^{(n)}(-h|d)$ monotonically approaches zero. The latter is possible because $P(-h|d) \leq 1 - P(h|d)$, and is bounded below by zero. It follows from (21) that

$$\begin{aligned}
\lim_{n\to\infty} \frac{\boldsymbol{x}_{-h}^{(n)} \cdot \boldsymbol{x}_d^{(n)}}{\boldsymbol{x}_h^{(n)} \cdot \boldsymbol{x}_d^{(n)}} &= \lim_{n\to\infty} \frac{P^{(n)}(-h,d)}{\sqrt{P^{(n)}(-h)}} \frac{\sqrt{P^{(n)}(h)}}{P^{(n)}(h,d)} \\
&= \lim_{n\to\infty} \sqrt{\frac{P^{(n)}(h)}{P^{(n)}(-h)}} \frac{P^{(n)}(-h|d)}{P^{(n)}(h|d)} \\
&= \lim_{n\to\infty} \sqrt{\frac{P^{(n)}(h)}{P^{(n)}(-h)}} \lim_{n\to\infty} \frac{P^{(n)}(-h|d)}{P^{(n)}(h|d)} = 0
\end{aligned}$$

The last line follows because the limits of both terms are well-defined because $P(-h|-d)$ is unchanged. Moreover, $\frac{P^{(n)}(-h|d)}{P^{(n)}(h|d)}$ decreases monotonically to zero, whereas $\sqrt{\frac{P^{(n)}(h)}{P^{(n)}(-h)}}$ is bounded. Therefore, there exists an N such that $\frac{\boldsymbol{x}_{-h}^{(n)} \cdot \boldsymbol{x}_d^{(n)}}{\boldsymbol{x}_h^{(n)} \cdot \boldsymbol{x}_d^{(n)}} < 1$ for $n > N$, which is a contradiction. □

However, if h becomes sufficiently associated with $-d$, then the agent will rank h below any of the alternatives, inconsistent with Bayesian reasoning.

Theorem 4. *Keeping $P(\cdot|d)$ fixed (with $P(-h|d) > 0$), for $P(h)$ sufficiently close to 1, $\boldsymbol{x}_h \cdot \boldsymbol{x}_d < \boldsymbol{x}_{-h} \cdot \boldsymbol{x}_d$.*

Proof. Suppose that the statement were false. Then there is a sequence of samples, indexed by n, such that as $P^{(n)}(h)$ monotonically approaches 1 and $P^{(n)}(-h)$ mono-

tonically approaches zero, the inequality is violated. It follows from (21) that

$$\begin{aligned}
\lim_{n\to\infty} \frac{\boldsymbol{x}_h^{(n)} \cdot \boldsymbol{x}_d^{(n)}}{\boldsymbol{x}_{-h}^{(n)} \cdot \boldsymbol{x}_d^{(n)}} &= \lim_{n\to\infty} \frac{P^{(n)}(h,d)}{\sqrt{P^{(n)}(h)}} \frac{\sqrt{P^{(n)}(-h)}}{P^{(n)}(-h,d)} \\
&= \lim_{n\to\infty} \sqrt{\frac{P^{(n)}(-h)}{P^{(n)}(h)}} \frac{P^{(n)}(h|d)}{P^{(n)}(-h|d)} \\
&= \lim_{n\to\infty} \sqrt{\frac{P^{(n)}(-h)}{P^{(n)}(h)}} \lim_{n\to\infty} \frac{P^{(n)}(h|d)}{P^{(n)}(-h|d)} = 0
\end{aligned}$$

The last line follows because both terms are well-defined, and indeed $\frac{P^{(n)}(h|d)}{P^{(n)}(-h|d)}$ is constant. Moreover, $\sqrt{\frac{P^{(n)}(-h)}{P^{(n)}(h)}}$ monotonically decreases to zero. Therefore, there exists an N such that $\frac{\boldsymbol{x}_h^{(n)} \cdot \boldsymbol{x}_d^{(n)}}{\boldsymbol{x}_{-h}^{(n)} \cdot \boldsymbol{x}_d^{(n)}} < 1$ for $n > N$, which is a contradiction. □

For example, consider Example 2. $P(\textsc{Orange})n$ is the count of orange items in the list, and $P(\textsc{Orange}, \textsc{Number})n$ is the count of orange numbers. Feature Orange will retrieve a context vector with $P(\textsc{Orange})n$ nonzero elements. Because the context vector must have norm 1, each of these elements equals $1/\sqrt{P(\textsc{Orange})n}$. Analogous results hold true for feature Number. Now consider the inner product of these two context vectors. Because the inner product sums the nonzero elements, of which there are at most $\min(P(\textsc{Number}), P(\textsc{Orange}))n$. In this example, there are exactly $P(\textsc{Orange})n$ nonzero elements, 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 $P(\textsc{Orange})n$ nonzero elements. Regardless, 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 $P(\textsc{Orange}, \textsc{Number})/\sqrt{P(\textsc{Number})P(\textsc{Orange})}$, where $P(\textsc{Orange}, \textsc{Number})$ is the joint likelihood of orange and number.

The discussion in this section highlights the importance of using L^2 scaling, namely, the standard Euclidean norm. If we were to use L^1 scaling, the inner product would actually fall with the sample size, a counterintuitive result. Each of the elements in the

context vector associated with feature ORANGE would equal $1/(P(\text{ORANGE})n)$, whereas each item in the context vector associated with NUMBER equals $1/(P(\text{NUMBER})n)$. The inner product would equal $P(\text{NUMBER}, \text{ORANGE})/(n\sqrt{P(\text{NUMBER})P(\text{ORANGE})})$, 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.

3.3 Qualitative results

We first show that this model fits the results of Bordalo et al. (2021), Study 1 of which forms the basis of Example 2 above. In this experiment, associations are formed only during the presentation of the items, ruling out the possibility that the subject learned biased conditional probabilities outside of the experiment. This controlled setting also allows the experimenter to vary the number of targets and decoys and assess how this affects likelihood judgments. For these reasons, we emphasize the results in this study when we fit the model to the data. We first examine the qualitative predictions of the model.

To control for possible differences in the mental database for orange and blue, 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 (*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 "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?" They find the following two results:

R1 Participants are significantly more likely to say that orange is the more likely color in the blue treatment versus the gray treatment.

R2 The percentage of participants who say that orange is the most likely color declines with k, as k blue words are replaced with k orange words.

Bordalo et al. (2021) also ask questions regarding the probability and number of orange words. Appendix A shows how the model can be extended to handle probabilistic questions, and then shows that our model matches results analogous to Results 1 and 2 for probability judgments.

Table 1 illustrates the mapping between this setting and our model. The cue $d = \textsc{Number}$ and the decoy $-d$ is shape in the gray treatment and word in the blue treatment. The targets are the colors $h = \textsc{Orange}$ and $-h = \textsc{Blue}$.[7]

At the beginning of this section, we showed how the model can produce differential results for the blue versus gray treatment in a toy example. We now use the numbers from the Bordalo et al. (2021) experiment. We find, for the gray treatment,

$$\begin{aligned}
\boldsymbol{x}_{\textsc{Blue}} \cdot \boldsymbol{x}_{\textsc{Number}} &= \frac{P_g(\textsc{Blue}, \textsc{Number})}{\sqrt{P_g(\textsc{Blue})P_g(\textsc{Number})}} = \frac{15}{\sqrt{15}\sqrt{15+10}} = 0.77 \\
\boldsymbol{x}_{\textsc{Orange}} \cdot \boldsymbol{x}_{\textsc{Number}} &= \frac{P_g(\textsc{Orange}, \textsc{Number})}{\sqrt{P_g(\textsc{Orange})P_g(\textsc{Number})}} = \frac{10}{\sqrt{10}\sqrt{15+10}} = 0.63
\end{aligned}$$

Under the gray treatment, the agent correctly recalls a greater likelihood that the number is blue.

In contrast, in the blue treatment,

$$\begin{aligned}
\boldsymbol{x}_{\textsc{Blue}} \cdot \boldsymbol{x}_{\textsc{Number}} &= \frac{P_b(\textsc{Blue}, \textsc{Number})}{\sqrt{P_b(\textsc{Blue})P_b(\textsc{Number})}} = \frac{15}{\sqrt{15+25}\sqrt{15+10}} = 0.47 \\
\boldsymbol{x}_{\textsc{Orange}} \cdot \boldsymbol{x}_{\textsc{Number}} &= \frac{P_b(\textsc{Orange}, \textsc{Number})}{\sqrt{P_b(\textsc{Orange})P_b(\textsc{Number})}} = \frac{10}{\sqrt{10}\sqrt{15+10}} = 0.63
\end{aligned}$$

The introduction of blue words weakens the association between the color blue and numbers, so that, under the blue treatment, the agent incorrectly recalls a greater

[7] In the gray treatment, gray could also be considered as a third target. However, because no gray numbers were ever observed, its similarity to the cue is trivially zero, so we ignore it.

likelihood of the number being orange:

$$P_b(\textsc{Orange}|\textsc{Number}) > P_b(\textsc{Blue}|\textsc{Number})\sqrt{\frac{P_b(\textsc{Orange})}{P_b(\textsc{Blue})}}.$$

However,

$$P_g(\textsc{Orange}|\textsc{Number}) < P_g(\textsc{Blue}|\textsc{Number})\sqrt{\frac{P_g(\textsc{Orange})}{P_g(\textsc{Blue})}}.$$

To build intuition regarding Result 2, we add an orange word to our toy model from Section 3.1:

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

The list becomes:

$$\underbrace{\text{Blue } '5'}_{\begin{bmatrix}0\\0\\1\\0\end{bmatrix},\begin{bmatrix}1\\0\\0\\0\end{bmatrix}} \;//\; \underbrace{\text{Orange } '3'}_{\begin{bmatrix}0\\0\\0\\1\end{bmatrix},\begin{bmatrix}1\\0\\0\\0\end{bmatrix}} \;//\; \underbrace{\text{Blue } '\text{CAT}'}_{\begin{bmatrix}0\\0\\1\\0\end{bmatrix},\begin{bmatrix}0\\1\\0\\0\end{bmatrix}} \;//\; \underbrace{\text{Orange } '\text{DOG}'}_{\begin{bmatrix}0\\0\\0\\1\end{bmatrix},\begin{bmatrix}0\\1\\0\\0\end{bmatrix}} \tag{28}$$

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

Following the reasoning of (14), we have:

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

Note that the first three rows of M are the same as what we had before adding an orange word.

$$M = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}. \tag{29}$$

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

$$\boldsymbol{x}_{\text{NUMBER}} \propto M \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \propto \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \\ 0 \end{bmatrix}. \tag{30}$$

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

$$\boldsymbol{x}_{\text{BLUE}} = \begin{bmatrix} 1/\sqrt{2} \\ 0 \\ 1/\sqrt{2} \\ 0 \end{bmatrix}, \quad \boldsymbol{x}_{\text{ORANGE}} = \begin{bmatrix} 0 \\ 1/\sqrt{2} \\ 0 \\ 1/\sqrt{2} \end{bmatrix}.$$

Now, orange and blue are equally good matches for the cue 'Number':

$$\boldsymbol{x}_{\text{NUMBER}} \cdot \boldsymbol{x}_{\text{BLUE}} \quad = \quad \boldsymbol{x}_{\text{NUMBER}} \cdot \boldsymbol{x}_{\text{ORANGE}}.$$

In general, suppose we replace k blue words in the blue treatment with k orange words. Using numbers from the experiment:

$$\boldsymbol{x}_{\text{Blue}} \cdot \boldsymbol{x}_{\text{Number}} = \frac{P_b(\text{Blue}, \text{Number})}{\sqrt{P_b(\text{Blue})P_b(\text{Number})}} = \frac{15}{\sqrt{15+25-k}\sqrt{15+10}}$$

and

$$\boldsymbol{x}_{\text{Orange}} \cdot \boldsymbol{x}_{\text{Number}} = \frac{P_b(\text{Orange}, \text{Number})}{\sqrt{P_b(\text{Orange})P_b(\text{Number})}} = \frac{10}{\sqrt{10+k}\sqrt{15+10}}$$

At around $k = 5.4$, the model no longer predicts that participants will recall that orange is more likely.

We next show how the model addresses results in Kahneman and Tversky (1973), and in particular, Example 1. In this case, we do not know subjects' database (other than the base rates), which was an advantage we had in modeling the Bordalo et al. (2021) experimental framework.

Recall that the first group of subjects is not given the description of Tom W., and is asked to report the base rates of the 9 fields of study (Kahneman and Tversky, 1973, Table 2). This table gives a base rate for the humanities, for example, as 20%, and for computer science as 7%. The second and third groups are given the description and asked to rank the likelihood that Tom W. majors in a subject and the "similarity" (the words of Kahneman and Tversky) between Tom W. and the subject, respectively. For the moment, we ignore the similarity rank and focus on what Kahneman and Tversky term the likelihood rank of a major, which corresponds to the Bayesian posterior. The Bayesian calculation implies

$$P(\text{Field}|\text{Tom}) \propto P(\text{Tom}|\text{Field}) \underbrace{P(\text{Field})}_{\text{base rate}}, \quad \text{Field} \in \{\text{Computers}, \text{Humanities}\}. \tag{31}$$

If there were zero humanities majors that resembled Tom W. in subjects' database then ranking the probability that Tom W. is a computer scientist higher than a humanities major is consistent with Bayesian reasoning. This is clearly not what Kahneman and

Table 1: Mapping between cognitive model and experimental settings

Role in cognitive model	Theorem 1 notation	Kahneman and Tversky experiment	Bordalo et al. experiment
Cue	d	Tom W.	Number
Decoy	$-d$	Other people	Word
Target 1	h	Computer science	Orange
Target 2	$-h$	Humanities	Blue

Tversky have in mind, and they argue against this possibility. Nevertheless, without knowledge of the subjects' database, it is not possible to rule this scenario out. In this case, both the Bayesian model and our model would produce the observed result.

However, there is a wide range of probabilities under which our model would produce a different result than the Bayesian model. This range consists of cases that broadly resemble Study 1 of Bordalo et al. (2021), namely instances of Tom W. appear more frequently among humanities majors than computer science majors, but that they are difficult to recall because there are many other types of people majoring in the humanities. Table 1 maps the Tom W. experiment into the words/numbers experiment.

Specifically, consider Tom W. to be the cue and the targets to be the various fields of study (majors). As an illustration of where our model would make a different prediction than Bayesian inference, consider the case in which there are 7 computer science majors for every 20 humanities majors, therefore matching the proportions in base rates. Assume all computer science majors resemble Tom W., but half of the humanities majors do so. In this case, memory databases will be such that

$$\begin{aligned} \boldsymbol{x}_{\textsc{Computers}} \cdot \boldsymbol{x}_{\textsc{Tom}} &\propto \frac{P(\textsc{Computers}, \textsc{Tom})}{\sqrt{P(\textsc{Computers})}} = \frac{7}{\sqrt{7}} \\ \boldsymbol{x}_{\textsc{Humanities}} \cdot \boldsymbol{x}_{\textsc{Tom}} &\propto \frac{P(\textsc{Humanities}, \textsc{Tom})}{\sqrt{P(\textsc{Humanities})}} = \frac{10}{\sqrt{20}} \end{aligned}$$

and thus, $\boldsymbol{x}_{\textsc{Computers}} \cdot \boldsymbol{x}_{\textsc{Tom}} > \boldsymbol{x}_{\textsc{Humanities}} \cdot \boldsymbol{x}_{\textsc{Tom}}$. Assuming (as Kahneman and Tversky (1973) implicitly do) that the memory databases on average are the same among the

third group that assess likelihoods as the first group that assesses base rates, this would explain why computer science is judged as a more likely major for Tom W. than is the humanities. Note that Bayesian inference leads to the opposite conclusion in this case:

$$\begin{aligned} P(\textsc{Computers} \,|\, \textsc{Tom}) &\propto P(\textsc{Tom} \,|\, \textsc{Computers}) P(\textsc{Computers}) = P(\textsc{Tom}, \textsc{Computers}) = 7/27 \\ P(\textsc{Humanities} \,|\, \textsc{Tom}) &\propto P(\textsc{Tom} \,|\, \textsc{Humanities}) P(\textsc{Humanities}) = P(\textsc{Tom}, \textsc{Humanities}) = 10/27 \end{aligned}$$

Thus the model accounts for the Tom W. experiment in Kahneman and Tversky (1973).

Finally, recall that there is a second group that assesses similarities. While the analysis above does not require that we take a stance on what Kahneman and Tversky (1973) mean by "similarity," their use of the word is consistent with the general idea that it is contextual similarity that drives probabilistic judgement.

3.4 Quantitative results

The model above is stylized. It abstracts, for example, from initial associations M_0. It also abstracts from contextual dynamics, employing a minimalist version of autoregressive context. Finally, the temporal context model does not itself explain all aspects of memory. All of these are potential sources of errors in the model. We encapsulate these sources of error in a simple way by assuming that inner products are perceived with noise by the agent. Assuming perception occurs with noise is standard (Stanislaw and Todorov, 1999), see also (Kahana, 2012, Chapter 2). The addition of a single parameter allows the model to quantitatively fit the data of Bordalo et al. (2021).

In this extension, the criterion for stating orange is more likely is

$$(\boldsymbol{x}_{\textsc{Orange}} \cdot \boldsymbol{x}_{\textsc{Number}}) e^{\tilde{\epsilon}_{\textsc{Orange}}} > (\boldsymbol{x}_{\textsc{Blue}} \cdot \boldsymbol{x}_{\textsc{Number}})\, e^{\tilde{\epsilon}_{\textsc{Blue}}},$$

where $\tilde{\epsilon}$ are agent-specific random variables.[8] Substituting into the solution and ex-

[8]For gray, the inner product will be zero, so no subject would ever say gray is most likely.

pressing in terms of the error terms, this threshold is equivalently

$$\frac{P_b(\textsc{Orange}, \textsc{Number})}{P_b(\textsc{Blue}, \textsc{Number})}\sqrt{\frac{P_b(\textsc{Blue})}{P_b(\textsc{Orange})}} > e^{\tilde{\epsilon}_{\textsc{Blue}}-\tilde{\epsilon}_{\textsc{Orange}}}.$$

Taking the log and substituting in the values from Bordalo et al. (2021), we find that a participant in the gray treatment states that

$$\log\left(\frac{10}{15}\right) + \frac{1}{2}\log\left(\frac{15}{10}\right) \approx -0.20 > \tilde{\epsilon}_{\textsc{Blue}} - \tilde{\epsilon}_{\textsc{Orange}},$$

and a participant in the blue treatment will say that orange is more likely if

$$\log\left(\frac{10}{15}\right) + \frac{1}{2}\log\left(\frac{40}{10}\right) \approx 0.29 > \tilde{\epsilon}_{\textsc{Blue}} - \tilde{\epsilon}_{\textsc{Orange}}.$$

The latter criterion is more likely to be met, so those in the blue condition will tend to respond orange at a higher rate.

Suppose now that $\tilde{\epsilon}_{\textsc{Blue}}$ and $\tilde{\epsilon}_{\textsc{Orange}}$ are iid normally distributed. In this case, their difference

$$\tilde{\epsilon} \equiv \tilde{\epsilon}_{\textsc{Blue}} - \tilde{\epsilon}_{\textsc{Orange}} \sim N(0, \sigma^2)$$

for some variance σ^2. Under this assumption, the proportion of participants who judge orange numbers to be more likely than blue numbers will simply be given by the normal CDF. Let $\Phi(z)$ denote the standard normal CDF. The proportion in the gray treatment equals

$$\Phi\left(\frac{1}{\sigma}\left(\log\left(\frac{10}{15}\right) + \frac{1}{2}\log\left(\frac{15}{10}\right)\right)\right),$$

and the proportion in the blue treatment will equal

$$\Phi\left(\frac{1}{\sigma}\left(\log\left(\frac{10}{15}\right) + \frac{1}{2}\log\left(\frac{40}{10}\right)\right)\right).$$

The latter is larger than the former, as found in the experiment.

For Study 2, we want to know how the answer changes if we replace the k blue

words in the blue treatment with the k orange words. In this case, the participant will say that orange is more likely if

$$\log\left(\frac{10}{15}\right) + \frac{1}{2}\log\left(\frac{40-k}{10+k}\right) > \tilde{\epsilon},$$

which is less likely to be true for high values of k. The proportion of participants is then

$$\Phi\left(\frac{1}{\sigma}\left(\log\left(\frac{10}{15}\right) + \frac{1}{2}\log\left(\frac{40-k}{10+k}\right)\right)\right),$$

which falls in k, matching the experimental results.

3.4.1 Model fit

To fit the model to the data of Bordalo et al. (2021), we estimate σ by minimizing the distance between all moments in the data and model, where the data come from Figures 1 and 2 of Bordalo et al. (2021).[9] Specifically, we estimate σ to be

$$\hat{\sigma} = \underset{\sigma}{\operatorname{argmin}}\left\{\text{error}(\sigma)^\top \text{error}(\sigma)\right\},$$

where $\text{error}(\sigma)$ is a vector of (percent) errors with jth element

$$\text{error}_j(\sigma) = \frac{\text{data moment}_j - \text{model moment}_j(\sigma)}{\text{data moment}_j}.$$

This procedure gives an estimate of

$$\hat{\sigma} \approx 1.3737.$$

Figures 3 and 4 plot the resulting moments, and show that the model provides an excellent fit across the experimental conditions.

[9] In Appendix C, we also consider a simpler calibration that exactly identifies σ from one moment, the gray treatment.

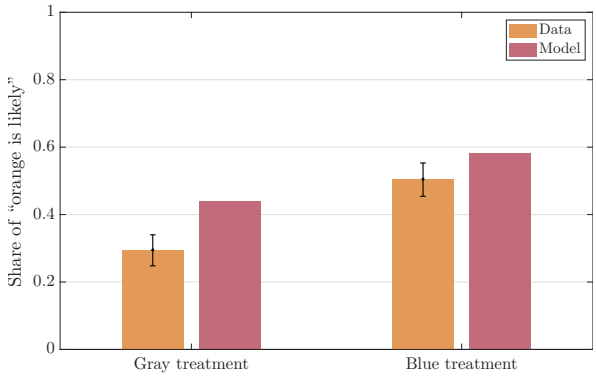
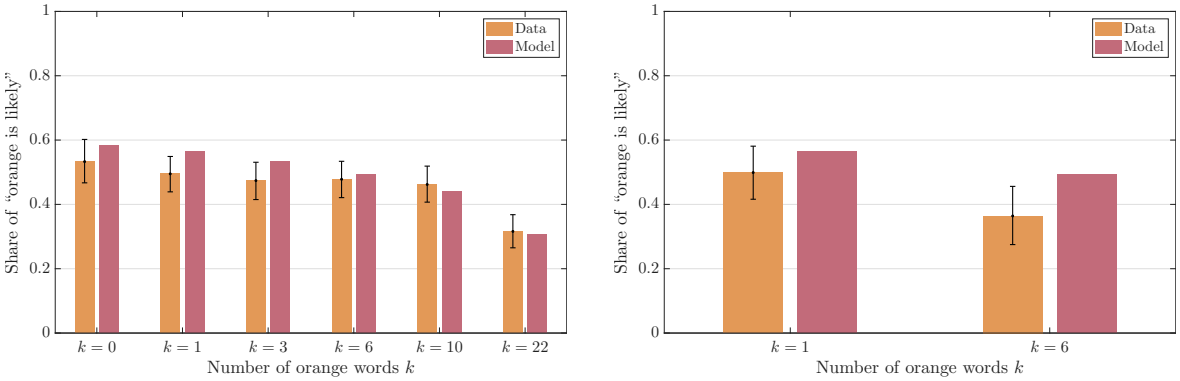


Figure 3: Study 1, Figure 1 of Bordalo et al. (2021). The noise parameter σ is estimated to match all moments in Figures 3 and 4.



(a) Bordalo et al. Figure 2a

(b) Bordalo et al. Figure 2b

Figure 4: Study 2, Figure 2 of Bordalo et al. (2021). The noise parameter σ is estimated to match all moments in Figures 3 and 4.

3.5 Comparison with a probabilistic model

The previous sections, together with Appendices A and B, show that the model can account for the evidence in Bordalo et al. (2021). Here, we compare our model to theirs.

Building on Gennaioli and Shleifer (2010) and Bordalo et al. (2016), Bordalo et al. (2021) specify a reduced-form model with the ability to capture deviations from Bayesian updating. They define a similarity function $S(d,h)$, to which they apply a Luce choice rule (see Appendix A) to determine the subjective probability:

$$\tilde{P}(h|d) = \frac{e^{S(d,h)}}{e^{S(d,h)} + e^{S(d,\neg h)}}, \tag{32}$$

where

$$S(d,h) = \alpha f(P(d,h)) - \gamma f(P(-d,h)), \tag{33}$$

and where f is an increasing function, parameterized by $f(x) = \log(c+x)$ for a constant c. Under this parameterization, the model captures Bayesian inference when $\alpha = 1$ and $\gamma = c = 0$. Note that in this model, by construction, the likelihood assessments increase in $P(h|d)$ (which is necessary for basic plausibility) and decrease in $P(h|{-d})$ (without which the model would ultimately resemble a Bayesian one). As described above, these are also key implications of our cognitive theory.

An important difference between the models is that, in the Bordalo et al. (2021) model, γ is a flexible parameter, introduced to fit the representativeness heuristic. In contrast, our model provides a cognitive foundation for γ, and in so doing eliminates γ entirely as a free parameter.[10] A second key difference is that the Bordalo et al. model is fundamentally based on probability distortion, whereas our model is based on recognition memory. To illustrate the benefits of a cognitive framework, we turn to the conjunction fallacy.

[10]This statement is an approximate one because the functional form (32–33) does not nest our results. Rather, our model offers a cognitive foundation for why $P(h,-d)$ would enter the probability assessment of $P(h|d)$.

Table 2: **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 Further applications

4.1 Tversky and Kahneman's conjunction fallacy

In the previous section, the possible targets were mutually exclusive: for example, a number could be either blue or orange, but not both. We now relax that assumption, and in the process address a second puzzle, the conjunction fallacy. Tversky and Kahneman (1983) present a wide range of experimental findings in which subjects judge a conjunction of two statements to be significantly more likely than one of the constituent statements, a violation of the laws of probability. We describe one of their experiments in detail.

Example 3 (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." Subjects were then asked to rank, in order of likelihood, the possibilities given in Table 2. Over 80% of participants rank, from more to less likely, F>T&F>T for Linda and A>A&J>J for Bill. Regardless 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 *conjunction fallacy*.

Here, we show how the model explains this fallacy. We first give an example and then a formal proof. Consider subjects asked to rank the probability that Bill is an accountant (A), a jazz player (J), or both (A ∩ J) (see Table 2). While Bill as presented in the Tversky and Kahneman (1983) experiment may be a composite representation, we simplify notation by representing Bill as basis vector f_{Bill}. The features space in this simplified example is:

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

Assume Bill is seen in a context that makes it clear that he is an accountant. Subjects

see jazz players, and they also see accountants who are not Bill. Thus:

$$\begin{array}{ccccc} \text{Bill as Accountant} & // & \text{Accountant} & // & \text{Jazz Player} \\ \underbrace{\begin{bmatrix}1\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\end{bmatrix}} & & \begin{bmatrix}0\\1\\0\end{bmatrix} & & \begin{bmatrix}0\\0\\1\end{bmatrix} \\ \boldsymbol{x}_1 & & \boldsymbol{x}_2 & & \boldsymbol{x}_3 \\ \begin{bmatrix}1\\0\\0\end{bmatrix} & & \begin{bmatrix}0\\1\\0\end{bmatrix} & & \begin{bmatrix}0\\0\\1\end{bmatrix} \end{array} \tag{34}$$

The mental representation (34) implies that the memory matrix is equal to:

$$M = M_0 + \begin{bmatrix}1\\0\\0\end{bmatrix}\begin{bmatrix}1 & 1 & 0\end{bmatrix} + \begin{bmatrix}0\\1\\0\end{bmatrix}\begin{bmatrix}0 & 1 & 0\end{bmatrix} + \begin{bmatrix}0\\0\\1\end{bmatrix}\begin{bmatrix}0 & 0 & 1\end{bmatrix},$$

With M_0 equal to the zero matrix,

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

In the experiment, the cue is the feature 'Bill':

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

which picks up the single context in which Bill is seen.

The targets are Accountant together with Jazz, Accountant, and Jazz:

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

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

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

From which we see that

$$\boldsymbol{x}_{\text{Bill}} \cdot \boldsymbol{x}_{\text{A}} > \boldsymbol{x}_{\text{Bill}} \cdot \boldsymbol{x}_{\text{A}\cap\text{J}} > \boldsymbol{x}_{\text{Bill}} \cdot \boldsymbol{x}_{\text{J}}.$$

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," this composite feature 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.

To generalize from this specific example, we translate the problem into the formal language of Section 3. Recall that $\Omega = \mathcal{H} \times \mathcal{D}$. Implicit in this definition is that members of the population are identified by a single hypothesis and a single cue. We now allow members to be potentially identified by two hypotheses, one in the set $\mathcal{H}_1$, and one in the set $\mathcal{H}_2$, so that $\Omega = \mathcal{H}_1 \times \mathcal{H}_2 \times \mathcal{D}$. Note that any probability measure P over Ω is also a probability measure over $\mathcal{H}_1$ and $\mathcal{H}_2$. Thus, $P(h_{1,j}) =$

$\sum_{h_{2,k}\in\mathcal{H}_2} P(h_{1,j}, h_{2,k})$, for $h_{1,j} \in \mathcal{H}_1$. Clearly $P(h_{1,j}, h_{2,k}) \leq P(h_{1,j})$, with a strict inequality provided that P allows some instances of $h_{1,j}$ to occur with an element of $\mathcal{H}_2$ other than $h_{2,k}$.

In this example, it is useful to allow agents to view elements that are in $\mathcal{H}_1 \times \mathcal{H}_2 \times \mathcal{D}$, in $\mathcal{H}_1 \times \mathcal{D}$, in $\mathcal{H}_2 \times \mathcal{D}$, or only in $\mathcal{H}_1$ and $\mathcal{H}_2$. Because all elements of Ω have attributes in each of these sets, we need to extend our definitions to allow the agent to view certain subsets Ω. For example, if an agent observes someone to be an accountant but does not know if the person also plays jazz, then the agent is observing a subset of Ω rather than a specific element. If it becomes known whether the individual is a jazz player or not, the subset resolves into a single element. We let $\wp(\Omega)$ denote this broader set of admissible subsets.

We maintain the assumption of Section 3 that $f_{h_{1,j}}$ and $f_{h_{2,j}}$, etc., are basis vectors. The following naturally extends to the model of Section 3 to account for the conjunction fallacy.

Model Summary (Conjunctions)**.** *Suppose that an agent views the n elements of $\wp(\Omega)$ as feature vectors f_i, $i = 1, \ldots, n$. If $f_i = f_{h_1}, f_{h_2}, f_d$, for $h_j \in \mathcal{H}_j$, $f_d \in \mathcal{D}$, then f_i is a basis vector; otherwise f_i is a composite representation. Feature vectors f_i combine with orthogonal contexts $\boldsymbol{x}_i$ to form memory M as in (7), with M_0 the zero matrix. Let f_j, $j \in \mathcal{H}_1 \cup \mathcal{H}_2 \cup (\mathcal{H}_1 \times \mathcal{H}_2) \cup \mathcal{D}$ represent a cue ($d \in \mathcal{D}$), a target ($h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2$), or a conjunction target ($h_1 \cap h_2 \in \mathcal{H}_1 \times \mathcal{H}_2$). In the case of the conjunction target, the agent forms the composite representation. When cued with d, the agent ranks the conjunction target $h_1 \cap h_2$ as more likely than h_1 (given cue d) if*

$$\boldsymbol{x}_d \cdot \boldsymbol{x}_{h_1 \cap h_2} > \boldsymbol{x}_d \cdot \boldsymbol{x}_{h_1}, \tag{38}$$

where $\boldsymbol{x}_j$ is a norm-1 retrieved context vector such that $\boldsymbol{x}_j \propto M f_j$.[11]

In this setting, the conjunction fallacy arises under the following conditions.

[11] We use the notation h_1 to denote a generic element of $\mathcal{H}_1$, and similarly h_2 to denote a generic element of $\mathcal{H}_2$ when the second subscript is not necessary for clarity.

Theorem 5. *Suppose that the agent's sample matches the population in that items appear in their correct proportions. Then*

$$\boldsymbol{x}_d \cdot \boldsymbol{x}_{h_1 \cap h_2} = \frac{P(h_1, d) + P(h_2, d)}{\sqrt{(P(h_1) + P(h_2) + 2P(h_1 \cap h_2))P(d)}} \tag{39}$$

$$\boldsymbol{x}_d \cdot \boldsymbol{x}_{h_1} = \frac{P(h_1, d)}{\sqrt{P(h_1)P(d)}}. \tag{40}$$

Moreover, (38) is equivalent to:

$$1 + \frac{P(h_2|d)}{P(h_1|d)} > \sqrt{1 + \frac{P(h_2)}{P(h_1)} + 2P(h_1|h_2)}. \tag{41}$$

Proof. Equation 38 follows from the Model Summary, with the composite representation $f_{h_1} + f_{h_2}$ forming the cue. Equation 40 follows from Theorem 1. Given (39), (41) follows from substituting (39) and (40) into (38), dividing both sides by $\sqrt{P(d)}$ to create conditional probabilities and then rearranging, noting that, by the laws of probabilities, $P(h_1 \cap h_2)/P(h_2) = P(h_1 \,|\, h_2)$.

Therefore, it suffices to prove (39). Consider that

$$\boldsymbol{x}_{h_1 \cap h_2} \propto M(f_{h_1} + f_{h_2})$$

The unscaled context vector $M(f_{h_1} + f_{h_2})$ will have entries equal to 1 corresponding to times when either h_1 or h_2 are observed (but not both) and a 2 corresponding to times when both h_1 and h_2 are observed. Therefore, $(M(f_{h_1} + f_{h_2}))^\top M(f_{h_1} + f_{h_2}) = \#\{h_1 \cap (-h_2)\} + \#\{h_2 \cap (-h_1)\} + 4\#\{h_1 \cap h_2\} = \#\{h_1\} + \#\{h_2\} + 2\#\{h_1 \cap h_2\}$. When divided by the total number of elements in the population, this becomes $P(h_1) + P(h_2) + 2P(h_1 \cap h_2)$. In the numerator, we see cases where d and h_1 coincide, and d and h_2 coincide. There are $\#\{h_1 \cap d\} + \#\{h_2 \cap d\}$ cases. Dividing by the number of elements in the population yields (39). □

Suppose, for instance, that there are no observations of Bill as a jazz player. The

assessed probability that Bill could be both a Jazz player and an accountant is still greater than zero, though less than an accountant alone.

This application highlights the role of the cognitive model in that one would not obtain these results by applying the probability rule (21) to the conjunction. To consider a simple example, suppose that someone like Bill was never observed to be a jazz player. According to (21), the subject would rank the probability as zero. In contrast, the model above allows the cue jazz and accountant to work together, so that relevant instances are still brought to mind.

4.2 The cross-section of stock returns

One of the most striking empirical facts in finance is the outperformance of value stocks, namely stocks with book-to-market ratios that are relatively high, relative to growth stocks, namely stocks with low book-to-market ratios (Fama and French, 1992; Barberis et al., 1998). Standard finance theory would attribute this difference to a difference in risk. Conventional risk measures, however, do not appear to differ along the book-to-market dimension (Lettau and Wachter, 2007). The field continues to seek risk-based explanations, but so far an explanation that (1) accounts for different expected returns while not (2) implying greater risk along conventional measures has proven illusive.

An alternative explanation posits that growth stocks exhibit low returns because they are overpriced. This raises the question of what kind of cognitive bias would survive the financial incentives needed to overcome this well-publicized anomaly. The representativeness heuristic, which is "almost right," is a candidate for an explanation. In fact, in an early economic application of the representativeness heuristic, La Porta (1996) sorts companies based on analysts' expectations, and specifically on the "long-term growth" (LTG) measure of the I/B/E/S (Institutional Brokers' Estimate System) dataset.[12] He finds that stocks with high long-term growth forecasts

[12]For the LTG measure, analysts are instructed to forecast the company's annual increase in operating earnings "over its next full business cycle, typically spanning three to five years."

exhibit underperformance that is unexplained by standard risk metrics, a result that Bordalo et al. (2019a) replicate in more recent data. To explicitly connect this bias to the value premium, Guo and Wachter (2025) sort stocks according to book-to-market ratios, the traditional metric to determine value and growth. They show that growth companies tend to have higher future growth forecasts, not just as measured by LTG but also as measured over the next several years. They also show that while all projections are upward-biased, this is especially true of growth companies. Here, we show how our framework can account for these findings.

Let Ω denote the population of firms. Each firm $i \in \Omega$ has an expected annual growth rate of earnings, which we denote λ_i. We assume that there are two possibilities: $\lambda_i \in \{\lambda_\ell, \lambda_h\}$, where $\lambda_\ell < \lambda_h$. Let $g_{i,t+1}$ denote the realized growth rate for firm i between time t and $t+1$. Then

$$\mathbb{E}[g_{i,t+1}] = \lambda_i,$$

where $\mathbb{E}$ denotes the physical expectation.

The analyst forms beliefs regarding λ_i based on the data but also subject to the representativeness heuristic. Specifically, let G_{iT} denote average growth for firm i between time 0 and T: $\frac{1}{T}\sum_{t=0}^{T-1} g_{i,t+1} \equiv G_{iT}$. For convenience, we assume the analyst categorizes average growth as either high ($G_{iT} \in \mathcal{G}_h \equiv d$) or low ($G_{iT} \in \mathcal{G}_\ell \equiv -d$).[13] The analyst also sees data on subsequent growth $\frac{1}{\tau}\sum_{t=T}^{T-1+\tau} g_{i,t+1}$ for some large τ. Because τ is large, $\frac{1}{\tau}\sum_{t=T}^{T-1+\tau} g_{i,t+1} \approx \lambda_i$, and thus observing the average is effectively the same as having observed λ_i for each firm. Analysts observe each firm and encode the firm in its own context: 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.[14]

[13] Specifically, the analyst chooses a threshold $\bar{G} \in (\lambda_\ell, \lambda_h)$ such that $\mathcal{G}_\ell = \{G \in \mathbb{R} : G < \bar{G}\}$ and $\mathcal{G}_h = \mathbb{R} \setminus \mathcal{G}_\ell$. Note that the analyst could partition this set into arbitrarily many bins, which our theory accommodates, so this is without loss of generality.

[14] This memory database could arise from the analyst organizing and looking at historical data in a spreadsheet at the time of analysis. It could also arise through the specific memory mechanism of

Table 3: Mapping between cognitive model and economic forecasts

Role in cognitive model	Theorem 1 notation	Bordalo et al. experiment	Earnings expectation	Recession Forecasting
Cue	d	Number	High past growth $\mathcal{G}_{\hbar}$	Recession indicators $\mathcal{G}_{\imath}$
Decoy	$-d$	Word	Low past growth $\mathcal{G}_{\ell}$	Normal-times indicators $\mathcal{G}_{n}$
Target 1	h	Orange	High future growth $\lambda_{\hbar}$	Recession $z_{t+1} = 1$
Target 2	$-h$	Blue	Low future growth λ_{ℓ}	Normal times $z_{t+1} = 0$

We are now in the setting of Theorem 1. The matrix M contains $n = \|\Omega\|$ rows, each corresponding to the context in which a firm is viewed. Each row has a combination of ones and zeros, corresponding to whether $\mathcal{G}_{\hbar}$ or $\mathcal{G}_{\ell}$ is observed, and whether $\lambda_{\hbar}$ or λ_{ℓ} is observed. The cue is the past data of a given firm, and in particular $\mathcal{G}_{\hbar}$ or $\mathcal{G}_{\ell}$.[15] The cued context is determined using the basis vector corresponding to $G_{iT} \in \mathcal{G}_{\hbar}$, and the target contexts ($\boldsymbol{x}_{\lambda_{\hbar}}$ and $\boldsymbol{x}_{\lambda_{\ell}}$) are determined using the last two basis vectors. We can immediately apply Theorem 1 to see that the agent infers that the firm is more likely to be high-growth than low-growth if

$$\boldsymbol{x}_{\lambda_{\hbar}} \cdot \boldsymbol{x}_{\mathcal{G}_{\hbar}} = \frac{P(\lambda_{\hbar}, G \in \mathcal{G}_{\hbar})}{\sqrt{P(\lambda_{\hbar})}} > \frac{P(\lambda_{\ell}, G \in \mathcal{G}_{\hbar})}{\sqrt{P(\lambda_{\ell})}} = \boldsymbol{x}_{\lambda_{\ell}} \cdot \boldsymbol{x}_{\mathcal{G}_{\hbar}},$$

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

Table 3 maps this problem into the settings of Theorem 1 and Section 3.3. As in these previous results, the difficulty for the decision maker (in this case the analyst or investor) lies with $P(-h, d)$, or rather with the low value of $\boldsymbol{x}_{\lambda_{\ell}} \cdot \boldsymbol{x}_{\mathcal{G}_{\hbar}}$. The sample is not biased: There are $\mathcal{G}_{\hbar}$ firms with λ_{ℓ}. The problem is that there are also so many $\mathcal{G}_{\ell}$ firms with λ_{ℓ}. This makes it difficult to recall the $\mathcal{G}_{\hbar}$ firms with λ_{ℓ}.

The analyst then determines long-term growth as either λ_{ℓ} or $\lambda_{\hbar}$ according to which is a better match with the context retrieved by $\mathcal{G}_{i}$. Alternatively, the analyst might form a probability weighting as described in Appendix A, in which case firms that

thoughts becoming data (Wachter and Kahana, 2024). In this account, the analyst views G_{iT} together with the name of firm i. Subsequent presentation of firm i together with its data λ_i brings back the context in which i and G_{iT} are viewed together, thus associating λ_i and G_{iT}.

[15] The model predicts that analysts correctly associate $\mathcal{G}_{\ell}$ with λ_{ℓ}.

resemble high-growth firms have a weight of greater than 50% on λ_h. If investors also are subject to this bias, they will bid up the price of these firms, leading them to have higher market-to-book and price-earnings ratios. The probability of a firm being assigned to a high-growth category is equal to $P(\mathcal{G}_h)$. The subset of these firms for which $\lambda_i = \lambda_\ell$ will be assigned the incorrect (higher) growth forecast and will receive a higher valuation.

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 increases, the true conditional probability $P(\lambda_h|\mathcal{G}_h)$ will tend to 0 or 1. Hence, applying Theorem 3, the agents will recognize low-growth firms for what they are, and realized stock returns will tend to be predictably lower for firms which ex ante were expected to have higher long-term growth rates.

The model thus explains (1) the relation between past growth and future expected growth in the LTG data, as well as in the short-run expectations data; (2) the disappointing subsequent earnings of such firms; (3) the underperformance of high LTG firms in terms of returns; and (4) the underperformance of firms with high book-to-market or earnings-to-price ratios.

4.3 Excess volatility

The time series of stock prices, investment relative to GDP, and unemployment present their own sets of puzzles from the perspective of traditional finance and macroeconomics. Stock market valuations, real investment, and unemployment rates all share a level of volatility that traditional models of rational investors struggle to explain.[16]

Our model of the representativeness heuristic potentially offers a parsimonious explanation for this excess volatility. Models of stock market valuation, as well as investment, and unemployment (which stems from the decisions of firms to invest in hiring workers), have as a state variable a forecast for the future distribution of profitability.[17]

[16] See Campbell and Shiller (1988), Gourio (2012), Bordalo et al. (2019b), Hall (2017).

[17] See, for example, Guo and Wachter (2025), Bordalo et al. (2019b), Gomes et al. (2019), and Kilic

When profit opportunities are more favorable, valuations are higher, and firms have the incentive to invest both in physical capital and in hiring.[18]

Assume that the state of the economy is summarized by a latent binary variable $z_t \in \{0, 1\}$, with $z_t = 1$ indicating a recession. The analyst, investor, or manager views a panel of economic indicators $\mathcal{G}_t$. Similarly to the previous section, for simplicity we assume that $\mathcal{G}_t \in \{\mathcal{G}_\imath, \mathcal{G}_n\}$. Agents observe a previous time series of $\mathcal{G}_t$ followed by outcomes z_{t+1}. As in the previous section, the outcomes z_{t+1} bring to mind the prior $\mathcal{G}_t$, so that these are encoded in the same or similar contexts.[19] Let $\boldsymbol{x}_\imath$ (recession) denote the context retrieved when presented with the target $z_{t+1} = 1$ and $\boldsymbol{x}_n$ denote the context when presented with the target $z_{t+1} = 0$. We are again in the setting of Theorem 1. Agents judge the likelihood of a recession using:[20]

$$\boldsymbol{x}_{\mathcal{G}_t} \cdot \boldsymbol{x}_\imath > \boldsymbol{x}_{\mathcal{G}_t} \cdot \boldsymbol{x}_n,$$

Specifically, current economic indicators $\mathcal{G}_t$ evoke a mental context. Agents ask themselves whether this context is a better match with a recession or a normal context.

The difficulty is that economic growth is by and large unpredictable. Both $\mathcal{G}_t = \mathcal{G}_\imath$ and $\mathcal{G}_t = \mathcal{G}_n$ are usually associated in the data with normal times $z_{t+1} = 0$. However, occasionally $\mathcal{G}_t = \mathcal{G}_\imath$ is associated with $z_{t+1} = 1$. Even though a Kalman filter would infer a low probability of recession from $\mathcal{G}_t = \mathcal{G}_\imath$, agents are nonetheless reminded of a recession because it evokes the recession context.

These beliefs then filter through the economy. In forecasting a recession, agents

and Wachter (2018).

[18] Interest rates serve as a countervailing force. If agents' beliefs are in regard to corporate earnings only, as opposed to broader economic growth, then interest rates would remain roughly constant or would partially but not fully adjust to the point that the decline in interest rates would offset future cash flow expectations. Alternatively, a standard assumption in these models is to assume that beliefs affect interest rates equally, but that the elasticity of intertemporal substitution exceeds 1, so the interest rate is relatively stable.

[19] See note 14.

[20] This inequality could determine whether $\text{Prob}_t(z_{t+1} = 1) = 1$ or 0. Alternatively, a probability weight greater than 50% could be assigned to a recession using the methods in Appendix A.

lower their valuation of the stock market (Guo and Wachter, 2025).[21] Firms reduce hiring and capital expenditures. This generates volatile employment and investment that is correlated with stock prices. However, when the actual recession is not realized, stock valuations rise on surprisingly good earnings news. Investment and hiring, however, are more volatile than under rational inference, and, if the response is sufficiently strong, could create a self-fulfilling prophecy, especially if the belief regarding a recession or worse propagates through the economy.

5 Conclusion

This paper provides a cognitive foundation for Kahneman and Tversky's representativeness heuristic based on associate principles of memory. Our key insight is that, while apparently probabilistic in nature, the questions posed by the Kahneman and Tversky experiments are in fact about similarity. Thus, they are amenable to the tools of associative memory and, in particular, of retrieved context. Whether or not items are experienced under the same context governs whether the agent or subject views such items as similar, and ultimately drives likelihood judgements.

The associative memory approach leads to a theory of inference that is highly non-Bayesian but also tightly disciplined by both theory and experimental observation. We build on the foundational work of Bordalo et al. (2021) to show that our model can not only explain the qualitative and narrative-based Kahneman and Tversky results, but also provide a quantitative fit to the experimental results obtained in a controlled setting. We build on a previous model of the paired associates (specifically associative recognition) task. Indeed, offering a unified explanation of such seemingly disparate paradigms as paired associates and representativeness is part of the appeal of our approach.

[21] As Samuelson memorably puts it: "To prove that Wall Street is an early omen of movements still to come in GNP, commentators quote economic studies alleging that market downturns predicted four out of the last five recessions. That is an understatement. Wall Street indexes predicted nine out of the last five recessions!" (Samuelson, 1966).

The reason that contextual retrieval is able to explain the representativeness heuristic is that it naturally provides a role for observations in which targets co-occur with decoy distributions. What matters isn't, for example, how many high growth companies turn out to resemble Nvidia, but rather how many low growth companies do not. The futures of low-growth companies erode contextual similarity, even though they have no role in Bayesian reasoning. At the same time, this cognitive mechanism is "almost right" and is thus challenging to correct with conscious effort. This may help explain the persistence of financial anomalies.

Our model successfully demonstrates the role of the kernel of truth and the role of alternative associations, as well as providing an excellent quantitative fit to experimental data. The conjunction fallacy emerges naturally when composite representations simultaneously bring to mind multiple associations. We also illustrate how the model can be used to explain cross-sectional anomalies such as the value premium and time-series phenomena such as excess volatility.

Several avenues for future research emerge from this work. First, the model predicts that the representativeness heuristic arises from the dynamics of paired associates. Further experiments could explore this link. Second, while highly tractable, the model abstracts from essential principles of memory, such as temporal contiguity. Re-introducing these memory features will eventually be important in producing a unified theory. Third, the model may provide the potential to explain other behavioral anomalies, such as the disposition effect (Yeung et al., 2025).

Finally, while the representativeness heuristic leads to systematic errors in probability judgment, it may serve an adaptive function. The examples presented here have as a key feature that they involve asymmetric associations. It may be that in cases where associations are more symmetric, the heuristic does not face these problems and may have advantages in computability. Further work could also elucidate these advantages, pointing to a more general theory of cognition.

Appendix

A Questions regarding probability and number

Throughout the main text, we focus on what the model implies for the relative likelihood of hypotheses—that is, whether an agent judges one hypothesis to be more likely than another. However, the model can be naturally extended to explain a broader range of phenomena, including those in which agents state a probability.

As before, free recall is *not* required—the agent need not generate features from memory. For example, in the Bordalo et al. (2021) experiment, in which agents are rapidly shown a set of words or numbers, it would not be realistic for subjects to remember specific instances and then count them up. Rather, subjects translate the similarities into probabilities. The Luce choice rule (Luce, 1959) provides a natural way to do this.

Let $F(\cdot)$ be an increasing continuous function defined on the non-negative real numbers. For example $F(x) = x^\eta$ for $\eta > 0$. Assume the setting in the Model Summary. The Luce choice rule gives the following for elicited probability $\hat{P}$:

$$\hat{P}(h|d) = \frac{F(\boldsymbol{x}_h \cdot \boldsymbol{x}_d)}{\sum_{h' \in \mathcal{H}} F(\boldsymbol{x}_{h'} \cdot \boldsymbol{x}_d)} \tag{A.1}$$

If the question pertains to the number of instances, the agents estimate a quantity $\hat{N}$ and then apply $\hat{P}(h|d)\hat{N}$. Restricting attention to $F(x) = x^\eta$, the lower is η, the more elicited probabilities will shrink toward equal probabilities. Regardless of η, elicited probabilities will be equal if the inner products are equal and will be one or zero as the relative inner products diverge (preserving the results of Section 3.3).

To summarize:

1. Agents form judgments of relative likelihood based on the cosine similarity of the retrieved context vectors (see Model Summary).

2. Agents assess probabilities based on cosine similarity combined with a Luce choice rule.

3. Agents assess quantities by multiplying probability assessments with estimates of the total number of items.

These assumptions suffice to map the associative recognition task into probabilistic judgment. The model could be easily extended to account for processing noise, as in Section 3.4.

These assumptions only rely on subjects understanding that probabilities sum to one and knowing, roughly, the number of items they have seen. It is straightforward to show that, given these assumptions, the model satisfies Predictions 1–3 of Bordalo et al. (2021). The correlational patterns reported by Bordalo et al. further support the hypothesis of the same fundamental mechanism driving the qualitative likelihood statement and the quantitative probabilistic and numerical ones.[22]

However, there are circumstances where these assumptions may not be reasonable. One, described by Bordalo et al. (2023) and in recent work by Conlon and Kwon (2025), is that the list of possible hypotheses is not obvious. In that case, free recall most likely comes into play, and the probability of any given item may be biased upward or downward based on what the presence of features does to cue or suppress items during free recall. A second is when subjects are explicitly cued with a probability. Because so much of the literature on probabilistic reasoning involves a probability as a cue, we consider one such case in detail.

One example, frequently cited in the context of the representativeness heuristic is Casscells et al. (1978). Casscells et al. 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

[22]The link between judgments of likelihood, probability assessments, and numerical assessments is most likely enhanced when subjects are asked these questions as part of the same experiment, as in Bordalo et al.. In this case, the initial answer given by the subject becomes a data point (Wachter and Kahana, 2024; Mullainathan, 2002) in their memory database.

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 the participants, was 95%.[23] 18% of the participants gave the correct answer.

This and similar studies (Eddy, 1982) are subject to the criticism that they are asking agents to solve a probability problem, whereas in real-life settings, agents form judgements based on their experience. Suppose, however, we were to assume that the example was an accurate representation of cases considered by physicians and stored in their memory. Our model would predict that physicians would give an overstated probability of disease, but one that nonetheless is below 50%. However, there is no reason to believe that the example is an accurate representation; physicians may generally handle conditions with greater prevalence. Moreover, 95% itself represents a cue. Thus subjects, drawing from their memory, believe that the test makes disease more likely than otherwise, and then produce the number that is given to them. Bordalo et al. (2025) consider several examples of incorrect probabilistic reasoning, emphasizing the importance of the cue.

B Manipulation of recall through cues

Not surprisingly, this model which is designed to account for the role of cues, also accounts for cue-dependent recall as demonstrated by the third study of Bordalo et al. (2021). Study 3 manipulates Study 1 by introducing a third attribute to the items: font size.[24] All subjects are shown 10 large orange numbers, 15 small blue numbers, and 25

[23] This exhibits kernel of truth because $P(h|d) = P(h,d)/P(d) \approx 0.001/0.05 = 1/50$, 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 $P(d) = P(h,d) + P(-h,d) = 0.001 + 0.05 \times 0.999 \approx 0.05$. However, subjects are not reversing conditional probabilities. The model response of 0.95 is $P(h|d)$, nor $P(d|h)$, but rather $P(-d|-h)$, or one minus the false positive rate of $P(d|-h)$.

[24] In Study 1, subjects are shown 10 orange numbers, 15 blue numbers, and 25 blue words. After observing the sequence, subjects are asked a series of questions:

Q1 "An image was randomly drawn from the images that were just shown to you. The chosen

large blue words. There are then two treatments. Subjects in the color treatment are asked Q1 (what is the likely color of a number) and Q2 (what is the probability that a number is orange). Subjects in the size treatment are instead asked about size:

Q5 "An image was randomly drawn from the images that were just shown to you. The chosen image showed a number. What is the likely font size of a randomly drawn number?"

Q6 "An image was randomly drawn from the images that were just shown to you. The chosen image showed a number. What is the probability that a randomly drawn number is large?"

Regardless of the treatment, the Bayesian response should be the same. The goal of the study is therefore to study the importance of the cue, specifically to test whether the associations formed with the large blue words affect the judgments, despite their irrelevance to the questions. However, Bordalo et al. (2021) show in the data that

Bordalo et al. (2021) predict that the in the color treatment, agents are more likely to recall that the item is orange than they are that the item is large in the size treatment, which is in fact what the data show.

Our model makes the same prediction. Intuitively, in the color treatment, the presence of large *blue* words weakens the association between number and blue; whereas in the size treatment, the presence of *large* blue words weakens the association between number and orange (since both are large).

Formally, in the color treatment, the agent evaluates the inner products:

$$\boldsymbol{x}_{\text{Blue}} \cdot \boldsymbol{x}_{\text{Number}} = \frac{P(\text{Blue}, \text{Number})}{\sqrt{P(\text{Blue})P(\text{Number})}} = \frac{15}{\sqrt{15+25}\sqrt{25}},$$

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 that the number is orange?"

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

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

and

$$\boldsymbol{x}_{\text{ORANGE}} \cdot \boldsymbol{x}_{\text{NUMBER}} = \frac{P(\text{ORANGE}, \text{NUMBER})}{\sqrt{P(\text{ORANGE})P(\text{NUMBER})}} = \frac{10}{\sqrt{10}\sqrt{25}}.$$

In the size treatment, the agent instead evaluates

$$\boldsymbol{x}_{\text{SMALL}} \cdot \boldsymbol{x}_{\text{NUMBER}} = \frac{P(\text{SMALL}, \text{NUMBER})}{\sqrt{P(\text{SMALL})P(\text{NUMBER})}} = \frac{15}{\sqrt{15}\sqrt{25}}$$

and

$$\boldsymbol{x}_{\text{LARGE}} \cdot \boldsymbol{x}_{\text{NUMBER}} = \frac{P(\text{LARGE}, \text{NUMBER})}{\sqrt{P(\text{LARGE})P(\text{NUMBER})}} = \frac{10}{\sqrt{10+25}\sqrt{25}}.$$

In the color treatment, the ratio of inner products is

$$\frac{\boldsymbol{x}_{\text{ORANGE}} \cdot \boldsymbol{x}_{\text{NUMBER}}}{\boldsymbol{x}_{\text{BLUE}} \cdot \boldsymbol{x}_{\text{NUMBER}}} \approx 1.3,$$

whereas in the size treatment, the corresponding ratio is

$$\frac{\boldsymbol{x}_{\text{LARGE}} \cdot \boldsymbol{x}_{\text{NUMBER}}}{\boldsymbol{x}_{\text{SMALL}} \cdot \boldsymbol{x}_{\text{NUMBER}}} \approx 0.4.$$

Thus, the model predicts that subjects are more likely to judge the number to be orange in the color treatment versus large in the size treatment. Likewise, under a Luce choice rule, the subjective probability of orange/large will be higher in the color treatment.

C Alternative model fit to Bordalo et al. (2021)

In the main text, we estimate σ using all moments in Figures 1 and 2 of Bordalo et al. (2021). Here, we show that we can alternatively calibrate the value of σ^2 to a single moment: the proportion of subjects who judged orange to be more likely in the gray treatment. From the computations in the main text and Figure 1 of Bordalo et al. (2021), σ^2 solves

$$P(-0.2027 > \tilde{\epsilon}) = 0.295.$$

Letting $\Phi(z)$ denote the standard normal CDF, this means that[25]

$$\hat{\sigma} = \frac{-0.2027}{\Phi^{-1}(0.295)} \approx 0.3762.$$

Given this calibration, we can then compute the model-implied moments for the blue treatment in Figures 1 and 2 of Bordalo et al. (2021) from the normal CDF, as described above. Figure 5 compares the data in their Figure 1 to that implied by our model. Figure 6 plots the comparison for their Figure 2.

[25]The threshold -0.20 is $z = \Phi^{-1}(0.3)$ standard deviations σ away from zero, so $-0.20 = \sigma z$.

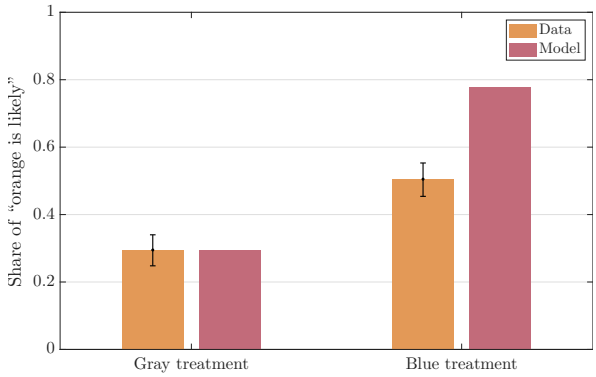
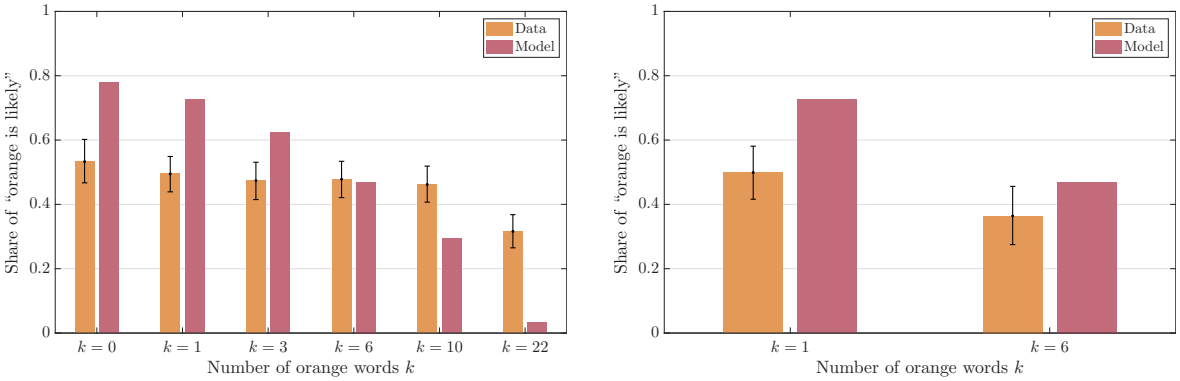


Figure 5: Study 1, Figure 1 of Bordalo et al. (2021) The noise parameter σ is calibrated to match the gray treatment.



(a) Bordalo et al. Figure 2a

(b) Bordalo et al. Figure 2b

Figure 6: Study 2, Figure 2 of Bordalo et al. (2021) The noise parameter σ is calibrated to match the gray treatment in Figure 1.

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., Conlon, J. J., Gennaioli, N., Kwon, S. Y., and Shleifer, A. (2025). How people use statistics. *Review of Economic Studies*.

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.

Campbell, J. Y. and Shiller, R. J. (1988). The dividend-price ratio and expectations of future dividends and discount factors. *Review of Financial Studies*, 1(3):195–228.

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.

Conlon, J. J. and Kwon, S. Y. (2025). Beliefs from cues. Working paper, Brown and Carnegie Mellon Universities.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In Kahneman, D., Slovic, P., and Tversky, A., editors, *Judgment under Uncertainty: Heuristics and Biases*, pages 249–267. Cambridge University Press.

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.

Fama, E. F. and French, K. R. (1992). The cross-section of expected returns. *Journal of Finance*, 47:427–465.

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

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

Gomes, J. F., Grotteria, M., and Wachter, J. A. (2019). Cyclical dispersion in expected defaults. *Review of Financial Studies*, 32(4):1275–1308.

Gourio, F. (2012). Disaster risk and business cycles. *American Economic Review*, 102(6):2734–2766.

Guo, H. and Wachter, J. A. (2025). "superstitious" investors. *Review of Asset Pricing Studies*, 15(1):1–45.

Hall, R. E. (2017). High discounts and high unemployment. *The American Economic Review*, 107(2):305–330.

Healey, M. K., Long, N. M., and Kahana, M. J. (2019). Contiguity in episodic memory. *Psychonomic Bulletin & Review*, 26(3):699—720.

Hockley, W. E. (1992). Item versus associative information: Further comparisons of forgetting rates. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18:1321–1330.

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

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

Kahana, M. J. (2012). *Foundations of Human Memory*. Oxford University Press, New York, NY.

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.

Kilic, M. and Wachter, J. A. (2018). Risk, unemployment, and the stock market: A rare-event-based explanation of labor market volatility. *The Review of Financial Studies*, 31(12):4762–4814.

La Porta, R. (1996). Expectations and the cross-section of stock returns. *The Journal of Finance*, 51(5):1715–1742.

Lettau, M. and Wachter, J. A. (2007). Why is long-horizon equity less risky? a duration-based explanation of the value premium. *Journal of Finance*, 62:55–92.

Luce, R. D. (1959). *Individual choice behavior*. Wiley, New York.

Manning, J., Hulbert, J., Williams, J., Piloto, L., Sahakyan, L., and Norman, K. (2016). A neural signature of contextually mediated intentional forgetting. *Psychonomic Bulletin & Review*, 23:1534–1542.

Metcalfe, J. (1991). Recognition failure and the composite memory trace in CHARM. *Psychological Review*, 98:529–553.

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 Dennis, S. (2024). Global matching models of recognition memory. In Kahana, M. J. and Wagner, A. D., editors, *Oxford Handbook of Human Memory*, volume 1, pages 895–922. Oxford University Press.

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.

Samuelson, P. A. (1966). Science and stocks. *Newsweek*, page 92.

Stanislaw, H. and Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior research methods, instruments, & computers*, 31(1):137–149.

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.

Yeung, T. L., Liu, R., Wachter, J. A., Kahana, M. J., and Zhang, Y. (2025). Navigating through fear and greed: The experience-driven disposition effect. Working paper, Tianjin University and Universita della Swizera Italiana and University of Pennsylvania.