Associative learning and representativeness*

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

November 20, 2024

Abstract

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

Preliminary, Comments welcome

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

*We thank Sudeep Bhatia, Pedro Bordalo, Spencer Kwon, Alice Healy, Peter Maxted, Nikolai Roussanov, John Sakon and participants at the 2020 Society of Financial Studies Cavalcade and at the Spring 2024 Memory Beliefs and Choice Meeting for helpful comments. We are grateful for financial support from NIH grant MH55687.

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

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

§Department of Finance, The Wharton School, University of Pennsylvania and the Securities and Exchange Commission; Email: jwachter@wharton.upenn.edu. The Securities and Exchange Commission disclaims responsibility for any private publication or statement of any SEC employee or Commissioner. This article expresses the author's views and does not necessarily reflect those of the Commission, the Commissioners, or other members of the staff.

1 Introduction

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

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

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

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

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

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

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

2 What is representativeness?

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

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

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

Example 2 (Casscells et al. (1978), as cited by Bordalo et al. (2019a))**.** *Doctors are*

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

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

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." Participants were then asked to rank, in order of likelihood, the possibilities given in Table 1. Over 80% of participants rank, from more to less likely, F>T&F>T for Linda and A>A&J>J for Bill. Regardless of beliefs, this likelihood ranking is a logical impossibility. Tversky and Kahneman

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

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

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

(1983) refer to the tendency of subjects to rank the conjunction as more likely than one of the constituents as the *conjunctive fallacy*.

Examples 1– 3 all display a *kernel of truth*: computer science students are more likely to be introverted than those in the general population. Those that have a condition are more likely to test positive for that condition. Social-justice-oriented young women are more likely to be feminists. 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. That is, while subjects exhibit evidence of non-rational expectations (the elicited subjective and the physical distributions clearly fail to match), the inference does not seem *that* irrational – we can see ourselves making a similar mistake. Many of these examples are so well-known because of a sense we have that they are getting at a very real cognitive deficit to which humans are subject, analogous to an optical illusion.

To formalize these notions, assume there is a population of interest, denoted Ω. The examples above generally relate involve pairs, so we think of each member of the population having a pair of features (for example, being a number and being blue, or being an introvert and being a salesman). As in a standard paired associates task,

one element of the pair will form the cue. Following Bordalo et al. (2021), we refer to the set of possible cues as $\mathcal{D} = \{d_1, \ldots, d_K\}$; this is because, under an analogy with Bayesian updating, the cue forms the "data" on which the agent conditions. The other element of the pair is the one subject to retrieval based on the cue. Again following Bordalo et al., we refer to this as the *hypothesis* and denote the set of possible hypotheses by $\mathcal{H} = \{h_1, \ldots, h_J\}$.[1] There is a physical probability measure P over the space $\mathcal{H} \times \mathcal{D}$.[2]The experiment or survey elicits a different probability measure, which we denote $\hat{P}$. We do not take a stance as to whether this elicited probability measure should be considered a subjective probability measure.[3]

From these examples we can formulate a general statement of the kernel of truth and the representativeness heuristic.

Definition. *The hypothesis-data pair* (h, d) *exhibits the* kernel of truth *if*

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

or equivalently

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

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

This is the same definition as in Bordalo et al. (2016). The definitions (1) and (2)

[1]In some cases, the possibilites in H are non-exhaustive, and thus there may be other possibilities in the subject's memory database. We will assume these do not intrude on the subject's consideration. This assumption is reasonable because either (a) the set of possibilities lies explicitly before the subject or (b) they are repeated frequently enough so the subject is unlikely to forget them, e.g. colors orange and blue in the experiment below.

[2]Notice that in Example 3 above (the conjunctive fallacy), the experiment elicits probabilistic assessments over pairs of hypotheses. In this case, elements of $\mathcal{H}$ should be viewed as pairs of professions and characteristics such as being a feminist or a jazz player. The hypothesis "Bill is an accountant" is then the set that consists of all pairs of which accountant is an element.

[3]The psychological theory implicit in standard decision models is that the agent carries a subjective probability measure in their head, and updates in a manner perhaps similar to Bayesian updating. Our view is different: rather the agent responds to cues and it is the associations that the agent carries in their head and updates.There are many potential differences, one of which is that responses to cues need not be constrained by the laws of probability.

are equivalent because

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

namely $P(h)$ is a weighted average of $P(h|d)$ and $P(h|-d)$. $P(h|d)$ exceeds the average, if and only if it exceeds the other item in the average.

Definition. *Elicited probabilities for* $(h, d) \in H \times D$ *exhibit the* representativeness heuristic *if* (h, d) *exhibits the kernel of truth, and if*

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

while

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

In other words, observations of d raise the elicited probability of h being true (Inequalities 1,2); however, they do not go so far as to say that h is more probable than the alternative $-h$ (Inequality 4)). The agent, however, thinks that they do (Inequality 3). Inequality 4 is not strictly speaking necessary. Namely, the "heuristic" might lead to an over-response to data, but the agent might be directionally correct.

Note that if $P(d) \geq P(-d)$, the kernel of truth (2) can be rewritten as

$$P(h|d)\frac{P(d)}{P(h)} > P(h|-d)\frac{P(-d)}{P(h)}$$

It then follows from Bayes rule that, given h, one can distinguish between d and $-d$:

$$P(d|h) > P(-d|h).$$

In other words, (3) entails reversal of the conditioning argument.[4] It is correct to infer

[4]Tversky and Kahneman (1974) argue against a link between reversing the conditioning argument and representativeness. Their point is not, however, that representativeness does not sometimes entail the reversal of the conditioning argument, but rather it doesn't always have such an implication. Their

the probability of d from knowledge of h. It is not correct to infer the probability of h from observing d.

3 A cognitive theory of the representativeness heuristic

The forgoing examples raises the question of whether the representativeness heuristic might emerge from social interactions, reinforced by repeated exposures to certain images in the media (actively promoting a simplified view of reality) as opposed to a basic cognitive mechanism. While Bhatia (2017) takes the former view, Bordalo et al. (2021), succeed in reproducing the representativeness heuristic in a laboratory setting, in which there are no prior biases. Their results suggest the latter view. Our approach is therefore to seek a basic cognitive mechanism underlying the representativeness heuristic.

Here, we offer an account of one such basic cognitive mechanism, building on our earlier work (Wachter and Kahana, 2024). Wachter and Kahana take the view that lifetime portfolio decisions were best modeled as free recall, in which the agent simulates future states by freely recalling prior experiences, with weights determined by present context and by the associations between features and context, In contrast, this paper is based on recognition. Indeed, a main insight of this paper is that probabilistic judgement is best thought of as *recognition*, a subject in which there is a long literature (Kahana et al., 2005).[5] Unlike a free recall or serial recall task that involves the production of words (and, secondarily, a decision as to whether to speak the word), the probabilistic judgement task involves no production – the cue provides the possibilities– and only a judgement as to which is more likely. While probabilistic judgement is not modeled in the memory literature directly, it shares a fundamental similarity with

point is that an alternative model in which an agent is Bayesian but has a specific confusion regarding conditioning is not one that accounts for the range of their findings.

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

recognition memory tasks, and most specifically, associative recognition (Kahana et al., 2005).

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

As in Wachter and Kahana (2024), agents, over a lifetime, view features of the environment, which we represent as column vectors f_t, with t being the time at which the agent views the features. We adopt the standard view of assuming that features are basis vectors in a very high dimensional space (Howard and Kahana, 2002). The agent associates these features with an internal mental context x_t, also a column vector. Associations are stored in the memory matrix, which is formed by taking the outer products of features and context:

$$M_t \equiv M_{t-1} + x_t f_t^\top \tag{5}$$

$$= M_0 + \sum_{s=1}^{t} x_s f_s^\top \tag{6}$$

that encodes contexts and features together. Features retrieve context:

$$x_t^{\text{in}} \propto M_{t-1} f_t, \tag{7}$$

which is scaled so that it is a unit vector. Context is updated based on retrieved context (7) and on past context x_{t-1}.

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

The cueing recovers all the contexts previously associated with that feature:

$$\begin{aligned} x_t^{\text{in}} = M_{t-1} f_t &\propto M_0 f_t + \sum_{s=1}^{t-1} (x_s f_s^\top) f_t \\ &\propto M_0 f_t + \sum_{s=1}^{t-1} x_s (f_s^\top f_t). \end{aligned} \tag{8}$$

In the benchmark case of features that are orthonormal basis vectors, $f_s^\top f_t$ equal either to zero or one. It is zero if $f_s \neq f_t$; it is one if $f_s = f_t$. Equation 8 shows that features evoke the past contexts under which they are experienced. A context x_s appears in the sum in (8) if the corresponding f_s equals f_t; otherwise it does not. Thus context is a weighted average of past contexts under which that feature was experienced. This basic contextual retrieval enabled the jump-back-in-time that drove decision making in Wachter and Kahana (2024), and it will drive probabilistic assessments in this paper.

In the associative recall task, subjects are cued with one item of a pair and asked if another, given, item was paired with it on a list (this item is known as the target). Kahana and Jin (2024) evaluate predictions of this model for an associative recall task, assuming the following decision rule: the agent compares the current context to the context under which the target is experienced. That context can be extracted using (7):

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

The current context is that of the cue. The agent determines if there is a match if the inner product, namely cosine similarity, between the cue and the target, exceed some lower bound:

$$x_{\text{Cue}} \cdot x_{\text{Target}} > B$$

for $B \in (0, 1)$. Retrieved contexts are scaled so as to be norm 1. Unlike Wachter and Kahana (2024), these contexts will not be interpreted as conditional probabilities, and so their elements need not be scaled to sum to 1. Rather, we scale these using the more standard L^2-norm, so that the inner product has the interpretation of cosine similarity,

and is bounded between zero and one, zero when the vectors are orthogonal, and one when they are the same. We discuss the importance of this scaling below.

Our argument is that probability judgement tasks, while seemingly concerning probabilities, are in fact questions of associative recognition. There is one difference, which may be superficial but we nonetheless must address. Probabilistic judgement questions take two forms, one is: which [of two or more options] is most likely (or, rank in order of likelihood); the second is: what is the probability [of two or more options]. To the question "which is more likely?" the agent chooses the target with the greatest inner product. Because the inner product is between zero and one, the subject could plausibly translate it into a probability, and use it to answer "what is the probability of [target]?." However, there is no requirement that the inner product across the targets sums to one. Subjects may realize that probabilities sum to one, and thus use this fact to constrain their answers. Therefore, we make the standard assumption that the subject applies a simple Luce choice rule to determine probabilities. The simplest version of this rule takes the value of the inner product of a given target and divides by the sum of the other remaining targets, also provided in the experiment.

Key to our modeling assumptions is that free recall is *not* required – the agent need not generate any features. This has an important implication for interpreting these experiments. Cases in which agents are asked to rank the likelihood of outcomes, as is standard in the Kahneman and Tversky methodology, require no production – one can simply evaluate the best match with the target. Cases in which one is producing probabilities from a given set of possibilities (or the possibilities are obvious because there is a small number and they are repeated many times) can be handled with the Luce choice rule (see Equation 15). However, as Bordalo et al. (2023) emphasize, if asked to produce a probability and one is not given the possibilities, then free recall inevitably comes into play, and the probability of any given item may be biased upward or downward based on what the presence of the features does to cue or suppress items during free recall.[6]

[6]For example, if one were to apply the Luce choice rule to examine probabilities related to the

A key question is what happens when an object has multiple features. Fortunately autoregressive context provides the answer. Multiple features are observed so close in time that we say that they are experienced under the same context (as an approximation for the continuous-time limit of an autoregressive context process, in which contexts would be virtually indistinguishable (and there would be no room for an unrelated feature to intrude on the pair). We thus leave aside the question of the correct way to model perceptual foundations: do we observe multiple features at exactly the same moment in time, or in an infinitesimally close series?

A final assumption, clear in our examples below, is that the agent orthogonalizes context between paired associates, and again, when receiving the cue. While performance in the associative recognition does exhibit dynamics that form the signature of the Howard and Kahana (2002) temporal context model, these dynamics are not the focus of this paper. One implication is that every element in the list has a unique, orthogonal context. A second implication is the that the current context is in fact the retrieved context of the cue. Evidence for suppression of context comes from Osth and Fox (2019), who find little evidence of recency and temporal contiguity. The fact that recency and temporal contiguity still appear in the cued recall task, in which subjects are asked to recall, as opposed to recognize, the second element of the pair (Davis et al., 2008) suggests that a full account of paired associates, and thus probability judgement, will eventually invoke autoregressive context given sufficient data. However, in light of the Osth and Fox evidence, and the fact that autoregressive dynamics are unnecessary to explaining the heuristic, we assume the agent orthogonalizes context between pairs and prior to receiving the cue.

This is a strikingly different view of probabilities than is implied by Bayesian updating. Indeed, in Bayesian updating, agents store the probabilities and update them over time, in a process that, at first glance, resembles (5). An alternative Bayesian

conjunctive fallacy (which Tversky and Kahneman (1983) do not do), it would seem most natural to divide by the elements in the relevant column of Table 1, rather than the implied set of possibilities (e.g. accountant but not a jazz player, accountant and a surfer but not a jazz player).

framework is to assume perfect recall, in which the agent recalls and counts up instances and divides by a total. Tversky and Kahneman (1974, 1983) take a different view – that agents rely on a set of heuristics to compute probabilities.

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

3.1 Application to Bordalo et al. (2021)

Below, we show how to apply the above ideas to the three studies that Bordalo et al. (2021) conduct. In order to make explicit calculations, we start by considering a very stylized version of these studies that will allow us to examine the mechanism in detail.

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

Q2: [same scenario, but then asked] "What is the probability the number is orange?" Finally, on the next screen, participants are asked:

Q3: "How many orange numbers were shown to you?", and

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

Bordalo et al. (2021) shows that their model makes four predictions:

1. The blue treatment reduces the assessed likelihood that the randomly drawn number is blue in Q1 and Q2 relative to the gray treatment. [A Bayesian model predicts no effect.]

2. The blue treatment reduces the share of blue numbers recalled in Q3 and Q4 relative to the gray treatment. There is also a positive correlation at the individual level between the share of blue numbers recalled from Q3 and Q4 and the assessed likelihood of the number being blue in Q1 and Q2.

3. "...the assessed probability that a random number is orange in Q1 and Q2 and the share of recalled orange numbers in Q3 and Q4 decrease with the number of orange words k in the decoy distribution." [A Bayesian model predicts no effect.]

4. "...the share or recalled orange/large numbers and the assessment of the probability that a random number is orange/large is higher in the color treatment than in the size treatment." [A Bayesian model predicts no effect.]

They confirm these in the data. A necessary step, therefore, is to show that our model would make the same predictions.

3.1.1 Study 1

In the first study participants are randomly assigned to one of two treatments. Each participant sees a sequence of 50 of these abstract images. In the first condition, they see 10 orange numbers, 15 blue numbers, and 25 gray shapes (the *gray* treatment). In the second condition, they see 10 orange numbers, 15 blue numbers, and 25 blue words (the *blue* treatment). Participants are then asked several questions concerning the likelihood of a number being orange. Bordalo et al. (2021) found that participants in the blue treatment were significantly more likely to state that the likely color of a randomly-drawn number was orange (50% versus 30%). They also reported a higher probability of the number being orange.

In this experiment, it appears that the blue words "interfere" with the blue numbers. It is almost as if participants say to themselves "words are supposed to be blue; numbers are supposed to be orange." This example replicates a Tversky and Kahneman (1974) notion of representativeness using an experimental paradigm more reminiscent of Estes

(1972), who shows, in several cases, how probability judgements converge to the truth. One can think of the orange numbers as the introverted librarians, who drive the blue numbers (the introverted salesman) out of memory. Note the subject reverses the conditioning argument – all orange items are numbers, but not all numbers are orange. The agent ignores the base rate – in saying that numbers tend to be orange, the agent neglects the fact that the base rate of blue items is higher than that of orange items.

We illustrate how the model captures the result of Study 1 with a short list, where x_t refers to the context.

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

We model these items in the features space as:

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

Features can be both a number and blue at the same time. We capture this by assuming that the perception of number and blue occur so close in time that they are associated with the same context. For the purpose of this example, all numbers are the same feature and all words are the same feature. We also assume that there is a distraction between each item in the list, so that contexts are orthogonal. This is a realistic assumptions: Agents can introduce their own distractions for the purpose of better memory of the pairs; see Davis et al. (2008). The presence of the distractor im-

plies that, while features that appear jointly in an object are remembered together, the context for the next feature pair is quite different and in fact can be nearly orthogonal. While the idea of a perfectly orthogonal context between lists is an approximation, it is one that allows us to illustrate the main mechanism, as well as derive an analytical formula. It does mean that we cannot the more complex memory dynamics that govern associations among disparate pairs, but this is not our focus.

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

$$\underbrace{\begin{bmatrix}0\\0\\1\\0\end{bmatrix}, \begin{bmatrix}1\\0\\0\\0\end{bmatrix}}_{\text{Blue } '5'} \;//\; \underbrace{\begin{bmatrix}0\\0\\0\\1\end{bmatrix}, \begin{bmatrix}1\\0\\0\\0\end{bmatrix}}_{\text{Orange } '3'} \;//\; \underbrace{\begin{bmatrix}0\\0\\1\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\\0\end{bmatrix}}_{\text{Blue } 'CAT'} \tag{9}$$

$$x_1 = \begin{bmatrix}1\\0\\0\end{bmatrix} \qquad x_2 = \begin{bmatrix}0\\1\\0\end{bmatrix} \qquad x_3 = \begin{bmatrix}0\\0\\1\end{bmatrix}$$

The mental representation (9) implies that the context-to-features matrix equals

$$
\begin{aligned}
M_T^\top &= M_0^\top + \sum_{t=1}^{T} f_t x_t^\top \\
&= M_0^\top + \left(\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right) \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} + \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right) \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \\
&\quad + \left(\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \right) \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} .
\end{aligned} \tag{10}
$$

We abstract away from pre-experimental associations by setting all elements of M_0 to zero. Then, suppressing time subscripts,

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

and

$$
M = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} \tag{11}
$$

Participants are first asked: What is the likely color of the number? The cue is the

feature 'Number,' which retrieves context

$$x_{\text{NUMBER}} \equiv \frac{Mf_{\text{NUMBER}}}{||Mf_{\text{NUMBER}}||} \propto M \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \propto \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{bmatrix}. \tag{12}$$

As in Kahana and Jin (2024), the participant assesses the similarity of 'Number' to 'Blue' and to 'Orange' by comparing their retrieved contexts:

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

and

$$x_{\text{ORANGE}} \equiv \frac{Mf_{\text{ORANGE}}}{||Mf_{\text{ORANGE}}||} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \tag{14}$$

Note how:

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

with the difference being due to the blue words. Participants for whom this is true report that orange numbers are more likely.

The model also translates into probabilities, as in Bordalo et al. (2021) Q2: "What is the probability that the number is orange?" with a Luce choice rule:

$$p_{\text{ORANGE}} = \frac{F(x_{\text{NUMBER}} \cdot x_{\text{ORANGE}})}{F(x_{\text{NUMBER}} \cdot x_{\text{ORANGE}}) + F(x_{\text{NUMBER}} \cdot x_{\text{BLUE}})} \tag{15}$$

where $F(\cdot)$ is a monotonically increasing function. Typical examples include power or exponential.

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

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

with corresponding mental representation:

$$\begin{array}{ccccc} \text{Blue } '5' & // & \text{Orange } '3' & // & \text{Gray } \square \\ \underbrace{\begin{bmatrix}0\\0\\1\\0\\0\end{bmatrix}, \begin{bmatrix}1\\0\\0\\0\\0\end{bmatrix}} & & \underbrace{\begin{bmatrix}0\\0\\0\\1\\0\end{bmatrix}, \begin{bmatrix}1\\0\\0\\0\\0\end{bmatrix}} & & \underbrace{\begin{bmatrix}0\\0\\0\\0\\1\end{bmatrix}, \begin{bmatrix}0\\1\\0\\0\\0\end{bmatrix}} \\ x_1 & & x_2 & & x_3 \\ \begin{bmatrix}1\\0\\0\end{bmatrix} & & \begin{bmatrix}0\\1\\0\end{bmatrix} & & \begin{bmatrix}0\\0\\1\end{bmatrix} \end{array} \tag{16}$$

Applying the same reasoning as in (10) implies that the memory matrix equals

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

with

$$M = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix} \tag{17}$$

Participants are asked: "what color is the number?" The cue is number. Retrieved number context is identical to (12). The target is color. Each color retrieves a basis vector context:

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

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

3.1.2 Study 2

Bordalo et al. (2021) then examine what happens to probability assessment when they introduce orange words to the list. To examine the implication of our model for this manipulation consider lengthening the short list to add one orange word. The list is

now:

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

with corresponding representation:

$$\underbrace{\begin{bmatrix}0\\0\\1\\0\end{bmatrix},\begin{bmatrix}1\\0\\0\\0\end{bmatrix}}_{\text{Blue } '5'} \;//\; \underbrace{\begin{bmatrix}0\\0\\0\\1\end{bmatrix},\begin{bmatrix}1\\0\\0\\0\end{bmatrix}}_{\text{Orange } '3'} \;//\; \underbrace{\begin{bmatrix}0\\0\\1\\0\end{bmatrix},\begin{bmatrix}0\\1\\0\\0\end{bmatrix}}_{\text{Blue } '\text{CAT}'} \;//\; \underbrace{\begin{bmatrix}0\\0\\0\\1\end{bmatrix},\begin{bmatrix}0\\1\\0\\0\end{bmatrix}}_{\text{Orange } '\text{DOG}} \tag{18}$$

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

Following the reasoning of (10), we have:

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

so that

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

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

$$M = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}. \tag{19}$$

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

$$x_{\textsc{Number}} \propto M \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \propto \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \\ 0 \end{bmatrix}. \tag{20}$$

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

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

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

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

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

3.1.3 Study 3

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

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

Number	Word	Blue	Orange	Small	Large
↕	↕	↕	↕	↕	↕
$\begin{bmatrix}1\\0\\0\\0\\0\\0\end{bmatrix}$	$\begin{bmatrix}0\\1\\0\\0\\0\\0\end{bmatrix}$	$\begin{bmatrix}0\\0\\1\\0\\0\\0\end{bmatrix}$	$\begin{bmatrix}0\\0\\0\\1\\0\\0\end{bmatrix}$	$\begin{bmatrix}0\\0\\0\\0\\1\\0\end{bmatrix}$	$\begin{bmatrix}0\\0\\0\\0\\0\\1\end{bmatrix}$

In Bordalo et al. (2021), the blue numbers are small, but everything else is large.

Therefore, consider the following list and corresponding mental representation:

$$\underbrace{\begin{matrix}\text{Small, Blue } '5' \\ \begin{bmatrix}0\\0\\0\\0\\1\\0\end{bmatrix}, \begin{bmatrix}0\\0\\1\\0\\0\\0\end{bmatrix}, \begin{bmatrix}1\\0\\0\\0\\0\\0\end{bmatrix}\end{matrix}}_{x_1} \quad // \quad \underbrace{\begin{matrix}\text{Large, Orange } '3' \\ \begin{bmatrix}0\\0\\0\\0\\0\\1\end{bmatrix}, \begin{bmatrix}0\\0\\0\\1\\0\\0\end{bmatrix}, \begin{bmatrix}1\\0\\0\\0\\0\\0\end{bmatrix}\end{matrix}}_{x_2} \quad // \quad \underbrace{\begin{matrix}\text{Large, Blue } '\text{CAT}' \\ \begin{bmatrix}0\\0\\0\\0\\0\\1\end{bmatrix}, \begin{bmatrix}0\\0\\1\\0\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\\0\\0\\0\end{bmatrix}\end{matrix}}_{x_3} \quad 1 \tag{21}$$

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

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

whereas M adds columns

$$M = \begin{bmatrix}1&0&1&0&1&0\\1&0&0&1&0&1\\0&1&1&0&0&1\end{bmatrix} \tag{22}$$

Because the first column is unaffected, context retrieved by feature 'Number' is the

same as (12):

$$x_{\text{NUMBER}} \propto Mf_{\text{NUMBER}} \propto \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{bmatrix}. \tag{23}$$

When asked, "what is the likely color of the number," the participant once again compares similarities of retrieved contexts of blue and orange to (23), obtaining the same answers as in Study 1:

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

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

If the participants were, however, remembering more orange numbers, one might think they might also remember more large numbers, since the orange numbers are large whereas the blue (forgotten) numbers are small. However, this is not what occurs.

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

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

and

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

and therefore

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

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

3.1.4 What happens when we increase the proportion of blue numbers in the list?

When adding a list item, we add a context, and therefore a row to M:

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

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

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

whereas

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

We see that

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

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

We can use this example to gain intuition for how elicited probabilities change with an increase in the sample size or the proportion of one type of item versus another as part of the observation set. Let N be the sample size and $\mathcal{L}$ for the empirical likelihood density so that, for example, $\mathcal{L}(\textsc{Orange})N$ is the count of orange items on the list, and $\mathcal{L}(\textsc{Number}, \textsc{Orange})N$ is the count of orange numbers. Note that feature ORANGE will retrieve a context vector with $\mathcal{L}(\textsc{Orange})N$ nonzero elements. Because the context vector must have length equal to unity, each of these elements equals $1/\sqrt{\mathcal{L}(\textsc{Orange})N}$. Analogous results hold true for feature NUMBER. Now consider the inner product of these two context vectors. Because the inner product takes the sum of the nonzero elements, the sum is taken over at most $\min(\mathcal{L}(\textsc{Number}), \mathcal{L}(\textsc{Orange}))N$ items. In Study 1, it is taken over exactly $\mathcal{L}(\textsc{Orange})N$ of them, not only because there are fewer orange items, but because in tall orange items are numbers. If there were some orange items that were not numbers, the context overlap would be less than perfect, and there would be fewer than $\mathcal{L}(\textsc{Orange})N$ nonzero elements. That, is not all orange items occur in the number context. In both cases, however, the number of nonzero elements scales with N. Thus the inner product does not depend on N. In the specific case of orange numbers, the inner product equals $\mathcal{L}(\textsc{Number}, \textsc{Orange})/\sqrt{\mathcal{L}(\textsc{Number})\mathcal{L}(\textsc{Orange})}$, where $\mathcal{L}(\textsc{Number}, \textsc{Orange})$ is the joint likelihood of an object being both a number and the color orange.

This reasoning highlights the importance of using L^2 scaling. For if we were to use L^1 scaling, the inner product would actually fall with the sample size, a counter-intuitive outcome. Each of the elements in the context vector associated with feature ORANGE would equal $1/\mathcal{L}(\textsc{Orange})N$, whereas each item in the context vector associated with feature NUMBER equals $1/\mathcal{L}(\textsc{Number})N$. The inner product would equal $\mathcal{L}(\textsc{Number}, \textsc{Orange})/(N\sqrt{\mathcal{L}(\textsc{Number})\mathcal{L}(\textsc{Orange})})$, a decreasing function of N. While one could redefine cognitive processes through cosine similarity (namely, use not the inner product but the scaled inner product), and thus recover the same result under

context vectors scaled under the L^1 norm, such an approach introduces an extra step without adding explanatory power.

3.2 General results

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

Our model makes the simple and elegant prediction that

$$\hat{P}(h|d) \propto F(x_h \cdot x_d) = F\left(\frac{P(h,d)}{\sqrt{P(h)P(d)}}\right), \tag{24}$$

and moreover, that the instance judged to be most likely is the one with the highest value for $P(h, d)/\sqrt{P(h)P(d)}$. Here $F(\cdot)$ is an increasing function that forms the input into the Luce choice rule, The first equality follows from our recall mechanism, whereas the second follows from the reasoning above that links the dot product to $\mathcal{L}$.

Note that

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

and that $P(d)$ is a constant of proportionality. Thus if F if separable, (24) implies

$$\hat{P}(h|d) \quad \propto \quad F\left(\frac{P(h,d)}{\sqrt{P(h,d) + P(h,-d)}}\right) \tag{25}$$

Note that F is strictly increasing. It follows that $\hat{P}(h|d)$ is strictly increasing in $P(h, d)$ and decreasing in $P(h, -d)$.[7] It then follows from arguments in (Bordalo et al., 2021, Appendix A) that the four predictions hold.

We now turn to the predictions of Bordalo et al. (2021). Prediction 1 states that

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

the subject is less likely to state that the color of the number is blue (Q1), and that the probability of the number is blue (Q2) in the blue treatment as compared to the gray treatment.

We apply (24), first using the values from the blue treatment: 10 orange numbers, 15 blue numbers, and 25 blue words. Recall that $h = \textsc{Blue}$ and $d = \textsc{Number}$. In particular, we find that when BLUE is the target:

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

Compare this with the case of ORANGE as target:

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

There are more blue numbers, but their probability is penalized by the large number of blue words. On the other hand, when we use the values from the gray treatment, in which the 25 blue words are replaced by 25 gray shapes:

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

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

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

While comparing inner products determines which color serves as a better match,

and hence which is recognized as appearing with number, the inner products do not sum to 1. Thus to determine elicited probabilities, we assume the agent uses some version of the Luce choice rule. It suffices that $F(\cdot)$ be increasing to obtain the result that the likelihood of blue numbers is lower under the blue treatment as compared with the gray treatment. This shows the model satisfies Prediction 1.

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

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

$$\hat{P}(h \,|\, d) \propto \left(\frac{P(h,d)}{\sqrt{P(h)P(d)}} \right)^{\eta}, \tag{26}$$

for some parameter $\eta > 0$. We can calibrate η using Study 1 of Bordalo et al. (2021), who find that participants in the blue treatment estimated a 7-percentage-point higher probability of orange than did those in the gray treatment. In both the blue and gray treatments, there are 10 orange numbers and no orange words, so (26) implies

$$\hat{P}(\textsc{Orange} \,|\, \textsc{Number}) \propto \left(\frac{P(\textsc{Orange}, \textsc{Number})}{\sqrt{P(\textsc{Orange})P(\textsc{Number})}} \right)^{\eta} = \left(\frac{10}{\sqrt{10}\sqrt{15+10}} \right)^{\eta}. \tag{27}$$

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

$$\hat{P}_{\text{gray}}(\textsc{Blue} \,|\, \textsc{Number}) \propto \left(\frac{P(\textsc{Blue}, \textsc{Number})}{\sqrt{P(\textsc{Blue})P(\textsc{Number})}} \right)^{\eta} = \left(\frac{15}{\sqrt{15}\sqrt{15+10}} \right)^{\eta}. \tag{28}$$

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

$$\hat{P}_{\text{blue}}(\textsc{Blue} \mid \textsc{Number}) \propto \left(\frac{P(\textsc{Blue}, \textsc{Number})}{\sqrt{P(\textsc{Blue})P(\textsc{Number})}} \right)^{\eta} = \left(\frac{15}{\sqrt{15+25}\sqrt{15+10}} \right)^{\eta}. \quad (29)$$

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

$$\frac{(10/\sqrt{10})^{\eta}}{(10/\sqrt{10})^{\eta} + (15/\sqrt{40})^{\eta}} - \frac{(10/\sqrt{10})^{\eta}}{(10/\sqrt{10})^{\eta} + (15/\sqrt{15})^{\eta}} = 0.07. \quad (30)$$

Solving for η implies the estimate

$$\eta \approx 0.57.$$

We now turn to Prediction 3, which states the the assessed probability that a random number is orange decreases with the number of orange words k in the decoy distribution (see also Section 3.1.2 which discusses Bordalo et al. (2021) Study 2, designed to address this prediction). We can easily see the model satisfies this prediction by using the formula (24). In particular,

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

and

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

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

Prediction 4 contrasts the *color* treatment with a *size* treatment, and relies on the results of Bordalo et al. (2021) Study 3 i (see Section 3.1.3). Study 3 shows a sequence

of 50 images as before. Except in this case, the images vary according to a third attribute, name size. All orange numbers are large, and all blue numbers are small, while all words are blue and large.

The color treatment consists of the Q1 and Q2 defined above. The size treatment replaces Q1 with "What is the likely font size of a randomly drawn number?" and "What is the probability of a randomly drawn number being large?" This question presents a direct contrast with Bayesian inference, in that all the large numbers are orange and all the small numbers are blue. Thus, under unbiased recall, the actual objects elicited by the questions in the size and the color treatments are identical. However, as Section 3.1.3 shows, the cue determines which items the agent recalls.

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

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

and

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

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

3.3 Comparison with the Bordalo et al. (2021) Model

Bordalo et al. (2021) posit that the agent assesses the probability of h using a Luce

choice rule:[8]

$$\tilde{P}(h \mid d) = \frac{S(h,d)}{\sum_{h' \in H} S(h',d)}, \tag{31}$$

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

$$S(h,d) = P(h,d)^{\alpha} P(-h,d)^{-\beta} P(h,-d)^{-\gamma}, \tag{32}$$

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

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

$$\tilde{P}(h \mid d) \propto P(h \mid d) \left[\frac{P(h \mid d)}{P(h \mid -d)} \right]^{\gamma} \tag{33}$$

which reduces to the case of Bayesian updating for $\gamma = 0$. In contrast, interference by alternative data $-d$ occurs for $\gamma > 0$.

Thus in the Bordalo et al. (2021) model, γ is a flexible parameter, introduced for the purpose of fitting the representativeness heuristic. By contrast, our model provides a cognitive foundation for γ, and, in so doing, eliminates γ entirely as a free parameter.[9] It also connects two previously disparate fields of study: the study of recognition, and the study of probability assessment. By drawing this connection, our hope is to stimulate further study in both areas.

4 Conjunctive fallacy

Here we show how the model can capture the conjunctive fallacy of Tversky and Kahneman (1983). Consider subjects asked to rank possibilities for Bill (see Table 1). We

[8] We slightly deviate from their notation here for convenience. The original paper specifies that $\tilde{P}(h \mid d) \propto e^{S(h,d)}$ and then defines $S(h,d)$ as the log of (32).

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

model the relative rankings of Accountant (A), Jazz player (J), and both (A&J).

While Bill as presented in the Tversky and Kahneman (1983) experiment may be a composite feature, we keep the notation simply by assuming it is a basis vector, denoted f_{BILL}. We assume f_{BILL} tends to be seen along with f_{A} and only rarely with f_{J}. Meanwhile, f_{J} is seen frequently in different contexts. As in Section 3.1, we consider the very simplest list that illustrates these ideas:

$$\begin{array}{ccccc} \text{Bill} & & \text{Accountant} & & \text{Jazz player} \\ \updownarrow & & \updownarrow & & \updownarrow \\ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} & & \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} & & \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \end{array}$$

$$\begin{array}{ccccc} \text{Bill who is an accountant} & // & \text{Jazz player} & // & \text{Accountant} \\ \underbrace{\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}} & & \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} & & \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \\ x_1 & & x_2 & & x_3 \\ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} & & \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} & & \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \end{array} \tag{34}$$

The mental representation (34) implies that the context-to-features matrix equals:

$$\begin{aligned} M_T^\top &= M_0^\top + \sum_{t=1}^{T} f_t x_t^\top \\ &= M_0^\top + \left(\left[\begin{array}{c} 1 \\ 0 \\ 0 \end{array} \right] + \left[\begin{array}{c} 0 \\ 1 \\ 0 \end{array} \right] \right) \left[\begin{array}{ccc} 1 & 0 & 0 \end{array} \right] + \left(\left[\begin{array}{c} 0 \\ 0 \\ 1 \end{array} \right] \right) \left[\begin{array}{ccc} 0 & 1 & 0 \end{array} \right] \\ &\quad + \left(\left[\begin{array}{c} 0 \\ 1 \\ 0 \end{array} \right] \right) \left[\begin{array}{ccc} 0 & 0 & 1 \end{array} \right] . \end{aligned} \tag{35}$$

Abstracting from M_0 and suppressing time subscripts gives us:

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

with

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

The cue is the feature 'Bill':

$$x_{\text{Bill}} \equiv \frac{M f_{\text{Bill}}}{||M f_{\text{Bill}}||} \propto \left[\begin{array}{c} 1 \\ 0 \\ 0 \end{array} \right],$$

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

together with Jazz, Accountant, and Jazz):

$$x_{\text{A\&J}} \equiv \frac{M(f_{\text{A}} + f_{\text{J}})}{\|M(f_{\text{A}} + f_{\text{J}})\|} = \begin{bmatrix} 1/\sqrt{3} \\ 1/\sqrt{3} \\ 1/\sqrt{3} \end{bmatrix} \tag{36}$$

$$x_{\text{J}} \equiv \frac{Mf_{\text{J}}}{\|Mf_{\text{J}}\|} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \tag{37}$$

$$x_{\text{A}} \equiv \frac{Mf_{\text{A}}}{\|Mf_{\text{A}}\|} = \begin{bmatrix} 1/\sqrt{2} \\ 0 \\ 1/\sqrt{2} \end{bmatrix} \tag{38}$$

From which we see that

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

Davis, O. C., Geller, A. S., Rizzuto, D. S., and Kahana, M. J. (2008). Temporal associative processes revealed by intrusions in paired-associate recall. *Psychonomic Bulletin & Review*, 15(1):64–69.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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