Cognition 194 (2020) 104057

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

What comes to mind?

Adam Bear[a,*], Samantha Bensinger[d], Julian Jara-Ettinger[b], Joshua Knobe[c], Fiery Cushman[a]

[a] Department of Psychology, Harvard University, Cambridge, MA 02138, United States
[b] Department of Psychology, Yale University, New Haven, CT 06511, United States
[c] Program in Cognitive Science, Yale University, New Haven, CT 06511, United States
[d] Yale Law School, New Haven, CT 06511, United States

ARTICLE INFO

Keywords:
Sampling
Decision-making
Consciousness
Computation

ABSTRACT

When solving problems, like making predictions or choices, people often "sample" possibilities into mind. Here, we consider whether there is structure to the kinds of thoughts people sample by default—that is, without an explicit goal. Across three experiments we found that what comes to mind by default are samples from a probability distribution that combines what people think is likely and what they think is good. Experiment 1 found that the first quantities that come to mind for everyday behaviors and events are quantities that combine what is average and ideal. Experiment 2 found, in a manipulated context, that the distribution of numbers that come to mind resemble the mathematical *product* of the presented statistical distribution and a (softmax-transformed) prescriptive distribution. Experiment 3 replicated these findings in a visual domain. These results provide insight into the process generating people's conscious thoughts and invite new questions about the value of thinking about things that are both likely and good.

1. Introduction

Of all the amounts of TV that a person might watch in a day, think of one particular amount. Go ahead, give it a try: What is the first *amount of television watching per day* that comes to your mind? Was it one hour? Two? Maybe five or six? There are no right or wrong answers to this question. Clearly, however, some amounts (three) are more likely than others (seventeen).

Our goal is to understand what comes to mind by default—that is, in the absence of an explicit task. Of course, without any task, people's thoughts may be unstructured and uninteresting. In contrast to this possibility, we find structure that is both consistent and unique.

Specifically, we ask whether the things that first come to mind are usefully represented as samples from a probability distribution. Consider again the example of an *amount of television watching per day*. Intuitively, the set of possible amounts that we could call to mind ranges from zero to twenty-four hours. Within this range, some amounts are much more likely to be sampled into mind than others. When sampling, on what basis do people weight *three* more highly than *seventeen*?

One obvious hypothesis is that the mind simply defaults to one of the standard weightings used in ordinary tasks. The existing literature offers two likely candidates. First, for many tasks, people sample from sets of things weighted by a representation of the probability (i.e., frequency) of those things. Such sampling processes are an important starting point for many methods of statistical inference over generative models. For instance, in such a model, a person might encode that 40% of people exercise zero hours a week, 15% exercise one hour, 10% exercise two hours, and so forth—in other words, their generative model is designed to approximate the actual probability with which these types of events occur, are encountered, or are performed in the world (or perhaps would be, given hypothetical conditions). Much prior research shows that people are able to sample from such statistical distributions (e.g., Griffiths & Tenenbaum, 2006; Vul, Goodman, Griffiths, & Tenenbaum, 2014), and indeed may estimate distributional properties precisely by sampling from them (Gershman, Vul, & Tenenbaum, 2012; Icard, 2016; Vul & Pashler, 2008; though see Lieder, Griffiths, & Hsu, 2018). In sum, as a first step in trying to predict or explain events, we often engage in sampling weighted by *probability*.

Second, in order to make choices, people represent the prescriptive *value* of different types of things.[1] For instance, perhaps exercising zero hours a week is pretty bad, three hours is really good, and 60 hours is really bad. To turn values into choice, however, people cannot directly sample some option from their set of represented values, as these values

[*] Corresponding author at: Department of Psychology, Harvard University, 33 Kirkland Street, Cambridge, MA 02138, United States.
E-mail address: adambear@fas.harvard.edu (A. Bear).

[1] To avoid confusion, we use the term "value" throughout the manuscript only to refer to *prescriptive* value. When simply referencing a quantity, we use words like "number" or "amount."

https://doi.org/10.1016/j.cognition.2019.104057
Received 3 April 2019; Received in revised form 6 August 2019; Accepted 20 August 2019

are not probabilities. Rather, choices can be modeled as value-weighted samples from a *translation* of the value function into a probability distribution (Luce, 1959). The most obvious such translation of value into choice probabilities would simply assume that agents always choose the option with maximum value and never choose anything else (the *hardmax* function). For a variety of reasons, though, most models of value-guided decision-making instead make use of a softmax function, which exponentiates the value of all options and then normalizes these values into a probability distribution of choice probabilities that range from 0 and 1. In economics, this transformation typically reflects the experimenter's own uncertainty about the true option values that people are using to make their choices; people are assumed to be deterministically choosing the option with highest value (i.e., performing *hardmax*) but, because the experimenter can only noisily estimates people's option values, the softmax function captures this uncertainty under certain assumptions about the noise in the estimates (McFadden, 1973). In contrast, in reinforcement learning, individual agents are assumed to implement something like softmax over their option values for the purpose of exploring possible options that are not *currently* valued as best (e.g., a new restaurant that just opened across the street), but may turn out to be better than the option that is currently most valued (e.g., a classic favorite restaurant; Sutton & Barto, 1998). Whichever of these explanations is at work, it is clear that, when our task is decision-making, our decisions seem to be approximated by softmax-weighted samples of *value*.

For different tasks, then, people sample according to different weightings—they focus on probability for prediction tasks, but on value — converted to softmax choice probabilities — for decision-making. How might these task-specific sampling strategies influence what comes to mind in the absence of any explicit task?

One uninteresting possibility is that people uniformly default to one or the other task-specific approach—everybody samples according to probability, or everybody samples according to value. A related possibility is that people simply view their task as ambiguous, and everybody resolves the ambiguity by defaulting to a single task-specific approach, but different people default to different tasks. If so, the distribution of what comes to mind in a population will be some linear combination of the responses of people engaging in a purely statistical task of prediction and the responses of people engaging in a purely normative task of decision-making.

Here, we consider a different, more surprising proposal: What comes to mind is not a simple *mixture* of the task-specific sampling processes involved in prediction and decision-making, but is a distinctive *compromise* between the two. Specifically, we suggest that what comes to mind by default are things that are simultaneously probable *and* valuable (i.e., things that are proportional to the mathematical *product* of these two task-specific sampling distributions). In contrast, according to the uninteresting mixture view, an amount comes to mind simply when it is either valuable *or* probable (i.e., is an additive mixture of the two). If the compromise view is right, then what comes to mind by default is governed by a specific and unique weighting function of its own, and stands out as an interesting object of study in its own right.

Our proposal is inspired by a diverse body of recent empirical findings. Remarkably, across tasks that are superficially very dissimilar, people combine probability and value information into a *sui generis*, hybrid representation when engaged in many basic forms of reasoning (e.g., Bear & Knobe, 2017; Icard, Kominsky, & Knobe, 2017; Phillips & Cushman, 2017; Wysocki, 2018). For example, when participants are asked what amount of television is 'normal,' their answers are intermediate between purely probabilistic and value-based judgments (Bear & Knobe, 2017). Similar results have been found across numerous other domains (Wysocki, 2018) and using several other measures, such as causal judgment (Icard et al., 2017), modal reasoning (Phillips & Cushman, 2017), counterfactual reasoning (Kahneman & Miller, 1986), gradable adjectives (Egré & Cova, 2015), and concept prototypes (Bear & Knobe, 2017). These experiments are not amenable, however, to a precise quantitative characterization of the manner in which probability and value information are integrated.

In three experiments, we tested the hypothesis that what first comes to mind is a blend of statistical probability and prescriptive value. Initially, we explored this question qualitatively in a naturalistic setting (Experiment 1; cf. Bear & Knobe, 2017). We then designed two artificial settings to quantitatively model the sampling distribution of what comes to mind. One of these settings asked about the number of minutes that people engaged in a fictional hobby (Experiment 2), while the other asked people to visually imagine a fictional tool of a certain length (Experiment 3). Despite the different modalities in which these intuitions were probed, we found strong support for the hypothesis that people sample from a unique probability distribution that combines statistical and prescriptive information.

2. Experiment 1

In this experiment, we examined how people's intuitions about average and ideal amounts of various ordinary behaviors or events relate to what numbers spontaneously come to mind. We developed a list of 40 such behaviors or events (such as *amounts of TV watching per day*), 20 of which were borrowed from a similar design from Bear and Knobe (2017). We hypothesized that the numbers that first come to mind would be influenced not only by what was considered average, but also what was considered ideal.

2.1. Method

This study proceeded in two parts on Amazon's Mechanical Turk. One set of 100 participants was randomly assigned to judge either the average or ideal amount of a set of 20 randomly chosen behaviors or activities, which were randomly selected from the total set of 40. (This sample size, and the one reported below, was chosen on the basis of past work (Bear & Knobe, 2017) and past pilot data, which suggested that we would have sufficient power to detect an influence of average and ideal judgments at the item level.) These 20 items were presented in random order to participants. Thus, for 20 of the 40 domains, approximately 50 participants were asked to fill in responses like "Average number of hours of TV that a person watches in a day", and approximately 50 other participants were asked to fill in responses like "Ideal number of hours of TV for a person to watch in a day". To avoid demand characteristics, participants were always asked only about either averages or ideals, never both in the same session.

A separate group of 100 subjects participated in the sampling part of the experiment, in which they gave amounts that first came to mind. Participants were instructed to simply "enter the first number that comes to mind" when reading the presented phrase, and it was emphasized that there was no "correct" answer. In order to encourage participants to give a spontaneous judgment, we instructed them to try to give each response in under 5 s. However, responses were still solicited after this time delay. After completing two practice trials, the participants were presented with a random 20 out of 40 domains, presented in random order. Each page simply displayed a phrase like "NUMBER OF HOURS OF TV FOR A PERSON TO WATCH IN A DAY" and a timer counting down from 5 s, along with a box for subjects to give their response.

2.2. Results

Participants' responses in each condition were averaged for each of our 40 domains (Table 1). All responses from three participants who failed an attention check were excluded from further analysis. In addition, 89 item-level responses that were 3 standard deviations away from the mean answer for that item's dependent measure were eliminated.

Since our questions involved very different kinds of quantities

Table 1
Mean Average, Ideal, and Sample Judgments across Domains.

Domain	Average	Ideal	Sample	Domain	Average	Ideal	Sample
Hours TV/day	3.38	1.63	2.87	*Drinks frat bro consumes/wkend*	11.12	6.63	15.64
Sugary drinks/wk	9.17	2.41	5.91	*Times honk at drivers/wk*	2.67	0.72	2.53
Hours Exercise/wk	4.00	5.58	6.33	*Mins on social media/day*	60.57	35.40	59.10
Cals consumed/day	2225.91	1900.00	1859.24	*Times parent punishes child/month*	6.58	2.28	3.25
Servings fruits & veggies/month	40.00	94.96	39.16	*Miles walked/wk*	9.79	12.96	9.96
Lies told/wk	9.57	1.17	8.44	*% people drive drunk*	11.30	1.23	9.45
Mins late for appointment	14.22	3.04	13.60	*Times cheat on partner in life*	1.52	0.00	1.73
Books read/yr	7.22	17.40	8.45	*Times snooze alarm/day*	2.13	0.76	1.98
Romantic partners in life	6.09	5.77	8.06	*Parking tickets/yr*	1.67	0.04	1.37
Country's international conflicts/decade	11.67	1.36	4.15	*Times car wash/yr*	10.77	12.85	11.31
$ cheated on taxes	437.45	82.00	350.32	*Cups coffee/day*	2.21	1.84	2.72
% students cheat on HS exam	33.00	2.17	19.50	*Desserts/wk*	3.85	2.92	4.04
Times checking phone/day	28.57	7.68	16.57	*Loads of laundry/wk*	3.42	2.70	3.75
Mins waiting on phone for customer service	20.21	3.88	13.29	*% smokers*	22.81	6.16	20.79
Times called parents/month	5.00	5.50	7.04	*% HS students underage drink*	35.81	13.71	32.96
Times clean home/month	5.78	4.35	6.24	*% lie on dating website*	50.56	13.40	47.20
Times computer crash/wk	3.07	0.12	1.14	*Servings carbs/day*	62.43	16.13	33.23
% HS dropouts	10.67	1.29	11.49	*Txt msgs sent/day*	27.18	12.88	18.10
% middle schoolers bullied	17.59	0.81	19.46	*Times lose temper/wk*	2.60	0.56	2.20
Hrs slept/night	6.69	7.84	7.32	*Times swearing/day*	8.69	5.88	11.26

(hours, calories, etc.), assumptions of normality were violated. To address this problem, mean responses for each measure were converted to log scale (after adding 1 to each number, to avoid taking the log of 0).

To examine how judgments of averages and ideals affect sampling judgments, we compared a regression model in which only average judgments predict sampling judgments to a model in which both average and ideal judgments predict these judgments. The latter model reveals that both judged averages, $\beta = 0.77$, SE = 0.05, $p < .001$, and judged ideals, $\beta = 0.19$, SE = 0.05, $p = .001$, significantly predict sampling judgments. Moreover, the Akaike Information Criterion (AIC) for this model (3.03) is markedly lower than that for a model in which only judged averages predict normality judgments (14.08), suggesting that it is a more appropriate model of the observed data.[2]

We also conducted non-parametric analyses to explore whether the means of people's samples were intermediate between the mean of judged averages and ideals. For a given sample amount to be intermediate, it must be both on the ideal side of the average and the average side of the ideal. For the 40 domains, 29 were on the ideal side of average (binomial $p = .006$), and 37 were on the average side of ideal (binomial $p < .001$). Further, 26 out of 40 of the mean samples met both of these criteria — i.e., they were intermediate between average and ideal judgments. Thus, although many sample amounts were not intermediate, the proportion that were intermediate was considerably greater than what would be expected by chance (binomial $p < .001$ with a null hypothesis of 1/3, since there are two possible ways that an item can be non-intermediate).[3]

Although the format of this study makes it difficult to qualitatively assess the shape of the distributions of individual responses, the distributions of what comes to mind do not, at first appearance, seem to be a simple linear combination of two distributions centered around average and ideal judgments (Fig. 1S). We consider this question in more rigorous detail in the studies that follow.

3. Experiment 2

Experiment 1 provided some initial evidence that what first comes to mind is a blend of statistical and prescriptive information, but it was limited in several ways. For example, it is possible that we unintentionally selected real-world domains in which it happens to be the case that the most salient quantities that come to mind are a blend of what is common and what is desirable — even if this is not inevitable. In addition, our analysis showing that both estimated averages and judged ideals significantly predict what comes to mind could be explained by some unmeasured third variable that is correlated with the two predictors of interest.

To alleviate these and other concerns, we moved to a more controlled setting in Experiment 2, in which the entire distributions of statistical and prescriptive information that participants were exposed to were varied, so we could explore how these full distributions were functionally combined to produce samples that came to mind.

3.1. Method

One thousand and two hundred participants from Amazon's Mechanical Turk were randomly assigned into one of six conditions in a 2 × 3 pre-registered design. (Previous pilot data collection suggested that this would offer sufficient power to adequately distinguish our computational models.) We orthogonally manipulated the statistical distribution of quantities presented to participants (unimodal vs. bimodal) and the prescriptive goodness of those quantities (high ideal, low ideal, intermediate ideal).

Participants were first presented with a description of the fictional hobby of "flubbing". In the low ideal condition, participants were told that "although it is safe to flub for a few minutes every week, doctors warn that there are serious health risks associated with flubbing for longer periods of time." The high ideal condition, in contrast, stated that "doctors advise their patients to flub as much as possible" and that the more people flub, the healthier they are. The intermediate ideal condition stated that "doctors advise their patients to flub a moderate amount each week."

Participants were then told that they would be presented with

[2] As a robustness check, we also ran a regression where instead of taking the log of average, ideal, and sample means reported in Table 1, we *first* took the log of participants' individual responses (+1), and then computed the means of these log-transformed responses. In this new regression, judged averages, $\beta = 0.79$, SE = 0.04, $p < .001$, and judged ideals, $\beta = 0.17$, SE = 0.03, $p < .001$, continue to significantly predict what comes to mind. Moreover, the AIC is much lower for this model that includes ideal, compared to an average-only model (ΔAIC = 18.87).

[3] Given that several of these items have ideals that are essentially 0, which makes it unlikely that mean samples would be on the non-average side of ideal, we evaluate how many items are on the average side of ideal in a restricted set of 30 items whose ideals are not obviously at floor. Within this set, 27/30 items are on the average side of ideal (binomial $p < .001$), and 19/30 items have mean samples that are intermediate between ideal and average (binomial $p < .001$ for null of 1/3).

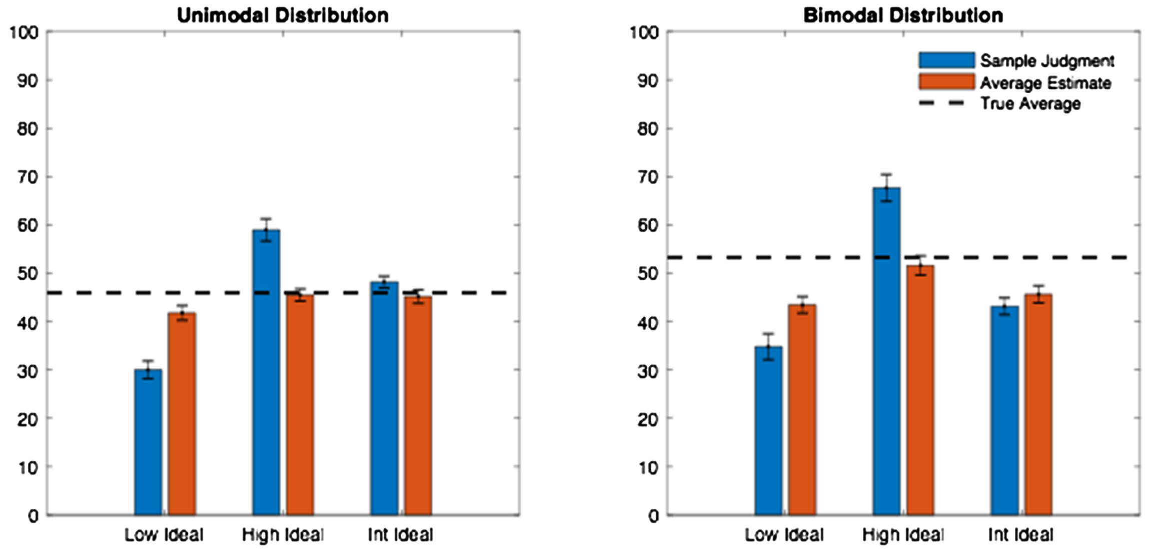

Fig. 1. Mean samples (blue) and estimates of average amount of flubbing (orange) for the unimodal (left) and bimodal (right) conditions from Experiment 2. Also shown are the true average amounts of flubbing presented (dashed black lines). Error bars are 95% CIs of the means. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

amounts of time (in minutes) that 100 people flubbed in the past week (one at a time, on separate pages), along with health grades, ranging from A+ to D−, that indicated the healthiness of flubbing for each of these amounts of time.

Before continuing, participants were also asked two comprehension questions: “Is it healthier for a person to flub for a long time or a moderate amount of time?”, and “Is it healthier for a person to flub for a short time or a moderate amount of time?”. Participants who answered either of these questions incorrectly were excluded from data analysis.

Grades were calculated in the following way. In the high ideal condition, all amounts of flubbing greater than 80 min were given an A+, and all amounts less than 20 were given a D−. The opposite was the case in the low ideal condition. Then, within the 20–80 range, grades were spaced linearly in intervals of 5, such that 75–80 corresponded to A+, 70–75 A, and so on for the high ideal condition, and the reverse for the low ideal condition. In the intermediate ideal condition, the most ideal amount of flubbing was set to 50, and quality of a given amount of flubbing x was scaled linearly based on the absolute distance from 50.[4] Because x was constrained to the 0–100 range, the grades in this condition were spaced in intervals of 4, rather than 5. Thus, x amounts within 4 of 50 were given an A+ grade, amounts less than 46 or greater than 54 were given an A grade, amounts less than 42 or greater than 58 were given an A− grade, and so on.

The amounts of flubbing were sampled from a normal distribution with $\mu = 45$ and $\sigma = 15$ in the unimodal condition and a sum of normal distributions with $\mu = 35$ and 75, and $\sigma = 5$, in the bimodal condition. (Note that the modes were intentionally selected to be slightly off-center from 50 so that we could observe interesting asymmetries that would be uniquely predicted by our models.) These numbers were rounded to the nearest integer. Within each of these conditions, all participants were given the exact same 100 numbers (i.e., we only sampled from these distributions once per condition), presented in a different random order for each participant.

After viewing all 100 amounts of flubbing, participants were asked, without forewarning, what was the first number of minutes of flubbing that came to mind. As in Experiment 1, they were told that there was no need to deliberate about this and that we were not looking for a particular correct answer. Participants were also asked afterwards what they thought the average amount was.

3.2. Computational framework

We consider several models that combine statistical and prescriptive information to produce a probability distribution of possible samples. Importantly, these models capture a population-level distribution of samples of what *first* comes to mind (what we measure in our experiments), rather than a distribution of samples that a single individual might make over time. In the models below, we use the notation $P(x)$ to refer to the statistical probability of observing x minutes of flubbing and $V(x)$ to refer to the prescriptive value of flubbing for that amount of time. In our experiments, $P(x)$ is specified by the generative unimodal and bimodal PDFs that we sampled from (see above), and $V(x)$ is computed linearly based on the grades that participants were presented. In the low- and high- ideal conditions, the grades varied over the range of 20–80 min (even though numbers could be presented from 0 to 100). Thus, there were 60 possible prescriptive values, with $V(x) = x - 20$ if $20 < x < 80$ in the high-ideal condition, and $V(x) = 80 - x$ if $20 < x < 80$ in the low-ideal condition. Values were then capped at 0 or 60 for all $x < 20$ or > 80. In the intermediate-ideal condition, $V(x) = 50 - |x - 50|$ for all x, such that $V(0) = 0$, $V(100) = 0$, and $V(50) = 50$.

The two simplest possible models assume that participants' samples are drawn proportional to either just the statistical probabilities $P(x)$ or just the prescriptive values $V(x)$ that we presented. We call these two models Statistics-Only (SO) and Value-Only (VO), where the probabilities[5] of sampling a given x on each of these models is defined as

$$SO(x;\,C) = P(x) + C$$

and

$$VO(x;\,C) = V(x) + C$$

where C is a constant that accounts for uniform non-relevant factors that contribute to sampling.

[4] To avoid overcomplicating the design for participants, we do not consider a bimodal ideal condition in which it best to flub either very low or very high amounts. However, it is an interesting direction for future research to consider how people sample from more complex value functions such as this.

[5] The equations presented here are not strictly probabilities or PDFs because they are not normalized. For simplicity, we present the unnormalized equations here.

Next, we consider two simple ways in which statistical and prescriptive information could be combined. First, as described in the Introduction, both of these factors might independently influence what quantities people are likely to sample, such that amounts that are more common are more likely to be sampled, and amounts that are more desirable are also more likely to be sampled, but the interaction between these two pieces of information does not play any role. In other words, people default to sampling from *either* a statistical distribution or a prescriptive distribution. According to this Additive model (Add), the probability of sampling is a weighted sum of statistical and normative factors:

$$Add(x;\, w,\, C) = wP(x) + (1 - w)V(x) + C.$$

Alternatively, samples may be proportional to the *product* of statistical probability and normative value, such that what comes to mind are amounts of flubbing that are both common and good. In other words, on this model, statistical and prescriptive information are not independent factors that contribute to what comes to mind, but are both necessary. We call this the Multiplicative (Mult) model:[6]

$$Mult(x;\, C) = P(x)V(x) + C.$$

Lastly, inspired by past work (Bear, Bensinger, Jara-Ettinger, & Knobe, 2018), we consider a variant of the multiplicative model, which exponentiates the prescriptive values. Given that prescriptive value is useful for choice, and choice probabilities are often modeled by taking the softmax transformation of prescriptive values (Luce, 1959; McFadden, 1973; Sutton & Barto, 1998), this is a sensible way of converting prescriptive values into choice probabilities. We call this the Softmax (SM) model:

$$SM(x;\, \tau,\, C) = P(x)e^{\frac{V(x)}{\tau}} + C$$

where τ is a softmax temperature parameter that modulates the influence of $V(x)$ relative to $P(x)$. (Note that, as in the other models above, the normalizing constant that is typically presented in the softmax equation is omitted here; see footnote 5.) We predicted that this model would best capture participants' responses.

3.3. Results

In total, 1197 participants completed the experiment (deviating slightly from our 1200 target), and 945 passed the comprehension questions. All reported analyses were performed on only those participants who passed comprehension.

We first compared participants' mean sample judgments to their mean estimates of the average amount of flubbing that they saw in the 100 numbers we presented. These means, for each of the two statistical distributions, are presented in Fig. 1. As shown, although participants did not perfectly estimate the true average amounts of flubbing in either distribution (dashed black lines), these judgments were significantly different from the mean sample judgments in the low ideal, paired $t(331) = 11.98$, $p < .001$, and high ideal, paired $t(293) = 16.55$, $p < .001$, conditions. As expected, sample judgments and average estimates did not significantly diverge in the intermediate ideal condition, $t(318) = 0.085$, $p = .93$. Thus, participants' samples were pulled towards prescriptively good amounts of flubbing.

Next, we fit each class of model described above to the data from the six conditions using maximum likelihood estimation. This was implemented with MATLAB's *fmincon* function. Parameters were estimated separately for each condition.

As predicted, the softmax model outperformed all other models in log likelihood, AIC, and BIC (see Supplement). Moreover, as shown in Fig. 2, this model successfully captured much of the structure of the distributions across all conditions, suggesting that it is a plausible computational account of what spontaneously comes to people's minds.

4. Experiment 3

Thus far, we have explored numerical quantities that come to mind. But if softmax sampling is a general feature of the way that the mind operates by default, similar principles should apply in an entirely different modality. In Experiment 3, we measure visual imagery that comes to mind after people have been presented with images of a fictional hunting tool in order to explore whether people sample in similar ways in a completely non-numerical context.

4.1. Method

One thousand and eight hundred participants from Amazon's Mechanical Turk were randomly assigned into one of six conditions in a 2 × 3 pre-registered design. (Given the additional task demands of manipulating a visual stimulus and the new exclusion criteria described below, we increased our sample size from Experiment 2. Pilot data suggested that this would offer enough power to distinguish our computational models.) We orthogonally manipulated the statistical distribution of quantities presented to participants (unimodal vs. bimodal) and the prescriptive values of those quantities (high ideal, low ideal, intermediate ideal).

Participants were first presented with a description of a fictional hunting tool called a "stagnar" (Bear & Knobe, 2017). In the low ideal condition, participants were told that "shorter stagnars are better than longer ones [for hunting] because they are easier to handle." The high ideal condition, in contrast, stated that "longer stagnars are better… because they are able to inflict the most damage." The intermediate ideal condition stated that "stagnars of moderate length are better than long or short ones because they are both small enough to handle and large enough to inflict serious damage."

Participants were then told that they would be presented with pictures of 100 different stagnars of varying lengths (one at a time, on separate pages), along with grades, ranging from A+ to D−, that indicated how good the stagnars were for hunting.

Stagnars were presented on the screen with a width of $300 + 400x$ pixels, where x corresponded to the minutes of flubbing used in Experiment 2. (The heights of the stagnar images were scaled proportionally to the original image.) Grades were also the exact same as those used in Experiment 2.

After reading the instructions, participants were asked a comprehension question: "The best stagnars are of what length?", which could be answered with "Short Length," "Moderate Length," or "Long Length." They were also asked to complete two screen calibrations, where they used a slider to adjust the length of a stagnar image to match the length of another image on the screen, which was either 400px or 600px wide.

After seeing the 100 stimuli, participants were asked to "imagine the first stagnar that comes to mind." We emphasized that they did not need to deliberate about this and that we were not looking for a specific "correct answer." Once the participants felt they had this image in mind, they advanced to a page where they could adjust an image of a stagnar with a slider to the image they were imagining. This slider had no markings on it, and could be dragged to create any length between 300px and 700px (corresponding to the limits of 0 and 100 min in Experiment 2) in increments of 4px. Afterwards, participants were asked to use the same slider to create the average length they saw from the 100 stimuli.

[6] In our online code available at OSF (see Preregistration and Code Availability), we also explore a more flexible version of this model, which we call *MultLS*. This more flexible model allows $V(x)$ to be linearly transformed, i.e., $V'(x) = aV(x) + b$, where a and b are free parameters, and $V'(x) \geq 0$ for all x. Model fits for this more flexible model end up being virtually identical to those for the simpler model above. For further information about the results from this model, see Supplementary Materials.

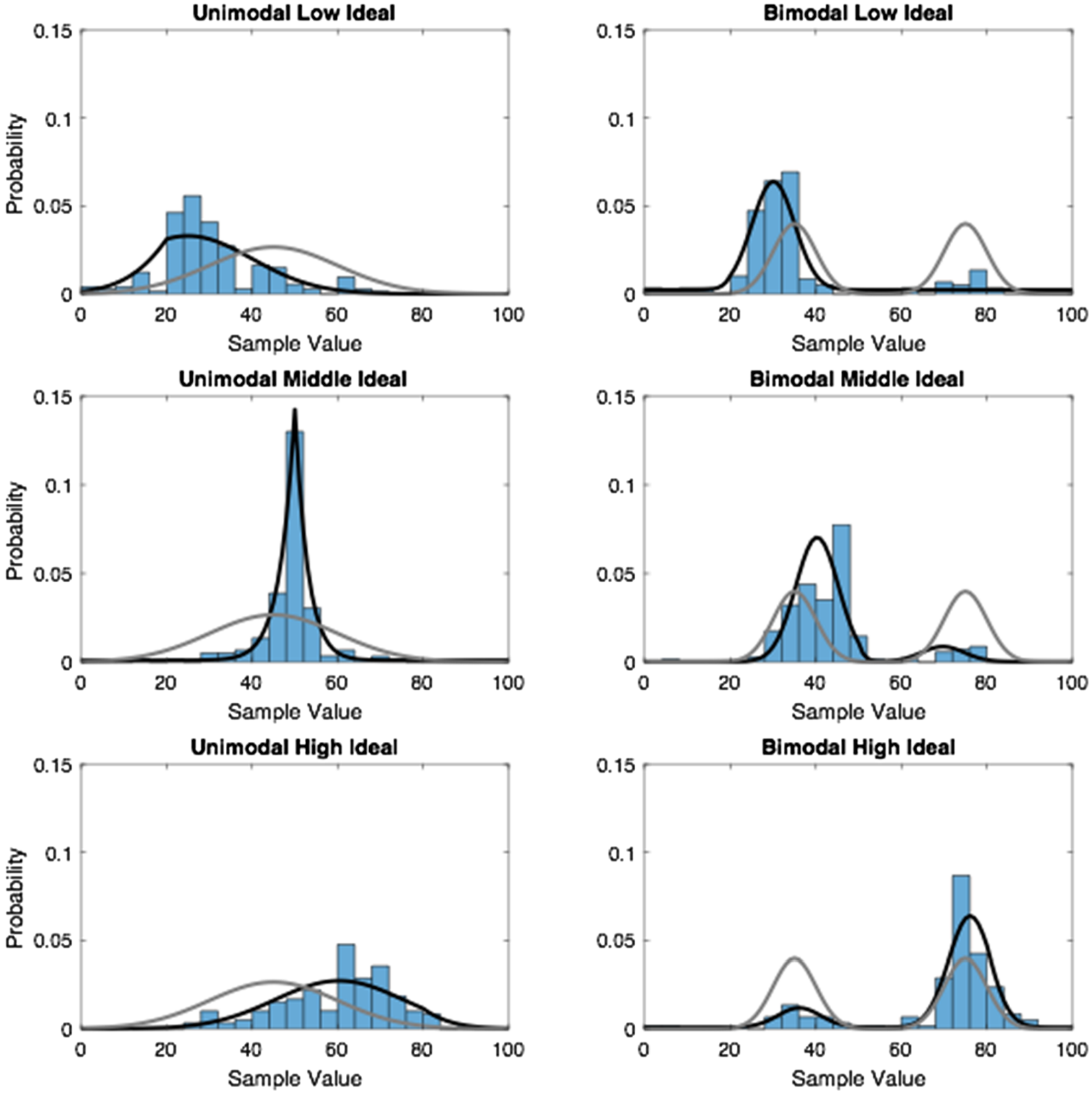

Fig. 2. Distributions and model fits for sample amounts of minutes from Experiment 2. Vertical bars show proportion of amounts sampled by participants, and black lines show softmax models with best fitting parameters for each condition. Also shown in gray are the generative statistical distributions (unimodal or bimodal) of amounts of flubbing.

4.2. Results

People who failed the comprehension question or were off by 20px in either direction for either screen calibration were excluded from data analysis. Participants who indicated that they had trouble with the slider or used a mobile device for the study were also excluded. In the end, 1129 participants passed these stringent requirements for analysis.

As in Experiment 2, we first compared participants' mean sample judgments (the stagnar length that came to mind) to their mean estimates of the average stagnar length that they saw in the 100 images we presented. These means, for each of the two statistical distributions, are presented in Fig. 3. As shown, although participants did not perfectly estimate the true average lengths in either distribution (dashed black lines), these judgments were significantly different from the mean sample judgments in the low ideal, paired $t(347) = 10.91$, $p < .001$, and high ideal, paired $t(390) = 20.77$, $p < .001$, conditions. As expected, sample judgments and average estimates did not significantly diverge in the intermediate ideal condition, $t(389) = 1.15$, $p = .250$.

Next, as in Experiment 2, we fit each class of model described above to the data from the six conditions using maximum likelihood estimation. This was implemented with MATLAB's *fmincon* function in the same way as in Experiment 2. Parameters were estimated separately for each condition.

As predicted, the softmax model outperformed all other models in log likelihood, AIC, and BIC (see Supplement). Moreover, as shown in Fig. 4, this model successfully captured much of the structure of the distributions across all conditions, suggesting that it is a plausible computational account of what spontaneously comes to people's minds.

5. General discussion

Our findings suggest that what first comes to mind, in the absence of an explicit goal, is a compromise between what is statistically probable and what is valuable. Experiment 1 demonstrated this across a number of real-world domains, while Experiments 2 and 3 provided quantitative support that people sample thoughts to mind from a probability distribution that combines information from these two dimensions of frequency and value.

These experiments rule out the hypothesis that what comes to mind by default is simply a mixture of what would come to mind during two

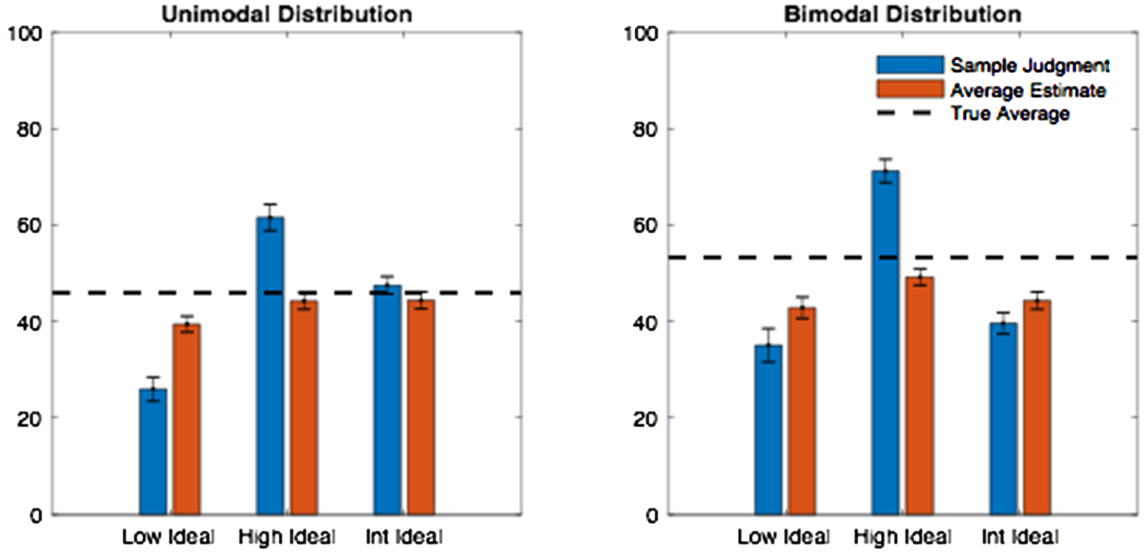

Fig. 3. Mean samples (blue) and estimates of average stagnar length (orange) for the unimodal (left) and bimodal (right) conditions from Experiment 3. (Note that all lengths are scaled to the 0–100 range, where 0 corresponds to 300px and 100 corresponds to 700px.) Also shown are the true average lengths of stagnars presented (dashed black lines). Error bars are 95% CIs of the means. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

important kinds of tasks: prediction and decision-making. At the same time, the default probability distribution that people do appear to sample from is closely related to those distributions involved in prediction and choice. Specifically, the default probability distribution in our best fitting model is the mathematical *product* of a distribution that could be used for prediction, $P(x)$, and a distribution that could be used for choice, softmax of $V(x)$.

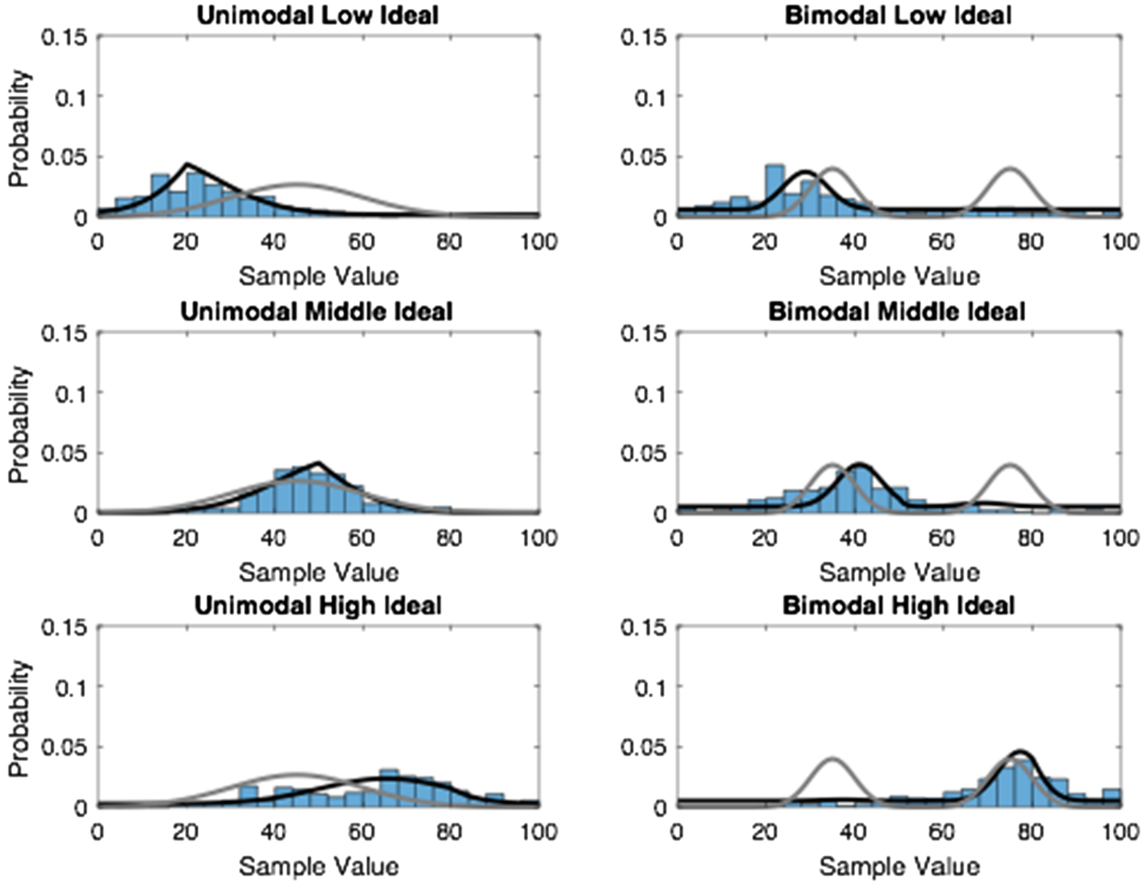

Fig. 4. Distributions and model fits for sample stagnar lengths from Experiment 4. (Note that all lengths are scaled to the 0–100 range, where 0 corresponds to 300px and 100 corresponds to 700px.) Vertical bars show proportion of amounts sampled by participants, and black lines show softmax models with best fitting parameters for each condition. Also shown in gray are the generative statistical distributions (unimodal or bimodal) of lengths presented. For comparison, we use the same y-axis as Fig. 2.

5.1. Conceptions of value

The present studies suggest that the probability of being sampled is impacted in some way by representations of prescriptive value. However, existing research has distinguished a number of different conceptions of value, and one might ask which of these conceptions is actually at work in these phenomena.

To begin with, people can represent either an object's expected value (i.e., a probability-weighted sum of the goodness values of many possible situations) or its value in and of itself. These might then come apart in some cases. For example, fancy food tends to be expensive and might therefore be represented as having low expected value, but all the same, it might be represented as having high value in and of itself.

Second, people can represent either the degree to which an object is valuable in general or the degree to which it is valuable given the goals they have right at the moment. For example, a person might think that steak dinners are generally very good, but if she happens not to be hungry, she might not value having one right now.

Finally, as Barsalou (1985) has emphasized, people often represent categories in terms of some distinctive ideal. People associate knives with the ideal of cutting well, teachers with the ideal of teaching well, and so forth. Here again, the result is that we may sometimes arrive at two distinct notions of value. Thus, in thinking about burglars, people might feel that there is some straightforward sense in which the best sort of burglar is one who hardly ever steals anything, but people might also have a Barsalou-style representation according to which there is some sense in which a ‘good burglar’ is one who burgles especially well.

Future research could explore the question as to which of these many different conceptions of value is actually impacting the probability that a given possibility will be sampled.

5.2. Relation to previous work on sampling

Within existing research, there has been a great deal of important work on the ways in which people use sampling to accomplish one or another specific task, including decision-making, judgment, prediction, and a variety of others (e.g., Fiedler, 2000; Frydman & Lawrence, 2019; Stewart, Chater, & Brown, 2006). Normative models have also been developed to explain how agents *should* sample in various contexts (e.g., Callaway & Griffiths, 2019; Icard, Cushman, & Knobe, 2018; Lieder et al., 2018). A question arises as to how the methods of sampling explored in this past work might relate to the kind of default sampling that we have explored here.

While past work has focused on task-specific sampling, we consider what comes to mind in the *absence* of any explicit task. Thus, it is possible that the kind of sampling people do without an explicit goal is simply distinct from other types of sampling that have been considered in the literature. Nevertheless, it is also possible that the default sampling distribution we have uncovered might subtly influence other types of sampling. For example, perhaps even when people are trying to sample options only proportional to their reward value in order to make the best decision possible, they are biased to think about options that are statistically probable (even if these options are not desirable). Or, conversely, perhaps when people are trying to make a merely probabilistic judgment or prediction, their samples are swayed by prescriptive value. We hope to explore these connections between what comes to mind by default and what comes to mind in task-specific settings in follow-up work.

While distinct in many ways, two lines of past research seem particularly noteworthy in relation to the present findings. First, Lieder et al. (2018) have developed a model of *utility-weighted sampling*, inspired by the mathematical technique of *importance sampling*. According to their model, when people consider possible future outcomes, they oversample outcomes that would have a large impact on their utility—either positive or negative—relative to equally probable outcomes that have a smaller impact on utility. Utility-weighted sampling is advantageous when one must estimate the expected value of an event by sampling its consequences. This specific context was not, however, the focus of our experiments. Rather, we queried the form of sampling that occurs in the absence of any particular task. And, in this context, we find clear evidence for an oversampling of high values, but no corresponding evidence for an overweighting of low values. This highlights two important questions for further study. First, what are the contexts in which people oversample both high and low values (consistent with utility-weighted sampling), and what are the contexts in which they selectively oversample high values (consistent with our findings)? Second, what are the distinct functions that favor one or the other approach in different contexts? As we have emphasized, our principle contribution is to demonstrate the striking structure endemic to “default” sampling, but considerable uncertainty about its ultimate function remains.

A second line of research has explored the role of sampling in *memory recall*. Given that we can only recall a limited number of experiences at a time, what factors should determine the specific memories that come to mind? In the context of spatial navigation, Mattar and Daw (2018) show that an optimal agent will recall past experiences that are high on two dimensions — *gain* and *need*. *Gain* roughly corresponds to the *value* that the memory would provide the agent for acquiring future rewards in the situation that is recalled, while *need* represents the *probability* that the agent will encounter this remembered situation in the future. Similar to the present results, the authors find that an optimal agent should recall memories that maximize the *product* of these two factors. That is, the most useful experiences to call to mind are those that are both instructive for generating reward and likely to be encountered in the future. In contrast, it is not useful to ruminate on past situations that are highly instructive, but unlikely to be encountered in the future, nor is it useful to recall past experiences that are likely to be encountered in the future, but have already been optimized for reward.

In the context of our tasks, it is unlikely that participants are literally sampling from episodic memories of past events. We find this unlikely because, for instance, the specific amounts of “flubbing” that come to mind in Experiment 2 were often quantities that were never presented during training. Nevertheless, it is intriguing that the basic motif of a product of statistical and value information arises in both these contexts. This may indicate a common functional design, even if it does not indicate a common mental or neural mechanism.

5.3. Default sampling and adaptiveness

Why might the brain sample by default from a distribution shaped by both probabilistic and prescriptive considerations? One obvious hypothesis would be that there is something adaptive about this approach to default sampling. We briefly consider two possible explanations along these lines.

First, one obvious potential function of sampling from a combination of these two distributions by default is that what comes to mind could aid in *both* a future prediction and a future decision. For example, two hours of television watching in a day might be a reasonable expectation of what another person might do and also be a reasonable target to watch for oneself. This amount could therefore serve as a useful anchor point for either of these two very different goals, as they might arise. By analogy, if you are unsure whether your friend prefers to eat meat or fish on any given night, you might default to choosing “surf & turf” for them, which would be a highly-valued option in either event. Biasing towards “compromise” dishes of this kind may be more adaptive than sampling from a mixture of pure-meat and pure-fish dishes.

A different kind of account might not explain the distributions we observe in terms of a compromise, but instead provide a more unified explanation for the adaptiveness of combining both of these features into a hybrid probability distribution. To see how this might work, we offer a specific, illustrative example of the general kind. Specifically, it is possible that the default representation we identified is actually quite tailored to the task of predicting the outcome of a choice process across variable environments. Often, we cannot consider every possible choice we could make in a given context, but rather must choose among a set of feasible options (Phillips, Morris, & Cushman, in press). For instance, for a variety of reasons (work, family, roommates, etc.) we may not be able to watch our preferred amount of TV every day. Thus, the feasible options on any given day will be a subset of the full range of 0–24 hours. In this case the set of feasible options is drawn from a *statistical probability distribution*, while choice among the set of feasible options is dictated by a *prescriptive value function*, converted to a choice probability distribution via softmax. The actual amount of TV that we are likely to choose, then, will be dictated by the product of these two distributions. Put simply, what we are likely to end up with on any given day is the product of the specific options we are likely to have (governed by descriptive probabilities) and our preference among that set of options (governed by value). What comes to mind by default, then, may be drawn from the distribution of things that are likely to be obtained, given both extrinsic feasibility constraints and personal preferences. This distribution could function as a useful baseline approximation of the average quality of goods or events that are obtainable in one's environment.

6. Conclusion

Although the ultimate functional explanation for people's default sampling tendencies remains uncertain, the present work offers a first step in describing the informational and computational factors that contribute to a largely unexplored psychological phenomenon: thoughts entering conscious awareness by default. Our results suggest that these thoughts are not simply a recapitulation of thoughts that would arise in other contexts, but are an interesting topic of study in their own right. Further work should continue to explore the nature of this blending and its role in downstream cognitive processes.

7. Preregistration and code availability

Experiment 1 was not formally preregistered; the preregistration for Experiment 2 can be accessed at http://aspredicted.org/blind.php?x=3bc27u, and the preregistration for Experiment 3 can be accessed at http://aspredicted.org/blind.php?x=ie74cf. De-identified data for all experiments along with a code-book and the data analysis scripts are posted on OSF at https://osf.io/2dzmf/?view_only=38c4c755930d4d89bceddc218fed21ad. The materials used in these studies may be requested at any time.

Acknowledgements

This research was supported by grant N00014-19-1-2025 from the Office of Naval Research and grant 61061 from the John Templeton Foundation.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2019.104057.

References

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 629–654.

Bear, A., & Knobe, J. (2017). Normality: Part descriptive, part prescriptive. *Cognition, 167*, 25–37.

Bear, A., Bensinger, S., Jara-Ettinger, J., & Knobe, J. (2018). *What comes to mind? A mix of what's likely and what's good. Proceedings of the fortieth annual conference of the cognitive science society*.

Callaway, F., & Griffiths, T. (2019). Attention in value-based choice as optimal sequential sampling [March 4]. https://doi.org/10.31234/osf.io/57v6k.

Egré, P., & Cova, F. (2015). Moral asymmetries and the semantics of many. *Semantics and Pragmatics, 8*, 1–45.

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review, 107*, 659–676.

Frydman, C., & Lawrence J. (2019) Efficient coding and risky choice. Unpublished manuscript.

Gershman, S. J., Vul, E., & Tenenbaum, J. B. (2012). Multistability and perceptual inference. *Neural Computation, 24*, 1–24.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science, 17*, 767–773.

Icard, T. (2016). Subjective probability as sampling propensity. *Review of Philosophy and Psychology, 7*, 863–903.

Icard, T., Cushman, F., & Knobe, J. (2018). *On the instrumental value of hypothetical and counterfactual thought. Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition, 161*, 80–93.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 93*, 136.

Lieder, F., Griffiths, T. L., & Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological Review, 125*, 1–32.

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

Mattar, M. G., & Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nature Neuroscience, 21*, 1609.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.). *Frontiers in econometrics*. New York: Academic Press.

Phillips, J., & Cushman, F. (2017). Morality constrains the default representation of what is possible. *Proceedings of the National Academy of Sciences, 114*, 4649–4654.

Phillips, J., Morris, A., & Cushman, F. (in press). How we know what not to think. *Trends in Cognitive Sciences*.

Stewart, N., Chater, N., & Brown, G. D. (2006). Decision by sampling. *Cognitive Psychology, 53*, 1–26.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: an introduction.* Cambridge, MA: MIT Press.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science, 19*, 645–647.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science, 38*, 599–637.

Wysocki, T. (2018). Normality: A two-faced concept. Unpublished manuscript.