Decoding silence in free recall

Francesco Fumarola [1,*] Zhengqi He,[1] Łukasz Kuśmierz,[1,2] and Taro Toyoizumi [1,3,†]

[1]*Laboratory for Neural Computation and Adaptation, RIKEN Center for Brain Science, Wako, Saitama 351-0198, Japan*

[2]*Department of Neurobiology, Duke University, Durham, North Carolina 27710, USA*

[3]*Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo, Tokyo 113-8656, Japan*

(Received 12 January 2022; accepted 16 May 2022; published 1 August 2022)

This article is part of the Physical Review Research *collection titled* Physics of Neuroscience.

In experiments on free recall from lists of items, not all memory retrievals are necessarily reported. Previous studies investigated unreported retrievals by attempting to induce their externalization. We show that, without any intervention, their statistics may be directly estimated through a model-free analysis of inter-response times—the silent intervals between recalls. A delay attributable to unreported recalls emerges in three situations: if the final item was already recalled ("silent recency effect"); if the item that, within the list, follows the latest recalled item was already recalled ("silent contiguity effect"); and in sequential recalls within high-performing trials ("sequential slowdown"). We endeavor to model all these effects through a stochastic process where the discarding of recalled items without reporting ("bouncing") occurs either if they are repetitious or, in strategically organized trials, if they are not sequential. Based on our findings, we propose various approaches to further probing the submerged dynamics of memory retrieval. This article is part of the Physical Review Research collection titled Physics of Neuroscience.

DOI: 10.1103/PhysRevResearch.4.033089

I. INTRODUCTION

Human memory recall has come to be increasingly regarded as a multilayered process, driven by the interplay of free association and higher cognitive operations [1–5]. Various neuroimaging tools are being deployed to study what has been termed *postretrieval* mechanisms, a set of mechanisms aimed at monitoring and evaluating retrieved memories [6–10]. Evidence points toward a wide-ranging involvement in the top-down control of memory recall by prefrontal [11,12], frontal [13], and parietal [14] areas. Ventrolateral prefrontal cortex (VLPFC), in particular, has been argued to implement a postretrieval selection process [10] that is stronger if task-irrelevant representations are dominant; electrodes implanted in VLPFC [5] were observed to increase their activity a few hundred milliseconds after the onset of the recall stage, consistent with a top-down postretrieval feedback. The general breadth and organization of postretrieval processes, however, remain unknown.

An ideal setting for addressing the question is the experimental paradigm known as *free recall* [15–19], which has provided insight into all facets of human episodic memory for over a century [20,21]. Free-recall experiments include a presentation stage and a recall stage; the former consists in the presentation of a list of items (often words), and the latter consists in the recall of those items (usually reported verbally by subjects) with no constraint on the ordering. The aim is to maximize performance, defined as the number of distinct items recalled correctly. (For a comprehensive description of both experiments and theories, see Ref. [22].)

On the basis of existent modeling approaches, results on free recall that have accumulated in the literature may be subsumed under two classes: results that have been conventionally modeled without invoking postretrieval mechanisms; and results that are, more or less implicitly, understood to originate in postretrieval feedback.

Within the former group, the most studied effects concern the *serial position* of recalled items, i.e., their order within the list presented. The *recency* and *primacy* effects [16] are the preferential tendency to recall items from the beginning and the end of the list, respectively. The difference in serial position between two consecutively recalled items is called "serial-position lag," and the *lag-recency* or *contiguity* effect [23] is a bias toward small lags, i.e., the tendency to recall contiguously items that are contiguous within the list [24,25]. The additional tendency to recall in forward order ("forward asymmetry") makes lag $L = +1$ the most frequent transition, which we will refer to as "sequential" [26]. All these results have been conventionally modeled without assuming any postretrieval mechanism [22].

In parallel, there is a body of work showing evidence for postretrieval mechanisms such as, notably, repetition avoidance. Repetition avoidance is indirectly evidenced by various experimental facts. For example, in serial recall, where items must be recalled in their order, fewer intrusions (words

*francesco.fumarola@riken.jp

†taro.toyoizumi@riken.jp

reported erroneously but not belonging to the lists) were found to correlate with better recall of the final item [27]; this was attributed to the fact that recall of the final item is boosted by repetition avoidance and becomes harder if fewer items are blocked as repetitious. More detailed descriptions of repetition avoidance were worked out with other models [28,29]. Since it is difficult to examine all possible models, it would be desirable to develop model-free approaches to extract information on unreported events from ordinary free-recall data.

Another topic that has been the focus of a large amount of attention is retrieval properties (such as recall strategies) not at the level of individual recalls but observed consistently within individual trials. This includes chunking, i.e., retrieving consecutive items in one group, and serializing, i.e., retrieving items in their serial order [1,4,30,31]. Such strategies are mostly consistent within a trial and considered to be a characteristic of trials rather than recalls. They were shown to be effective in boosting performance but are implemented by a minority of subjects and even by those only in a fraction of trials [30]. In fact, they are often developed and refined over the length of an experimental session as each subject learns from his performance in previous trials [32]. While some recall strategies are related to chunking, it has long since been known that a particularly advantageous strategy consists in waiving the *freedom* perk of free recall by serializing [30,31,33]. Sticking to such strategies should involve discarding any retrieval that does not fit the prescribed recall sequence.

One approach to unraveling the histories of unreported recalls consists in directly demanding that subjects externalize them. This type of interventionism was attempted as early as with the "unedited recall" experiments of Ref. [34], where subjects were instructed to express all items that came across their mind. Such experiments suggest that there are more items recalled than reported; moreover, they have provided further support to the idea that termination is predominantly triggered by repetition [35]. Although such an experimental procedure was shown to yield results coherent with the "edited" (i.e., regular) version of free recall [36] and was consequently proposed as a method to explore age dependence [37], memory capacity [38], and intrusions [29], self-reporting obviously involves physical bounds on the time resolution and, potentially, biases created by the additional instruction [39]. Whether or not such data offer an exact representation of thought processes taking place during ordinary free recall is also a question.

Here, we sidestep both the caveats of the externalization paradigm and, in our opening foray, the problem of selecting a model. We focus instead on the model-free analysis of one specific behavioral observable—the time that elapses between consecutive recalls, also known as the inter-recall interval or IRI [40–43].

During the recall stage of a typical free-recall trial, the first few recalled items are reported rather quickly; in contrast, towards the end of the trial the IRIs drastically increase [44,45]. Many features of this growth were shown to be reproduced by a pure-death process, in which items are sampled with replacement from a pool of listed items [45–47]. Crucially, in that model it is assumed that only items that have not been reported so far are admissible, and thus, when the other items

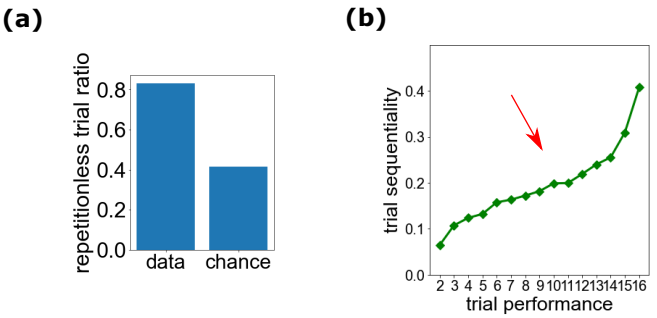

FIG. 1. Two features of free recall. (a) Paucity of observed output repetitions as compared with chance level (for details of chance level estimation, see Fig. S2). (b) Fraction of sequential recalls in a trial as a function of trial performance (i.e., number of recalled items). The mean curve is shown, while the standard error of the mean is too small to plot. The red arrow highlights the presence of an inflection point, where the slope starts to increase.

are drawn, they are simply discarded. Because towards the end of a trial there are many previously recalled (hence nonadmissible) items, many samples have to be drawn on average, and thus IRIs increase. Although the pure-death model is very simplistic, its structure suggests that a form of postretrieval editing takes place during the recall phase and that only some of the sampled (retrieved) items are reported.

We will argue that a careful but straightforward analysis of both the history of reported recalls and the recorded IRIs can allow us to uncover hidden retrievals that are suppressed by the participants. We will then buttress our model-free analysis of data by means of a toy model chosen for its simplicity and capability to explain simultaneously different types of unreported recalls.

II. TWO FREE-RECALL PHENOMENA

The data we considered, collected by the Computational Memory Laboratory at the University of Pennsylvania, concern experiments performed with lists of $N = 16$ words (see Appendix A). The data set displayed standard serial-position effects (with recency and primacy emerging as in Fig. S1 of the Supplemental Material [48]).

We begin by pointing out two general facts in the data.

(a) Although repetitions are not expressly forbidden and are duly recorded by the experimenter, the number of repetition-free trials is remarkably large. This statement can be made precise by adopting as chance level a basic Markov model of transitions among reported recalls [49] equipped with a sink state for termination (calculation of the chance level is fully detailed in the caption of Fig. S2). We thus find that the number of repetition-free trials is about twice the chance level [Fig. 1(a)]. This fact agrees with previous inquiries (see Introduction) and indicates that subjects have a spontaneous tendency toward not reporting twice the same recall (see Fig. S2 for a more detailed comparison of the repetition statistics to chance level in the data).

(b) We define "trial sequentiality" as the fraction of sequential recalls in the trial. Trial sequentiality is a peculiar function of trial performance; as shown in Fig. 1(b), not only does

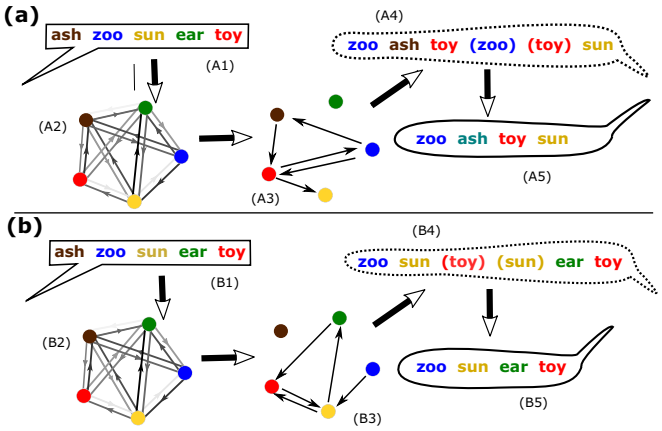

FIG. 2. Schematic depiction of bouncing. Intervention over the free-association process by "bouncing" simultaneously explains paucity of observed repetitions and higher sequentiality of high-performance trials. Here, a five-item list is presented in two separate trials leading to an instance of *repetition bouncing* (a) or one of *strategic bouncing* (b). During the presentation stage of free recall [(a1) and (b1)] a subject is presented visually or orally with a list of L items (here, $L = 5$ words). Memories of exposure to each of the items are stored in such a way that mental transitions among memories have unequal probabilities represented in (a2) and (b2) by the shades of gray of the arrows. During the recall stage (a3), the subject instantiates a mental trajectory through those memories [(a3) and (b3)]. Nothing prevents the sequence of retrieved memories from containing an unlimited number of repetitions [as, for example, in (a3) and (a4)], but the output is devoid of repetitions (a5), posing the problem of whether some retrievals are discarded. In addition, a tendency toward sequentiality in high-performing trials is also observed (b5), posing the problem of whether nonsequential transitions have been actively avoided [as when stepping back from "toy" to "sun" in (b3) and (b4)].

higher trial sequentiality correspond to higher performance but also this increasing curve appears to accelerate at an inflection point midperformance.

If one reaches for the simplest explanations by postretrieval processes, those two seemingly unrelated facts appear to share some key features.

(i) In both phenomena, the extra mechanism at work with free association can be described as the avoidance of options in principle provided by free association—repetitious retrievals [fact (a)] and nonsequential retrievals [fact (b)].

(ii) While the avoidance has different motives in the two cases and could be implemented in different brain areas, it may be assumed to operate in a *de facto* similar way, with each free association move being allowed or disallowed by a dedicated module depending on whether the retrieval meets a specific criterion (novelty or sequentiality).

These empirical observations lead to our working hypothesis, illustrated in Fig. 2. To denote the active avoidance of free-association recalls, we use hereinafter the word "bouncing." We distinguish two forms of bouncing: firstly, the avoidance of recalls that would yield repetition (cf. Introduction), which we call "repetition bouncing", and, secondly, the avoidance of recalls that would not obey a sequential ordering. We refer to this latter type as "strategic" bouncing because it is known that a bias toward sequentiality boosts performance [33] and, as will be seen below, it is observed selectively in high-performance trials.

In the rest of this paper, we report on our testing of the bouncing hypothesis from two complementary fronts—data mining and mathematical modeling. On the side of data mining, we exploit the large size of the data set to tease out information not explicitly reported by individual subjects, with the aim of uncovering what has not been verbally outputted. On the side of modeling, we articulate the above intuitions as a streamlined theory that represents free recall without bouncing as a Markov process and bouncing as a non-Markovian add-on, and compare predictions of the theory with observed features of the data.

III. MODEL-FREE ANALYSIS OF RECALL DELAYS

If a mechanism exists by which the free-association process is corrected, unwinded, or rebooted, given that this prevents the outputting of some retrieved items, it will not leave direct traces within the record of recalled items. Yet, it may produce temporal delays in recall transitions where the process is activated. In Fig. 2(a), for example, the word "sun" will be recalled more slowly because first the word "zoo" was retrieved and discarded. A natural approach is thus to hunt for traces of those delays in the recorded IRIs. To do so, the first obligatory step is an *a priori* analysis of the main potential confounders.

Following convention [22], we will call the ordered position of an item within the sequence of recalls reported during the recall stage the "output position." It can be a positive number if counted from the beginning of the recall process or a negative number if counted from the end (number of extant recalls to the end of the trial). For example, in the trial of Fig. 2(a), the word "ash" has output position $= +2$ or $= -3$. If trials with different performance are included, there is no fixed correspondence between positive and negative counting of outputs. It was shown in Ref. [45] that the distribution of IRIs is characterized by negative output position. This is indeed apparent in the data set under consideration, where dependence on the negative output position alone explains about 20% of the variance (Fig. S3). Note that the time interval from the onset of the recall stage to the first recall at output position $= +1$ is not an IRI but will be included for convenience among the IRIs and this does not qualitatively change our results.

Upon subgrouping all recall events by their negative output position, we compare events occurring before and after the item in a given serial position has been recalled [Fig. 3(a)]. After recall of an item, its repetitious retrieval becomes possible. If such a retrieval comes to mind, it would be associated with a positive time delay. In other words, systematic *bouncing* against the item in a given serial position will be expected to make the postrecall IRI larger than the prerecall IRI for every output position (see Appendix B for details of the analyses). This turns out to be true for the last item in the list regardless of the output position [Figs. 3(a) and 3(b)]. If the final item has already been recalled, all else being equal, a delay is found to occur in the recall process.

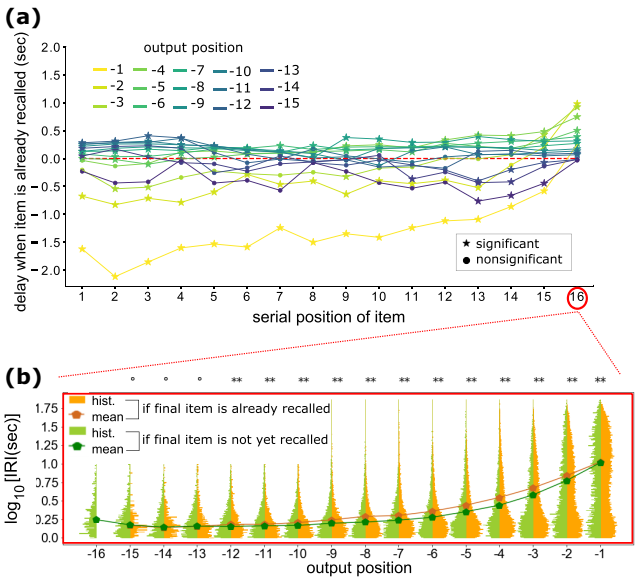

FIG. 3. Empirical evidence for a higher-order "silent recency effect" in free recall. (a) Difference between mean IRIs when subsampling on the basis of whether the item with a given serial position on the list has already been recalled or not. A histogram was made of the IRIs excluding recalls of the item with the given serial position also when it is not yet recalled. The serial position concerned is shown on the x axis, and plot markers encode the significance as defined by a Mann-Whitney U-test P value lower than 10^{-3} (star, significant; circle, nonsignificant). (b) Histogram (hist.) of the IRIs for recalls of items other than the final one, plotted separately for each output position and for the two conditions where the final item has already been recalled or not. The scale of the IRIs is logarithmic. The P value from the Mann-Whitney U test is encoded as follows: circle, $P > P^* = 10^{-3}$; one star, $P^* = 10^{-3} > P > 10^{-4}$; two stars, $10^{-4} > P$. A previously occurring recall of the final item is statistically associated with a delay in the recall of other items. Subjects are seen to rush for the final output because of the imposed time-out.

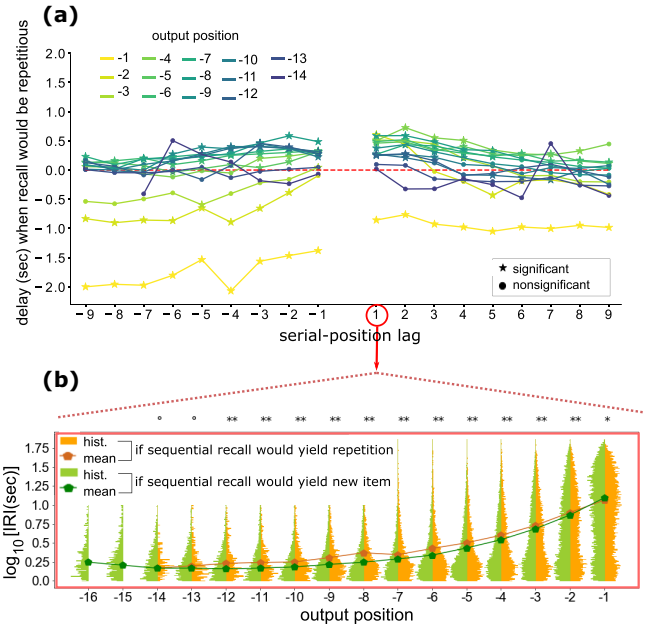

FIG. 4. Empirical evidence for a "silent contiguity effect." (a) Difference between mean IRIs when subsampling on the basis of whether the item situated at a certain fixed lag from the last recall has already been recalled or not. The lag position concerned is shown on the x axis, and plot markers encode significance as defined by a Mann-Whitney U-test P value lower than 10^{-3} (star, significant; circle, nonsignificant). (b) Histogram of the IRIs for recalls of items other than the sequential one (i.e., the one that was presented right after the latest recalled item) shown separately for each output position and for the two conditions where the sequential item has already been recalled or not. The scale of the IRIs is logarithmic. The P value from the Mann-Whitney U test is encoded as follows: circle, $P > P^* = 10^{-3}$; one star, $P^* > P > 10^{-4}$; two stars, $10^{-4} > P$. A previously occurring recall of the sequential item is statistically associated with a delay in the recall of other items.

The fact that the final item appears to be a preferential bouncing target is a nontrivial manifestation of the recency effect (cf. Introduction). The delay associated with potential repetition of the final item can thus be thought of as a "silent recency effect"; the quantitative strength of memory depends on serial position, and given that the attraction of the most recent memory from the list is the strongest, it stays such even after that memory has been recalled. Therefore subsequent recalls are slowed down through reversions to that item. The absence of a corresponding silent primacy, on the other hand, is an unexpected finding. It may be related to the observation that recency typically appears earlier than primacy in the recall process [Fig. S1(b)] as the early recall of an item makes it more likely to be bounced afterwards.

A similar analysis can be performed by comparing the conditions in which the retrieval transition with a given *serial-position lag* would pass the repetition screening or not. The mean IRI in these two conditions can again be compared, to check for a delay associated with the possible discard of such retrievals (technical aspects in Appendix B). A consideration that can guide expectations on the outcome is that, as mentioned in the Introduction, the most likely lag to occur is $L = +1$ (sequential recalls). If this type of transition remains the most probable one also when it can cause bouncing (i.e., when it would yield an item already recalled), we expect it to be associated with the largest delay.

The conjectures are confirmed by inspection of the data (Fig. 4). Thus we find that another silent serial-position effect is the predominance of delays associated with the potential discarding of sequential retrievals ("silent contiguity effect").

IV. A MINIMAL MODEL

To further test the bouncing hypothesis as applied to repetition avoidance, it is necessary to rely on forward models that describe it in a predictive fashion.

Models of free recall have covered a spectrum ranging from dual-store memory search models [3,50] to powerful theories of temporal context dependence [51]. It is not our ambition to provide a wide-ranging model of top-down control or in any way revise what is known from existing models of recall (see Introduction); because our focus is entirely on the concept of bouncing, we rely on the most drastic simplification compatible with the general features of free recall.

The three assumptions we make to simplify the problem are the following [see Fig. 2(a)]:

(i) In the absence of any bouncing, the reported recall process would be identical to the retrieval process, and both would be a Markov chain in the space of the list items.

(ii) In the presence of bouncing, the retrieval process can hit an undesired memory item and bounce back to the previously recalled item, and this step does not affect the sequence of reported items.

(iii) Every repetition bounce entails a probability q of terminating the process, and we will refer to the parameter q as impatience.

Thus the only parameters are the Markov transition matrix and the impatience q (one additional parameter will be introduced later when we study strategic bouncing).

We have explored alternative assumptions for both the termination mechanism (static sink states, abrupt termination thresholds) and the non-Markovian add-on (one example is analyzed in detail in Appendix E), and they have proven less predictive than the model we are presenting.

The non-Markovian contribution from bouncing against a recalled item is, in this framework, of a peculiarly simple type; namely, the transition probability to a given item becomes zero after that item has been recalled and stays zero up to the end of the recall process. This non-Markovianity can be described as a progressive masking of the transition matrix. Let $\hat{\pi}$ be the naked transition matrix (i.e., the zero-diagonal Markov transition matrix of the bounce-free model) with matrix elements $\pi(y|x)$ determining the transition probability from item x to item y and, for any set S of serial positions, let $\pi(S|x) \equiv \sum_{y\in S} \pi(y|x)$.

Further naming N the length of the presented lists, T the number of trials in the sample, m_α the performance of trial α measured by the number of distinct reported items, and x^α_k the serial position of the kth recall in the trial, and using the notation where $x^\alpha_{i:j}$ is the set of serial positions recalled in between output positions i and j, both included, the normalized log likelihood of the bouncing model is

$$L[\hat{\pi}, q] = \ln q + \frac{1}{T}\sum_{\alpha=1}^{T} \ln \pi\left(x^\alpha_{1:m_\alpha-1}|x^\alpha_{m_\alpha}\right) + \frac{1}{T}\sum_{\alpha=1}^{T}\sum_{n=1}^{m_\alpha-1} \ln \pi\left(x^\alpha_{n+1}|x^\alpha_n\right) - \frac{1}{T}\sum_{\alpha=1}^{T}\sum_{n=1}^{m_\alpha} \ln\left[1-(1-q)\pi\left(x^\alpha_{1:n-1}|x^\alpha_n\right)\right]$$

(for the derivation, see Appendix C). Fitting the data set with this likelihood (Appendix D) yields an impatience parameter $q \approx 0.1$ and a transition matrix that encapsulates all known serial-position effects including some degree of primacy [Fig. S4(a)]. The number of bounces underlying every recall event can then be calculated by simulating the model with these parameters [for sample trials, see Fig. S4(b)].

To test the model, we assume for simplicity that the IRI is proportional to the bouncing count. We first compare the mean of relevant variables characterizing a recall event. The dependence on negative output position, retrieved serial position, and the recall lag are all qualitatively recovered (Fig. 5).

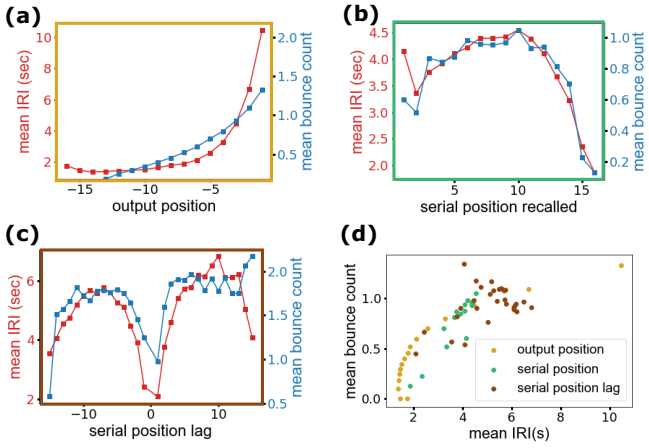

FIG. 5. Comparison between mean IRIs and the mean number of unreported retrievals of the model. The result of averaging is shown for fixed values of the recall event's output position (a), the serial position of the recalled item (b), and the serial-position lag from the preceding recall (c). In (d), a scatterplot of the values shown in (a)–(c) is presented.

We then test for the silent serial-position effects along the same lines followed for real data (Figs. 3 and 4), with the only difference that instead of measuring a mean temporal delay between two conditions, we now measure a difference in the mean number of bounces occurring within a recall event. We report that all main phenomenological features of silent serial-position effects, discussed above, are found in our pseudodata, both as concerns silent recency [Figs. 6(a) and 6(b)] and silent contiguity [Figs. 6(c) and 6(d)].

V. MODEL-FREE ANALYSIS OF SEQUENTIAL SLOWDOWN

We turn then to the distinctive shape of the sequentiality-performance curve [Fig. 1(b)]. We hypothesized that it might emerge if a partial discarding of nonsequential retrievals allows the implementation, in some trials, of a more markedly sequential recall strategy associated with higher performances—yielding the upturn of the curve for high-performing trials. [This interpretation would also justify the related fact that the output position of the final item becomes bimodally distributed for high-performing trials (Fig. S5), where the peak around the first output could be related to the recency effect and the peak around the final output could be related to the sequential strategy.] If that were the case, it should also leave traces in the statistics of the IRIs. Namely, we would expect that, at least among high performers, recalls performed by strategizing would be slower than those performed without strategizing.

To differentiate between strategic and nonstrategic trials, we can use the fact that the former have by construction a

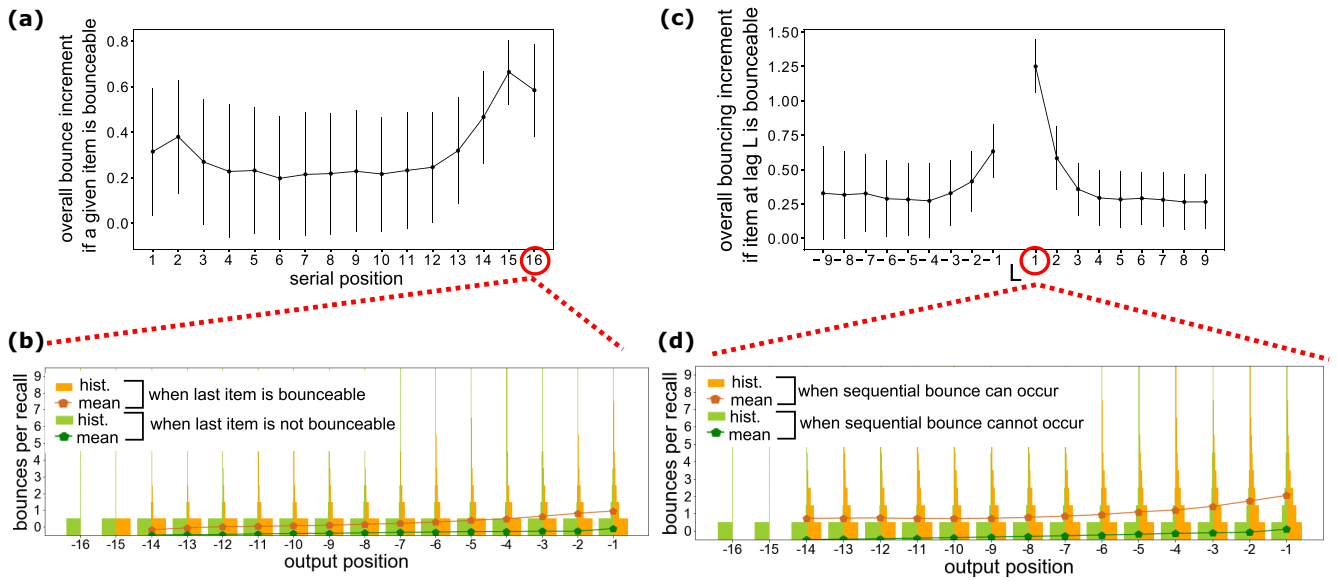

FIG. 6. Silent serial-position effects emerging from the bouncing model. (a) Selective variations in the mean bouncing counts estimated from model simulations. The mean number of bounces within a recall event has been averaged in the two subsamples of recalls where the list item in a given serial position is or is not bounceable, and the difference between the two means is plotted as a function of the serial position. Values on the x axis refer to the item by the bounceability of which subsamples are discriminated. (b) Bouncing-count histograms displayed side by side refer to recall events where the final item in the list is or is not bounceable. Corresponding mean curves show the mean bouncing increment after recall of the final item ("silent recency effect"). (c) Selective variations in the mean bouncing count per recall estimated from model simulations. The mean number of bounces within a recall event has been averaged in the two subsamples of recall events where a given lag does or does not lead to a bounceable item, and the difference between the two means is plotted as a function of the given lag. (d) Bouncing-count histograms displayed side by side refer to recall events where the sequential candidate for recall is or is not bounceable. Corresponding mean curves show the mean bouncing increment when the sequential item was already recalled ("silent contiguity effect").

higher trial sequentiality. An obvious approach is to separate each given sample of trials in the submedian and supramedian subsamples according to its distribution of trial sequentialities, and test whether recalls are faster in the submedian subsample.

We take into account three confounders. First, the dependence of the IRI on the negative output position can again play the role of a strong confounder. Second, since sequential strategies are used to achieve higher performances (see Introduction), the effect we are seeking may only be present in the more high-performing trials. We must thus proceed by considering separately event sets defined both by a given output position and by their occurring in trials with given performance. The third confounder is the difference in IRIs for sequential and nonsequential recalls within each trial. Indeed, sequential recalls are mostly faster than nonsequential ones across different performance levels and output positions [Fig. S6(a)].

For every condition defined by trial performance and output position, we consider thus the distribution of trial sequentiality among trials with that performance that contains a sequential recall in that output position [Fig. 7(a)]. We then split the sample by the median value of trial sequentiality and compare the IRIs of recall events in the two subsamples [Fig. 7(b)]. We only compare sequential recall events because highly sequential trials contain few nonsequential recalls. Thresholding out all sample pairs that do not pass a significance test, we calculate the sign of the delay associated with higher trial sequentiality. (Results are qualitatively unvaried whether we partition using a median computed over the sequentialities of all trials in a performance group or computed separately over those associated with sequential events at each given output position.)

We thus find the following [see Fig. 7(c)].

(i) In all cases where our significance requirement is met, the delay is positive, meaning that sequential recalls that occurred in sequentially biased trials have taken longer. In the following we refer to this phenomenon as *sequential slowdown*.

(ii) Strikingly, the phenomenon is concentrated in the high-performance regime. This is in fact just where it is to be found if the sequential slowdown is due to the implementation of recall strategies.

We will focus on the hypothesis that the slow sequential recalls are due to a subselection of the retrievals (for the problem with alternative explanations, see Discussion) or, equivalently, these delays are punctuated by the active rejection of what nonsequential retrievals are provided by free association, which is what we termed strategic bouncing.

VI. SEQUENTIAL SLOWDOWN IN THE BOUNCING MODEL

With model parameters coming from the abovementioned fit, we only add into the model a finite probability s ("strategicity") of bouncing back from a retrieved item if it is nonsequential. We are not necessarily assuming a common mechanism for strategic bouncing and repetition bouncing

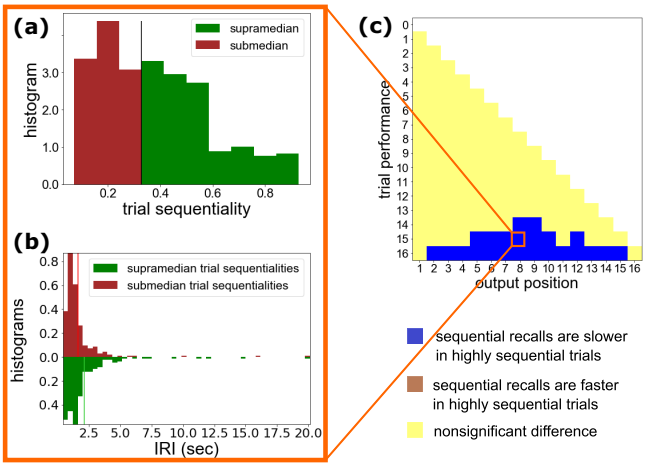

FIG. 7. "Sequential slowdown" in free-recall data. For each performance level, we list all trials that perform a sequential report at a given output position and study how its IRI changes in highly sequential trials. The sequentiality value associated with each trial is simply the fraction of sequential recalls in the trial. (a) The histogram of this quantity for output position 8 and trial performance 15 is further subdivided for the sake of the analysis into the set of trials above and below its median. (b) The corresponding distribution of IRIs in the two subsamples, with vertical lines with green and red colors indicating the respective mean values. (c) Sign of the difference in the mean values whenever the difference is significant by a Mann-Whitney test with threshold $P^* = 10^{-3}$. Sequentially biased trials are characterized by a slowdown of their sequential recalls, specifically in high-performing trials. Squares on the grid are colored as follows: *white* if data are not available from both subgroups (as is the case here because output position K does not exist for trials with performance $< K$); *yellow* if the Mann-Whitney P value between the two corresponding subsamples is larger than $P^* = 10^{-3}$; *blue* if the $P > P^*$ and the difference is positive, i.e., the supramedian trials perform significantly slower than sequential recalls; and *brown* for the opposite occurrence, which would be submedian trials performing significantly slower on sequential recalls, but strikingly, this is never observed.

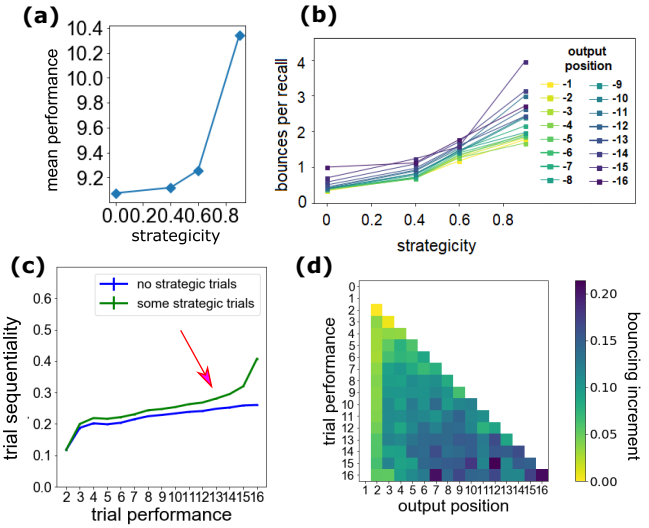

FIG. 8. Simulations of the bouncing model with strategizing. We simulated the model by adding in a given probability of strategic bouncing (i.e., the probability of rejecting a nonsequential retrieval), which we termed "strategicity." (a) Increasing strategicity also increases performance, coherently with the theory. (b) Increasing strategicity slows up the process, which agrees with the slowdown of sequential recall in highly sequential trials seen in the data. (c) The introduction of strategic bouncing leads to the emergence of the inflection point (red arrow). The features of real data that were shown in Fig. 1(b) are recovered here by merging a dominant population without strategicity, $s = 0$ (5×10^5 trials corresponding to the blue curve) with a smaller one having strategicity, $s = 0.5$ (5×10^3 trials). (d) Here, for each joint condition of performance and output position, sequential recalls are sampled above and below the median value of trial sequentiality. The mean difference in bouncing count proves to be always positive and increases in the area where sequential slowdown was detected in the data [compare Fig. 7(d)].

(the fact that strategic bouncing happens only in a minority of trials may point to the opposite) but only that the basic retrieval rejection step can also be described as a bounce. A priority needs to be established between the two types of bounce; we assume that subjects who are actively implementing a sequential strategy check first for sequentiality and then for repetitiousness and do not associate any impatience increment with strategic bounces. This proves to be the sensible choice from the requirement that the performance should increase as a function of strategicity [Fig. 8(a)] despite repetition bouncing (Fig. S1).

The analysis of sequential slowdown in data suggests that this increase in performance should come at the price of a slowdown in individual recalls, measured by the number of bounces. We test this statement on simulations of the model and find that indeed it holds true for simulated recalls with any possible value of their negative output position [Fig. 8(b)].

We then move on to replicating the results of data analysis for the dependence of trial sequentiality on performance that was presented at the outset [Fig. 1(b)]. Rather than aiming at numerical accuracy, we seek to understand the shape of the curve, namely, its inflection in midperformance, followed by a drastic rise for larger performances. First we note that repetition bounces alone do not explain this feature [Fig. 8(c)]. We then simulate a mixture of two populations, a dominant one without strategicity, $s = 0$ (5×10^5 trials), and a smaller one with strategicity, $s = 0.5$ (5×10^3 trials). When plotting the sequentiality of simulated trials versus their performance, we find that even such a small percentage of strategic trials is sufficient to inflect the curve [Fig. 8(c)].

Finally, we agnostically interrogate the mixed-set pseudodata model for the signatures of strategic slowdown we encountered in Fig. 7(c). We subsample recall events by the output position and trial performance characterizing them; for each joint condition, we use the sequentiality values of the relevant trials to subdivide the sample in a supramedian and submedian group and compute the bouncing increment, i.e., the difference between their mean bouncing counts.

This difference proves to be always positive [Fig. 8(d)], signifying a delay in time as found in real data. In Fig. 7(c), we also showed that the time delay found in real data grows into significance in the high-performance region. The modeling result is in agreement with real data because whenever the data show a statistically significant difference in IRIs, the

corresponding subsample of the model-generated pseudodata features relatively high values of the bouncing increment. This result supports the notion that sequential slowdown is due to strategic bouncing [Fig. 8(d)].

VII. DISCUSSION

Free-recall experiments are an essential behavioral probe in the study of human memory. In this paper, we undertook the rigorous mining of a large data set to uncover a number of features: (i) The number of repetitions observed in these experiments is vastly smaller than a basic expectation given by chance Markovian transitions; (ii) recalls are delayed if the final item has already been recalled ("silent recency effect"); (iii) recalls are delayed if sequential recall would yield a repetition ("silent contiguity effect"); (iv) the overall abundance of sequential recalls in the trials has a distinctively inflected shape when regarded as a function of trial performance; and (v) a slowdown in certain recall events is associated with the overall prevalence of sequential recalls in high-performing trials ("strategic sequential slowdown").

In addition to reporting on the above, we proceeded to suggest a unifying explanation coherent with the current body of knowledge on free recall (see Introduction) and based on the notion that free-association retrievals are sometimes "bounced," i.e., discarded. In a majority of cases this happens because they are repetitious (repetition bouncing) and in a minority because they are not sequential (strategic bouncing). To test the validity of our explanation, we used a minimal model with trajectories driven by heterogeneous transition probabilities and reset by bouncing. Bouncing of repetitious transitions occurs systematically for repetitious retrievals and adds to the "impatience" that, probabilistically, leads to termination of a recall trial; strategic bouncing of nonsequential transitions occurs in high-performing trials with a fixed probability. Upon fitting on the data and using the number of bounces as a proxy for the inter-recall time intervals (IRIs), the model was shown to qualitatively account for all the above features of free recall.

The minimal model accounts in one breath for repetition bouncing and strategic bounding, and may account also for a gradual release from repetition avoidance [52–54] by additionally incorporating what is known about its dependence on cognitive rather than physical time [55]. However, the features we uncovered by our model-free analysis of data should be simultaneously recoverable from a variety of models, such as the stochastic model of Ref. [56] and the generally valid framework of the context maintenance and retrieval model of Refs. [2,29]. All of the effects highlighted here can be explained on the basis of quantitative differences in the effective memory strength of different items; they do not require or rule out qualitative differences, such as the final item possibly being recalled from working memory, which may be faster than associative recall involving other sources.

Our work underscores the importance of unreported retrievals in determining behavioral observables of memory experiments. It is indeed easy to see that alternative explanations would run into several problems. For example, one might hypothesize that the strategy at play in certain high-performing trials consists merely in waiting longer before surrendering even if no new memory is being retrieved. This would predict a slowdown preferentially happening at later output positions, where surrendering is most prominent. However, such an explanation can be ruled out because the sequential slowdown is observed at a wide range of output positions [Fig. 7(c)].

In experiments where subjects are instructed to recall items not belonging to a specified category or items not beginning with certain letters [8], other types of bouncing might take place, but their detection can be attempted by the very means we have introduced.

Measurements performed in free recall are not limited to behavioral observables. A parallel study of unreported retrievals could be attempted through the decoding of suitable neurophysiological data. The timing and magnitude of postretrieval effects such as those already observed in electrocorticography (ECoG) of prefrontal cortex [5] could be compared with the timing and quantity of the bounces inferred by an analysis of behavioral data.

The issue of whether bouncing is consciously perceived by the participants is more delicate and could be partly addressed in the future through comparisons with externalized free recall [34,36,37] with separate controls for the accuracy and consistency of self-reporting.

Data used in this paper were generously made available at the University of Pennsylvania and can be downloaded from [57]. The PYTHON/PYTORCH library we developed to process free-recall data sets is available upon request.

ACKNOWLEDGMENTS

The study was supported by RIKEN Center for Brain Science, Brain/MINDS from AMED under Grant No. JP15dm0207001, and JSPS KAKENHI Grant No. JP18H05432. F.F., Z.H., Ł.K., and T.T. worked as a team; credit is not individualizable. All the authors reviewed the manuscript together. The authors declare no competing interests.

APPENDIX A: ARCHIVAL DATA

We base our analysis on a publicly available data set, collected as part of the Penn Electrophysiology of Encoding and Retrieval Study at the University of Pennsylvania (for a full account, see Ref. [43]). The data were acquired from consenting human subjects in compliance with the University of Pennsylvania's Institutional Review Board protocol.

Participants were given 75 s to attempt to recall aloud any of the presented items. Because of the limited amount of time, quantities pertaining to the first negative output position are indeed somewhat affected by the final rush to recall.

In our analysis, all trials (27 198 in total) were used. In the minority of trials containing some repetition [Fig. S2(a)], repetitions were ignored for the main analyses. Intrusions from outside the presented lists were all removed at the outset.

APPENDIX B: MODEL-FREE ANALYSES

We classified all recalls into the prerecall group or postrecall group with respect to a target item. The target item was characterized either by its serial position (Fig. 3) or by its

serial-position lag with respect to the latest item recalled (in which case the serial position of the target item varies from event to event; Fig. 4). Recall events are then classed as "postrecall" if they occur when the target item has already been recalled within the trial, and they are classed as "prerecall" if they occur at a moment of the trial when the target item has not yet been recalled.

As discussed in the main text, we aim at finding evidence of whether the target item may be retrieved and unreported during the recall events. Thus we are interested in the difference between the IRIs in the prerecall and postrecall conditions.

Moreover, we studied whether the above changes in the IRIs were output position dependent because IRIs are known to systematically increase with output positions. For this purpose, for each target item, we also divided recalls by their negative output positions and tested whether the IRI distributions of the prerecall and postrecall groups were significantly different at each output position. We checked whether the IRIs of the pre- and postrecall groups were statistically different using the Mann-Whitney U test because the IRI distributions were highly non-Gaussian.

To make the analysis of IRIs rigorous, we took the following precautions.

(i) We broke down the sampling of recall events by negative rather than positive output position to account for the character of the output position dependence highlighted in Fig. S3.

(ii) Events where the target item is reported should in principle be classed as prerecall, because it was not recalled yet while the recall process occurred. However, we excluded these events from the prerecall condition. Note that reported recalls of the target item could never figure in the postrecall sample for that target. Therefore including such events in the prerecall sample could unbalance the two comparison of the two groups.

(iii) In the case of unreported retrievals with a fixed lag, we count out all conditions where the lag would not yield an item within the list (due to the finite list length), as that makes recall unavailable rather than bounceable.

In all the subsample comparisons of the IRIs, we systematically filtered out information by testing for significance the difference between the IRIs in each pair of subsamples. Since the distribution of IRIs in any given subsample is highly non-Gaussian [see, e.g., Fig. 3(a)], rather than using a t test we adopted a Mann-Whitney U test [58].

For example, Fig. S6(a) should be strictly understood as stating that if we pool trials regardless of their sequentiality, the statement that sequential recalls are faster withstands a significance test with few exceptions on the triangular grid spanned by performance and output position values.

The P-value threshold we chose at the outset was $P^* = 10^{-3}$ and has been uniformly applied to all the analyses.

APPENDIX C: BOUNCING MODEL

In the bouncing model as set up in the main text, upon retrieving a repetitious item, the process reverts to the latest previously retrieved item and makes a new "try" from there. Here, we formalize the concept mathematically, writing down the log likelihood for the model.

Let us label the items by their serial position $x = 1, \ldots, N$, where N is the length of the memorized list. The first recalled item of every trial is picked according to a distribution $p_1(x) = \rho_{\text{init}}(x)$ which plays the role of an initial condition. The second item has probability $p_2(x) = \sum_{x_1} \pi(x|x_1)p_1(x_1)$, where $\pi(y|x)$ is the naked transition matrix. This matrix is assumed diagonal-free, so $\pi(x_1|x_1) = 0$, and no repetition can occur in the second recall.

At the third recall, for every previous history (x_1, x_2), there is a finite probability $\pi(x_1|x_2)$ of ending up in a repetition. If this happens, the trajectory goes back to x_2 without outputting either retrieval, which we refer to as a "bounce." In other words, there is a set J of undesired retrievals that contains in this case only item x_1 ($J = \{x_1\}$).

If an infinite number of bounces is allowed, the transition from the retrieval of x_2 to the next is governed by a matrix $\Pi_0(x|x_2; J)$ defined as

$$\Pi_0(y|x; J) = \frac{\pi(y|x)\mathbf{1}[y \notin J]}{\sum_{y \notin J} \pi(y|x)}, \tag{C1}$$

where $\mathbf{1}[X] = 1$ if statement X is true, and 0 otherwise.

In other words, the transition matrix is defined with the target items in the bounceable set masked away, and every other element is normalized so the columns sum to unity. Notice the special case $\Pi_0(y|x; \emptyset) = \pi(y|x)$.

This is, however, only correct if any number of bounces is allowed; in practice, it is realistic that the subject will give up after a certain number of bounces. Every repetition bounce is thus associated with a fixed surrender probability q we call "impatience" (although different modelings of impatience may also be devised).

Call $J_n = [x_1, x_2, \ldots, x_{n-1}]$ the (unordered) set of all items recalled up to time n. The probability $p_{n+1}(y)$ of item y being the $(n+1)$th recall is given by

$$\Pi(y|x_n; J_n) = \Pi_0(y|x_n; J_n)[1 - s_q(x_n; J_n)], \tag{C2}$$

where $s_q(x; J)$ is the probability of surrendering, i.e., terminating the process, during the recall step starting out from recall of item x, given that the bounceable set is J.

At each step the process has three possibilities: recalling, bouncing, or surrendering. Recall happens if a new word is retrieved before surrendering; surrendering occurs with probability q whenever a repetitious word is retrieved; and bouncing occurs with probability $1 - q$ when a repetitious word is retrieved, returning the process to the latest recalled item.

Calling the three corresponding probabilities s, b, and r, we clearly have

$$s_q(x, J) + b_q(x, J) + r_q(x, J) = 1. \tag{C3}$$

For a single retrieval step, their elemental values are

$$s_q^{(1)}(x, J) = \pi(J|x)q, \tag{C4}$$

$$b_q^{(1)}(x, J) = \pi(J|x)(1 - q), \tag{C5}$$

$$r_q^{(1)}(x, J) = \pi(\bar{J}|x), \tag{C6}$$

where $\bar{J}$ is the allowed set, i.e., the complement of J, and we used the notation $\pi(S|x) = \sum_{y \in S} \pi(y|x)$ for an arbitrary subset S of the N list items.

With a potentially infinite number of steps we must write

$$s_q(x;J) = \sum_{n=0}^{\infty} \left(b_q^{(1)}(x,J)\right)^n s^{(1)}(x,J)q = \sum_{n=0}^{\infty} [\pi(J|x)(1-q)]^n \pi(J|x)q. \tag{C7}$$

Merging Eqs. (C2) and (C7), we arrive at

$$\Pi(y|x;J) = \Pi_0(y|x;J)\left[1 - \sum_{n=0}^{\infty} [\pi(J|x)(1-q)]^n \pi(J|x)q\right]. \tag{C8}$$

We can now merge in Eq. (C1) to eliminate Π_0, obtaining

$$\Pi(y|x;J) = \frac{\pi(y|x)\,\mathbf{1}[y \notin J]}{1-(1-q)\pi(J|x)}, \quad \forall y \neq 0, \tag{C9}$$

where we made it explicit that this does not provide the probability of transition to silence, which can be described in terms of an effective sink state $y_{\text{sink}} \equiv 0$.

The only missing ingredient is now the equivalent of the probabilities (C9) for the case of transitions to the sink state. Transitions from the sink state have probability 1 of staying there and are therefore not problematic:

$$\Pi(y|0;J) = \mathbf{1}[y=0]. \tag{C10}$$

As for the eventuality of recall termination (transition from a list item to the sink), it corresponds to a probability

$$\Pi(0|x;J) = \frac{\pi(J|x)q}{1-(1-q)\pi(J|x)}. \tag{C11}$$

The normalized log likelihood is a function of the scalar parameter q and of the matrix parameter $\hat{\pi}$ and can be written as

$$\mathcal{L}[\hat{\pi}, q, \vec{\lambda}] = \frac{1}{T}\sum_{\alpha=1}^{T} L\left[x^{\alpha}_{1:N};\hat{\pi},q\right], \tag{C12}$$

where T is the number of trials and we used the notation

$$x^{\alpha}_{1:N} = \{x^{\alpha}_1, \ldots, x^{\alpha}_N\}$$

(note that in this convention the last item is also included).

The one-trial log likelihood appearing in Eq. (C12) is defined as

$$L[x_{1:N};\hat{\pi},q] = \sum_{n=1}^{N-1} \ln \Pi(x_{n+1}|x_n;x_{1:n-1}) \tag{C13}$$

with Π computed according to Eqs. (C9)–(C11) for the given point in the $(\hat{\pi}, q)$ parameter space.

Calling m_α the number of words recalled in the αth trial, we can merge the above formulas into

$$L[x_{1:N};\hat{\pi},q] = \sum_{n=1}^{m_\alpha-1} \ln \frac{\pi(x_{n+1}|x_n)}{1-(1-q)\pi(x_{1:n-1}|x_n)} + \ln \frac{\pi(x_{1:m_\alpha-1}|x_{m_\alpha})q}{1-(1-q)\pi(x_{1:m_\alpha-1}|x_{m_\alpha})}, \tag{C14}$$

from which, separating the arguments of the logarithms,

$$L[x_{1:N};\hat{\pi},q] = \ln q + \ln \pi(x_{1:m_\alpha-1}|x_{m_\alpha}) + \sum_{n=1}^{m_\alpha-1} \ln \pi(x_{n+1}|x_n) - \sum_{n=1}^{m_\alpha} \ln[1-(1-q)\pi(x_{1:n-1}|x_n)]. \tag{C15}$$

To account for the column-wise normalization of the stochastic matrix Π, instead of using Lagrange multipliers, we opt for adding a degree of freedom by writing $\pi(y|x) = u(y|x)/u(\text{all}|x)$ in terms of non-negative auxiliary variables $u(y|x)$ defined for $x \neq y$ without any normalization constraint, and with

$$u(\text{all}|x) = \sum_{y=1}^{N} u(y|x); \tag{C16}$$

the $\hat{u}$ matrix can be assumed to have a zero diagonal.

In terms of the matrix $\hat{u}$, the one-trial log likelihood (C13) becomes

$$L[x_{1:N};\hat{u},q] = \ln q + \ln u(x_{1:m_\alpha-1}|x_{m_\alpha}) + \sum_{n=1}^{m_\alpha-1} \ln u(x_{n+1}|x_n) - \sum_{n=1}^{m_\alpha} \ln[u(\text{all}|x_n) - (1-q)u(x_{1:n-1}|x_n)]. \tag{C17}$$

The corresponding q component of the log likelihood's gradient is

$$\mathcal{D}^q[\hat{u},q] = \frac{\partial}{\partial q}\mathcal{L} = \frac{1}{q} - \frac{1}{T}\sum_{\alpha=1}^{T}\sum_{n=1}^{m_\alpha} \frac{u\left(x^{\alpha}_{1:n-1}\middle|x^{\alpha}_n\right)}{u\left(\text{all}\middle|x^{\alpha}_n\right) - (1-q)u\left(x^{\alpha}_{1:n-1}\middle|x^{\alpha}_n\right)}. \tag{C18}$$

As for the u derivatives, we have

$$\mathcal{D}^u_{yx}[\hat{u},q] = \frac{\partial}{\partial u_{yx}}\mathcal{L} = -\frac{1}{T}\sum_{\substack{(\alpha,k):\\ x^{\alpha}_k = x}} \frac{1}{u(\text{all}|x) - (1-q)u\left(x^{\alpha}_{1:k-1}\middle|x\right)} + \frac{1}{T}\sum_{\substack{(\alpha,k,l):\\ x^{\alpha}_k = x, x^{\alpha}_l = y}} \left[\frac{\mathbf{1}(k=m_\alpha)}{u\left(x^{\alpha}_{1:m_\alpha-1}\middle|x\right)} + \frac{\mathbf{1}(l=k+1)}{u(y|x)} + \frac{(1-q)\mathbf{1}(k>l)}{u(\text{all}|x) - (1-q)u\left(x^{\alpha}_{1:k-1}\middle|x\right)}\right]. \tag{C19}$$

As a sanity check for this last formula, one can isolate the contribution from trials where $x_N^\alpha = x$ and find that it is exactly zero, as it should be because no transitions to the sink are observed from the last recalled item in perfect-recall trials ($m_\alpha = N$), and therefore the likelihood for such trials is not affected by the u_{yx_N} parameter for any y.

During the numerical search for an optimum, it is convenient to remove the lower bounds at zero from the elements of $\hat{u}$ and the [0,1] bounds from q. We do so by mapping both sets of variables to the full real axes. For the sake of convenience we used the transformation

$$s = \log\left(\frac{1}{q} - 1\right), \quad v = \log(u), \tag{C20}$$

so that the gradient becomes

$$\mathcal{D}^s[\hat{v}, s] = \frac{e^s}{(1+e^s)^2}\mathcal{D}^q\left[\exp(\hat{v}), \frac{1}{1+e^s}\right], \tag{C21}$$

$$\mathcal{D}^v_{yx}[\hat{v}, s] = \exp[v(y|x)]\,\mathcal{D}^u_{yx}\left[\exp(\hat{v}), \frac{1}{1+e^s}\right], \tag{C22}$$

or in terms of the physical variables,

$$\mathcal{D}^s = -q(1-q)\mathcal{D}^s_{yx}[\hat{u}, q], \tag{C23}$$

$$\mathcal{D}^v_{yx} = u(y|x)\,\mathcal{D}^u_{yx}[\hat{u}, q]. \tag{C24}$$

APPENDIX D: SIMULATIONS AND FITTING

To run simulations of the bouncing model with and without strategicity, we extract the normalized histogram of the full data set at output position = 1 and use it as our initial condition. Simulation of 10^5 trials takes minutes on an ordinary laptop.

To fit the model, we write down the gradient equations (C23) and (C24) and search for the minimum from a random initial condition using the limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm for bound constraints (L-BFGS-B algorithm) [59]. This converges in $< 10^2$ iterations to the matrix shown in Fig. S4(a) and to the optimal impatience $q \approx 0.1$.

Notice that sequentiality partially comes from the Markov transition probabilities and partially from strategicity. Our focus is not to achieve the best quantitative model but to explain how the strategicity sequentializes only high-performance trials, which yields a major difference between the two mechanisms of increasing sequentiality. For simplicity we used all trials to fit transition probabilities, which in principle may overestimate nonstrategic sequentiality.

Fitting and simulations based on an alternative "skipping model" are described in Appendix E.

APPENDIX E: THE SKIPPING MODEL

An alternative model through which one may attempt to reproduce the silent serial-position effects is what we will refer to as the skipping model, where the Markovian retrieval process is not interfered with by the higher-order process that merely censors the reporting of repetitious retrievals.

A repetitious retrieval is thus not discarded as in the bouncing model, but adopted as the starting point for the next transition (which will be described as a "skip" rather than as a "bounce"; see Fig. S7).

As in the bouncing model, an $N \times N$ Markov transition matrix π is defined such that π_{xy} is the probability of retrieving word x after retrieving word y. We assume again that this probability is independent of the time step throughout the retrieval process. We also include among the states of the Markov chain a sink state accounting for termination.

The transition from the nth to the $(n+1)$th recall is governed by a "recall propagator" $T_{\{x_1,\ldots,x_n\}}$, defined as a matrix that depends parametrically on the potentially repetitious (hence unreportable) set of serial positions $\{x_1, \ldots, x_n\}$, which is the set of words recalled up to that moment.

In other words, for any set of serial positions S, we are defining the matrix T_S such that $[T_S]_{xy}$ is the probability of recalling x after recalling y, given that the items in the set S cannot be reported, having already been recalled (though they can be retrieved). Since recalls do not happen at all retrieval events, skipping repetitious retrievals, for $x \in S$ or $y \notin S$ we have $[T_S]_{xy} = 0$ [Fig. S8(a)].

Since the data set samples transitions driven by the T_S matrix, the π matrix can be obtained via maximizing the likelihood of the T_S matrix given the data set. Unfortunately, even if infinite skips are allowed in a single recall event and no sink state is included, a closed formula such as Eq. (C1) is not available for T_S. Although a Dyson-like summation can be performed for this propagator, when computing the log likelihood on data, implementing it leads to a combinatorial blowup of the number of possible paths.

Thus, while the retrieval process in the skipping model is simpler than in the bouncing model (because it stays completely Markovian), the fitting is much less straightforward.

On the other hand, once the π matrix is given, the T matrix can always be calculated iteratively. In order to practically estimate the π matrix, we exploited the fact by making an extra assumption and putting an upper limit $n_{\max}$ on the possible number of bounces. We were thus able to build a recurrent neural network (RNN)-style computational graph as illustrated in Fig. S8(b).

For any set of serial positions S, let us define the matrix D_S such that

$$[D_S]_{yz} = \sum_{x \in S} \delta_{yx}\delta_{xz}.$$

Given a certain source item x and a repetitious set $S + x = S \cup x$, the T matrix can be calculated as

$$T_S = \sum_{n=0}^{n_{\max}} D_{\bar{S}}\,\pi\,(D_{S+x}\,\pi)^n, \tag{E1}$$

where $\bar{S}$ is the allowed output set and $S + x$ is the unreportable (i.e., repetitious) set; x, which is the source, is included in the unreportable set.

$\pi D_{\bar{S}}$ is here a skip-connect indicating the end of a bounce when an element recalled is in the reportable set. This graph now also allows transitions from x to itself, a hitch that can be solved by masking out diagonal elements of the π matrix. Even without masking, however, the diagonal elements of the π matrix vanish after fitting the data.

We implement the computational graph with PYTORCH. The negative log-likelihood loss function between the calculated source-to-target log probabilities with the current π matrix and observed target sample is used as the loss function. A certain batch number of unreportable sets and reported targets is randomly picked from the data set to carry out a stochastic-gradient-descent parameter update of the π matrix.

Let us call x_m^α the serial position of the word recalled in the mth recall of trial α. Call b the size of minibatches. Each minibatch selection consists of a set $\vec{\alpha}$ of b random trial indices α_k and of a set $\vec{m}$ of corresponding random output positions m_k. Here, $k = 1, \ldots, b$, and for each k, $1 \leqslant \alpha_k \geqslant N_T$ (N_T being the number of trials) and $1 \leqslant m_k \geqslant L$ (L being the list length).

The loss function for a particular batch is then

$$\mathcal{L}(\pi; \vec{\alpha}, \vec{k}) = -\left\langle \log \left[T_{\{x_1^{\alpha_k}, \ldots, x_{m_k-1}^{\alpha_k}\}} \right]_{x_{m_k+1}^{\alpha_k}, x_{m_k}^{\alpha_k}} \right\rangle_{k=1,\ldots,b},$$

where the average is used instead of a sum to keep the values of the loss function numerically low even with many batches. It takes approximately tens of minutes to obtain convergence on a conventional laptop.

The resulting matrix parameter π, shown in Fig. S8(c), can be used to simulate the model starting from the empirically observed initial condition. Both the silent recency effect and the silent contiguity effect are successfully reproduced (Fig. S9).

However, an additional "silent primacy effect" (where past recall of the first item in the list engenders a delay in subsequent recalls) emerges nearly as conspicuously as silent recency (Fig. 6). This is compatible with the existence of a corresponding primacy effect in reported recall frequencies [16] but is not compatible with the data (Fig. 3).

[1] S. D. Gronlund and R. M. Shiffrin, Retrieval strategies in recall of natural categories and categorized lists, J. Exp. Psychol.: Learn. Mem. Cognit. **12**, 550 (1986).

[2] S. M. Polyn, K. A. Norman, and M. J. Kahana, A context maintenance and retrieval model of organizational processes in free recall, Psychol. Rev. **116**, 129 (2009).

[3] J. G. Raaijmakers and R. M. Shiffrin, Search of associative memory, Psychol. Rev. **88**, 93 (1981).

[4] S. Becker and J. Lim, A computational model of prefrontal control in free recall: Strategic memory use in the California Verbal Learning Task, J. Cognit. Neurosci. **15**, 821 (2003).

[5] Y. Norman, E. M. Yeagle, M. Harel, A. D. Mehta, and R. Malach, Neuronal baseline shifts underlying boundary setting during free recall, Nat. Commun. **8**, 1301 (2017).

[6] H. R. Hayama, J. D. Johnson, and M. D. Rugg, The relationship between the right frontal old/new ERP effect and post-retrieval monitoring: Specific or non-specific? Neuropsychologia **46**, 1211 (2008).

[7] H. R. Hayama and M. D. Rugg, Right dorsolateral prefrontal cortex is engaged during post-retrieval processing of both episodic and semantic information, Neuropsychologia **47**, 2409 (2009).

[8] H. L. Roediger III and E. Tulving, Exclusion of learned material from recall as a postretrieval operation, J. Verbal Learn. Verbal Behav. **18**, 601 (1979).

[9] M. D. Rugg, J. D. Johnson, and M. R. Uncapher, Encoding and retrieval in episodic memory: Insights from fMRI, in *The Wiley Handbook on the Cognitive Neuroscience of Memory* (Wiley, New York, 2015), Chap. 5, p. 84.

[10] D. Badre, R. A. Poldrack, E. J. Paré-Blagoev, R. Z. Insler, and A. D. Wagner, Dissociable controlled retrieval and generalized selection mechanisms in ventrolateral prefrontal cortex, Neuron **47**, 907 (2005).

[11] N. M. Long, I. Öztekin, and D. Badre, Separable prefrontal cortex contributions to free recall, J. Neurosci. **30**, 10967 (2010).

[12] C. R. Savage, T. Deckersbach, S. Heckers, A. D. Wagner, D. L. Schacter, N. M. Alpert, A. J. Fischman, and S. L. Rauch, Prefrontal regions supporting spontaneous and directed application of verbal learning strategies: Evidence from PET, Brain **124**, 219 (2001).

[13] D. T. Stuss, M. P. Alexander, C. L. Palumbo, L. Buckle, L. Sayer, and J. Pogue, Organizational strategies with unilateral or bilateral frontal lobe injury in word learning tasks, Neuropsychology **8**, 355 (1994).

[14] R. Cabeza, E. Ciaramelli, I. R. Olson, and M. Moscovitch, The parietal cortex and episodic memory: An attentional account, Nat. Rev. Neurosci. **9**, 613 (2008).

[15] H. Ebbinghaus, *Memory: A Contribution to Experimental Psychology*, 1885 (*Teachers College*, Columbia University, New York, 1913).

[16] B. B. Murdock Jr., The serial position effect of free recall, J. Exp. Psychol.: Learn. Mem. Cognit. **64**, 482 (1962).

[17] E. Tulving, Subjective organization in free recall of "unrelated" words, Psychol. Rev. **69**, 344 (1962).

[18] L. Postman and L. W. Phillips, Short-term temporal changes in free recall, Q. J. Exp. Psychol. **17**, 132 (1965).

[19] A. K. Bousfield and W. A. Bousfield, Measurement of clustering and of sequential constancies in repeated free recall, Psychol. Rep. **19**, 935 (1966).

[20] A. Binet and V. Henri, Mémoire des mots, Ann. Psychol. **1**, 1 (1894).

[21] E. A. Kirkpatrick, An experimental study of memory, Psychol. Rev. **1**, 602 (1894).

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

[23] M. J. Kahana, Associative retrieval processes in free recall, Mem. Cognit. **24**, 103 (1996).

[24] M. W. Howard and M. J. Kahana, Contextual variability and serial position effects in free recall, J. Exp. Psychol.: Learn. Mem. Cognit. **25**, 923 (1999).

[25] M. K. Healey, N. M. Long, and M. J. Kahana, Contiguity in episodic memory, Psychon. Bull. Rev. **26**, 699 (2019).

[26] M. J. Kahana and J. B. Caplan, Associative asymmetry in probed recall of serial lists, Mem. Cognit. **30**, 841 (2002).

[27] S. Farrell and S. Lewandowsky, Response suppression contributes to recency in serial recall, Mem. Cognit. **40**, 1070 (2012).

[28] J. I. Vousden and G. D. Brown, To repeat or not to repeat: The time course of response suppression in sequential behaviour, in *4th Neural Computation and Psychology Workshop, London, 9–11 April 1997* (Springer, New York, 1998), pp. 301–315.

[29] L. J. Lohnas, S. M. Polyn, and M. J. Kahana, Expanding the scope of memory search: Modeling intralist and interlist effects in free recall, Psychol. Rev. **122**, 337 (2015).
[30] S. Romani, M. Katkov, and M. Tsodyks, Practice makes perfect in memory recall, Learn. Mem. **23**, 169 (2016).
[31] R. Gianutsos, The order of recall and the recall of order, Mem. Cognit. **4**, 627 (1976).
[32] G. Keppel, L. Postman, and B. Zavortink, Studies of learning to learn: VIII: The influence of massive amounts of training upon the learning and retention of paired-associate lists, J. Verbal Learn. Verbal Behav. **7**, 790 (1968).
[33] Q. Zhang, T. Griffiths, and K. Norman, Optimal policies for free recall, PsyArXiv (2021), doi: 10.31234/osf.io/sgepb.
[34] R. M. Hogan, Interitem encoding and directed search in free recall, Mem. Cognit. **3**, 197 (1975).
[35] N. Unsworth, G. A. Brewer, and G. J. Spillers, Understanding the dynamics of correct and error responses in free recall: Evidence from externalized free recall, Mem. Cognit. **38**, 419 (2010).
[36] H. L. Roediger and D. G. Payne, Recall criterion does not affect recall level or hypermnesia: A puzzle for generate/recognize theories, Mem. Cognit. **13**, 1 (1985).
[37] M. J. Kahana, E. D. Dolan, C. L. Sauder, and A. Wingfield, Intrusions in episodic recall: Age differences in editing of overt responses, J. Gerontol. Ser. B: Psychol. Sci. Social Sci. **60**, P92 (2005).
[38] N. Unsworth and G. A. Brewer, Variation in working memory capacity and intrusions: Differences in generation or editing? Eur. J. Cognit. Psychol. **22**, 990 (2010).
[39] C. Aguirre, C. J. Gómez-Ariza, and M. T. Bajo, Selective directed forgetting: Eliminating output order and demand characteristics explanations, Q. J. Exp. Psychol. **73**, 1514 (2020).
[40] H. R. Pollio, R. A. Kasschau, and H. E. Denise, Associative structure and the temporal characteristics of free recall, J. Exp. Psychol.: Learn. Mem. Cognit. **76**, 190 (1968).
[41] H. R. Pollio, S. Richards, and R. Lucas, Temporal properties of category recall, J. Verbal Learn. Verbal Behav. **8**, 529 (1969).
[42] K. E. Patterson, R. H. Meltzer, and G. Mandler, Inter-response times in categorized free recall, J. Verbal Learn. Verbal Behav. **10**, 417 (1971).
[43] M. K. Healey, P. Crutchley, and M. J. Kahana, Individual differences in memory search and their relation to intelligence, J. Exp. Psychol.: General **143**, 1553 (2014).
[44] B. B. Murdock and R. Okada, Interresponse times in single-trial free recall, J. Exp. Psychol.: Learn. Mem. Cognit. **86**, 263 (1970).
[45] D. Rohrer and J. T. Wixted, An analysis of latency and interresponse time in free recall, Mem. Cognit. **22**, 511 (1994).
[46] W. J. McGill, Stochastic latency mechanisms, in *Handbook of Mathematical Psychology* (Wiley, New York, 1963), Vol. 1, p. 309.
[47] D. Vorberg and R. Ulrich, Random search with unequal search rates: Serial and parallel generalizations of McGill's model, J. Math. Psychol. **31**, 1 (1987).
[48] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevResearch.4.033089 for details.
[49] F. Fumarola, A diffusive-particle theory of free recall, Adv. Cognit. Psychol. **13**, 201 (2017).
[50] J. G. Raaijmakers and R. M. Shiffrin, SAM: A theory of probabilistic search of associative memory, Psychol. Learn. Motiv.: Adv. Res. Theory **14**, 207 (1980).
[51] M. W. Howard and M. J. Kahana, A distributed representation of temporal context, J. Math. Psychol. **46**, 269 (2002).
[52] D. Rundus, Analysis of rehearsal processes in free recall, J. Exp. Psychol.: Learn. Mem. Cognit. **89**, 63 (1971).
[53] R. N. Henson, Item repetition in short-term memory: Ranschburg repeated, J. Exp. Psychol.: Learn. Mem. Cognit. **24**, 1162 (1998).
[54] D. Laming, Serial position curves in free recall, Psychol. Rev. **117**, 93 (2010).
[55] M. Duncan and S. Lewandowsky, The time course of response suppression: No evidence for a gradual release from inhibition, Memory **13**, 236 (2005).
[56] D. Laming, Failure to recall, Psychol. Rev. **116**, 157 (2009).
[57] http://memory.psych.upenn.edu/files/pubs/HealEtal14.data.tgz.
[58] T. P. Hettmansperger and J. W. McKean, *Robust Nonparametric Statistical Methods* (CRC, Boca Raton, FL, 2010).
[59] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, A limited memory algorithm for bound constrained optimization, SIAM J. Sci. Comput. **16**, 1190 (1995).