Detecting unreported recalls in memory experiments

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

[1]Laboratory for Neural Computation and Adaptation, RIKEN Center for Brain Science, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan

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

[*]taro.toyoizumi@riken.jp

Abstract

We investigate the role of discarded memory retrievals in experiments on free recall from lists of items. Such retrievals are not explicitly reported but testable predictions can be derived by assuming that they correlate with a delay in the timings of recall. We report on three situations where such a delay occurs: 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 highly performing trials ("sequential slowdown"). All these effects can be reproduced by a minimal model where the discarding of memories ("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.

Introduction

Recall of memories in humans is increasingly regarded as a multi-layered process emerging from the complex interplay of free association and higher cognitive functions (1, 2, 3, 4, 5). Various neuroimaging techniques are being deployed to study what has been termed *post-retrieval* processes, a set of processes aimed at monitoring and evaluating each retrieved memory (6, 7, 8, 9, 10). Evidence suggests a wide participation 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 hypothesized to support a post-retrieval selection mechanism (10) with selection demands increasing when task-irrelevant representations are prepotent; electrodes implanted in VLPFC (5) were found to increase their activity a few hundred milliseconds after the onset of the recall stage, consistent with a top-down post-retrieval feedback. The general breadth and organization of post-retrieval processes, however, remains unknown.

An ideal setting for addressing the question is the experimental paradigm known as *free recall* (15, 16, 17, 18, 19), which has been providing insight on all facets of human memory for over a century (20, 21). Free-recall experiments are split into a presentation stage and a recall stage; the former consists in the presentation of a list of items (often words), the latter 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 from the list. (For a comprehensive monograph on both experiments and theories see (22)).

Results on free recall accumulated in the literature may indeed be subsumed under two classes: results whose broad understanding does not seem to require invoking post-retrieval mechanisms; and results that are, more or less implicitly, underpinned by the existence of post-retrieval activity.

Among the former group, the most studied effects concern the *serial position* of recalled items, i.e. their order within the list presented. These include the *recency* and *primacy* effects (16), i.e., the tendency to more consistently recall items respectively from the beginning and end of the list. The difference in serial position between two consecutively recalled items is known as "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 presented list (24, 25). The additional tendency to recall in forward order ("forward asymmetry") makes lag $L = +1$ the most likely transition, which we will refer to as "sequential" (26).

A behavioral observable that is relevant to the latter group is the time that elapses between consecutive recalls (inter-recall interval, hereafter IRI) (27, 28, 29, 30). During the recall stage of a typical trial, the first few recalled items are reported rather quickly; in contrast, towards the end of the trial the IRIs drastically increase (31, 32). 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 (33, 34, 32). 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 are drawn, they are simply discarded. Because towards the end of a trial there are many previously recalled (hence non-admissible) 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 post-retrieval sampling takes place during the recall phase and that only some of the sampled (retrieved) items are reported.

The notion that retrievals are sometimes discarded rather than reported plays a crucial role also in advanced models based on attractor neural networks, which aspire to reproduce mechanistic nonlinear neural dynamics (35, 36, 37). These models were designed to calculate testable results of free association (defined as association between the latest previously recalled item and the next item to be retrieved), and in their simplest form they assume that a recall process never ends, but only reaches a state of perpetual self-repetition. This would be of course incompatible with data without including a subsequent top-down termination; moreover, not every repeated retrieval has the effect of confining the mental trajectory to such loops. This creates the need for a higher module that chooses surrender not at the first repetitious retrieval, but after the realization of having entered those loops. The nature of the higher module can stay unspecified, if only for the sake of analyzing mathematically the network, which leads to predicting the power law of recall performance with list length (38, 39) with extraordinary accuracy (40).

Another topic that has been a focus of a large amount of attention is the potential adoption of various retrieval strategies (1, 41, 4, 42). 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 often only in a fraction of trials (41). 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 (43). While some recall strategies are related to chunking, it is long since known that a particularly advantageous strategy consists in waiving the *freedom*'s perk of free recall and instead retrieve the items in their serial order (42). Sticking to such strategies should involve discarding any retrieval that does not fit the prescribed recall sequence, but it is not known to what extent this occurs in practice.

In spite of multiple models pointing toward the necessary presence of discarded retrievals, to our knowledge there has been no attempt to study them in a targeted fashion, nor to quantify their prominence in any specific setting. In fact, most studies to date have focused on what the participants report, bypassing any possible filter that intervenes between retrieval and outputting. It might seem indeed that discarded retrievals cannot be probed without explicitly demanding a report of subjective experience after each trial, a procedure obviously tricky and prone to inaccuracies.

In this study, we will show that a careful but straightforward analysis of both the history of reported recalls and the recorded IRIs can allow to uncover hidden retrievals that are suppressed by the participants. Most of the inferred instances of discarded retrievals are related to the avoidance of repetitions of

already-recalled items, but our analysis strongly suggests that some of the participants suppress even non-recalled items in order to adhere to their retrieval strategy. Besides providing a new methodology for behavioral data from memory experiments, our findings shed a novel light on post-retrieval mechanisms as a potentially ubiquitous ingredient of human memory.

Results

Two free recall phenomena

The data we considered, collected by the laboratory of Prof. Kahana at the university of Pennsylvania, concern experiments performed with lists of $N = 16$ words (see Materials and Methods for details). The dataset displayed standard serial-position effects (with recency and primacy emerging as in Fig. S1).

We begin by pointing out two general facts in this dataset:

a) Although repetitions are not expressly forbidden and are duly recorded by the experimenter, the number of repetition-free trials is about twice the chance level set by a fitted Markovian transition model (1A). This 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 and its caption for the description of the Markovian transition model).

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 (1B), not only higher trial sequentiality corresponds to higher performance but this increasing curve appears to have a distinctive form, with an inflection point in mid-performances.

We know of no work in the psychology literature hinting that that these two features can be explained by free association alone. If one reaches for the simplest explanations outside free association, those two seemingly unrelated facts appear to share some key features:

- 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 – respectively (a) repetitious retrievals and (b) non-sequential retrievals.

- 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. We will adopt the term "repetition bouncing" for the novelty-screening of free-association retrievals (i.e. the active avoidance of free association steps

that would yield repetition) and will use the term "strategic bouncing" for the sequentiality-testing of free-association retrievals (i.e. the enforced avoidance of free association steps that would not obey a sequential ordering of recalls).

In the rest of the paper, we report on our detailed 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 dataset 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 mathematical theory that represents free recall without bouncing as a Markov process and bouncing as a non-markovian add-on; after fitting the corresponding parameters we will compare the predictions of the theory to observed features of the data.

Silent serial-position effects

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 reported recalls. Yet, it may produced temporal delays in recall transitions where the process is activated. In both panels of Fig. 2, for example, the word "ear" will be recalled slower because first the word "toy" was retrieved and discarded. A natural approach is thus to hunt for traces of those delays in the recorded IRIs. And to do so, the first obligatory step is an a priori analysis of the main potential confounders.

Following convention (22) we will call "output position" the ordered position of an item within the sequence of recalls reported during the recall stage. 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. 2A, the word "sun" 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 (32) that the distribution of IRIs collapses if regarded as a function of negative output position. This is indeed apparent in the dataset 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 (the recall in output position $= +1$) is a latency time (during which, moreover, no repetition bouncing can occur) 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. 3B). After recall of an item, its repetitious retrieval becomes possible. If such a retrieval comes to mind, it would be associated to a positive time delay. In other words, systematic *bouncing* against the item in a given serial position will be expected to make the post-recall IRI larger than the pre-recall IRI for every output positions (see Materials and Methods for details

of the analysis). This turns out to be true for the last item in the list regardless of the output position. If the final item has already been recalled, all else being equal, a delay is found to occur in the recall process.

The fact that the final item appears to be a preferential bouncing target is not a surprise in retrospect. It has long been documented in the free recall literature (see Introduction) that the final items in a list are the most likely to be included in the output ("recency effect"). The delay associated to potential repetition of the final item can thus be thought of as a "silent recency effect"; if the attraction of the most recent memory from the list is the strongest, it can stay such after that memory has been recalled and we expect that 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 different ordering in which recency and primacy appear in the recall process (Fig. S1B) as the early recall of an item makes it the 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 or not the repetition screening. The mean IRI in these two conditions can again be compared, to check for a delay associated to the possible discard of such retrievals (technical aspects in Materials and Methods). As was mentioned in the Introduction, a widely ascertained property of reported recalls is that 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 to the largest delay.

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

A model for repetition bouncing

To further test the bouncing hypothesis it is necessary to possess a forward model that describes it in a predictive fashion. We proceed to build a streamlined theory that formalizes bouncing as applied to repetition avoidance (Fig. 2A).

Models of free recall have covered a spectrum ranging from dual-store memory search models (44, 3) to powerful theories of temporal context dependence (45). 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 (see Introduction); our focus is entirely on the concept of bouncing and 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:

- 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.
- 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 recalls.

- every 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 later introduced when we study strategic bouncing).

We have explored alternative assumptions for both the termination mechanism (static sink states, abrupt termination thresholds) and for the non-markovian add-on (one example is analyzed in detail in Supplementary Note 2), and they have proven less predictive than the model we are presenting.

The non-markovian contribution from bouncing to the recall process 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 simple masking of the transition matrix.

The corresponding loglikelihood as well as its gradient (Supplementary Note 1) are easily written down and applied to fitting the dataset. This 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. 5A). The number of bounces underlying every recall event can then be calculated by simulating the model with these parameters (for sample trials see Fig. 5B, while the fitting and simulating protocols are documented in Materials and Methods).

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 6).

We then test for the silent serial-position effects along the same lines followed for real data (Figs. 3-4), 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 (Fig. 7) and silent contiguity (Fig. 8).

Sequential slowdown

We turn then to the distinctive shape of the sequentiality/performance curve (Fig. 1B). We hypothesized that it might emerge if a partial discarding of free-association retrievals allows the implementation, in some trials, of a more markedly sequential recall strategy associated to 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 highly performing trials, Fig. S4).

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.

To differentiate between strategic and non-strategic trials we can use the fact that the former have by construction a 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 (41), the effect we are seeking may only be present in the more highly 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 of IRIs for sequential and non-sequential recalls within each trial. Sequential recalls are generally faster than non-sequential ones across different performance levels and output positions (Fig. S5A, except at high-performance).

For every condition of trial performance and output position, we thus consider the distribution of trial sequentiality among trials that contain a sequential recall in that performance level and output position (Fig. 9A). We then split the sample by the median value of trial sequentiality, and compare the IRIs of recall events in the two subsamples (Fig. 9B). We only compare sequential recall events because highly sequential trials contain few non-sequential recalls. Thresholding out all sample pairs that do not pass a significance test, we calculate the sign of the delay associated to 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 to sequential events at each given output position.)

We thus find that (see Fig. 9C):

- In all cases where our significance requirement is met, the delay is positive, meaning that sequential recalls occurred in sequentially biased trials have taken longer. In the following we refer to this phenomenon, not previously reported to our knowledge, as *sequential slowdown*.

- 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 non-sequential retrievals are provided by free association, which is what we termed strategic bouncing.

Sequential slowdown in the bouncing model

With model parameters coming from the above-mentioned fit, we only add into the model a finite probability s ("strategicity") of bouncing back from a retrieved item if it is non-sequential. We are not necessarily assuming a common mechanism for strategic bouncing and repetition bouncing (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 to strategic bounces. This proves to be the sensible choice from the requirement that the performance should increase as a function of strategicity (Fig. 10A).

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. 10B).

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. 1B). Rather than aiming at numerical accuracy we seek to understand the shape of the curve, namely its inflexion in mid-performances, followed by a drastic rise for larger performances. First we note that repetition bounces alone don't explain this feature (Fig. 10B). 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 small percentage of strategic trials is sufficient to inflect the curve (Fig. 10B).

Finally, we agnostically interrogate the mixed-set pseudodata model for the signatures of strategic slowdown we encountered in Fig. 9C. 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. 10D), signifying a delay in time as found in real data. In Fig. 9C, we also showed that the time delay found in real data grows into significance in the high performance region for sufficiently large output positions. The model would be agreement with real data if, 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 prediction is finally confirmed by the result of Fig. 10D, further supporting the notion that sequential slowdown is due to strategic bouncing.

Discussion

While free recall experiments are an essential behavioral probe in the study of human memory, much research has revolved around standard serial-position related effects (see Introduction). In this work, we undertook the rigorous mining of a large dataset to uncover a number of hitherto unnoticed features that substantially enlarge the current phenomenology:

- the number of repetitions observed in these experiments is vastly smaller than what a basic expectation given by chance Markovian transitions;
- recalls are delayed if the final item has already been recalled ("silent recency effect");
- recalls are delayed if sequential recall would yield a repetition ("silent contiguity effect");
- the overall abundance of sequential recalls in the trials has a distinctively inflected shape when regarded as a function of trial performance;
- a slowdown in certain recall events is associated to the overall prevalence of sequential recalls in the trial ("sequential slowdown").

In addition to reporting on the above, we proceeded to suggest a unifying explanation based on the notion that free-association retrieval are "bounced", i.e. discarded, in a majority of cases 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 model with trajectories driven by heterogeneous transition probabilities and reset by bouncing. Bouncing occurs systematically for repetitious retrievals and adds to the "impatience" that, probabilistically, leads to termination of a recall trial; bouncing of nonsequential transitions occurs in strategic trials with a fixed probability ("strategicity").

Fitting these parameters on 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. Our work strongly suggests the importance of unreported retrievals in determining behavioral observables of memory experiments.

It is easy to see indeed 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. If waiting accumulates impatience, that would make some trials both slower and more highly performing, without involving any retrieval rejection.

However, this would rather preferentially slow down non-sequential trials that have a higher risk of surrendering, in contradiction to data. Moreover, such explanation can be ruled out also because the sequential slowdown is not preferentially observed in the final recall of each trial, where surrendering is most prominent (Fig. 9C). In fact, a significant delay associated to high trial

sequentiality is not observed for the last output position of any performance group.

Neither our analysis nor our models have relied on specific assumptions on the source of the post-retrieval control, although active blocking of internally recalled items may be conjectured to emerge from higher-level processes that operate top-down over free association (see Introduction). However, our results can be taken rather as a starting point to tackling that question.

Measurements performed in free recall are not limited to behavioral observables and numerous studies have been performed using noninvasive neuroimaging techniques such as EEG, MEG, PET, fMRI ((5, 12, 46, 47, 48)) as well as invasive methods in epilepsy patients (49, 50). A parallel study of unreported retrievals could be attempted through the decoding of suitable neurophysiological data. The timing and magnitude of post-retrieval effects such as those already observed in ECoG of prefrontal cortex (5) could be compared to the timing and quantity of the bounces inferred by an analysis of behavioral data.

The issue whether bouncing is consciously perceived by the participants is more delicate, but can also be studied in a dedicated experimental setting, with the caution of providing separate controls as for the accuracy and consistency of self-reporting.

Our statistical inquiry on the IRIs dove to an unprecedented level of detail, yet it relied exclusively on mean values of the IRI over different subsamples. The full density profiles of subsampled IRI distributions contain a wealth of unexplored information that deserves to be the subject of a separate study.

In this work we focused on the two types of bounces – repetition bouncing and strategic bouncing to maintain sequentiality – merely because these two were compellingly suggested by Fig. 1. In a different dataset on human recall the criteria that drive post-selection will accordingly be different but are bound to leave detectable traces in the IRI statistics, specifically with tasks where some recalls are to be expressly excluded. For example in (8) subjects are instructed to recall items not belonging to a specified categories or not beginning with certain letters. While other types of bouncing might take place, their detection can be attempted by the very means we introduced.

In terms of complexity, our modeling of free association is situated halfway between the extremes of the pure-death model discussed in the Introduction (incapable of modeling contiguity effects, see (33)) and sophisticated models of free-associating neural networks such as (35). However, the features uncovered by our model-free analysis of data should be recoverable with little adjustment from a variety of models, notably from the standard Retrieved Context Model of episodic memory (2).

Finally, on the side of mechanistic neural dynamics, what remains to be understood is how the top-down post-retrieval actions that we described as bouncing is effectively implemented by a neuronal network. A natural candidate would be inhibitory control (response inhibition), and testing its implementation on a neuronal network with overlapping attractors will be a further step toward unraveling the biological substrate for post-retrieval operations.

Within the concept of free will, which has considerable social and legal im-

portance (51), *freedom of choice* separately requires the awareness of multiple options and the ability to discard them. Both components enter our definition of bouncing, which may warrant possible hybridizations of the free-recall paradigm with the class of neurophysiological tests devised for free will (52, 53, 54).

Our data mining and modeling approach, in conclusion, has shed light on unreported cognitive processes behind memory recall, and opens a new avenue for their exploration with modern recording technologies.

Materials and Methods

Data

We base our analysis on a publicly available dataset, collected as part the Penn Electrophysiology of Encoding and Retrieval Study at the University of Pennsylvania (for a full account see (30)). Participants consented according the University of Pennsylvania's IRB protocol and were compensated for their participation.

Each word was drawn from a pool of 1,638 words. Lists were constructed such that varying degrees of semantic relatedness occurred at both adjacent and distant serial positions in a way that randomized semantic effects.

For each list, there was a delay of 1.5 seconds before the first word appeared on the screen. Each item was on the screen for 3 seconds, followed by a variable time gap lasting between 8 and 12 seconds (uniform distribution). The recall stage began after 12–14 seconds from presentation of the last item.

Participant were given 75 seconds 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 (27198 in total) were used. In the minority of trials containing some repetition (Fig. S2A), repetitions were ignored for the main analyses. Intrusions from outside the presented lists were all removed at the outset.

Statistical Analyses

The set of potentially unreported retrievals was characterized by either the serial position of the retrieved item (Fig. 3) or the lag from the latest recall to the item (Fig. 4). In both cases, to make the analysis of IRIs rigorous we took the following precautions:

- We broke down the sampling of recall events by negative output position to discount the strong output position dependence highlighted in Fig. S3.
- We confronted pre- and post-recall conditions by comparing only the IRIs for the recall of items outside the putatively unreported set. Reported

recalls of an item from that set can appear in the pre-recall sample but not in the post-recall sample, so we excluded them from both.

- 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 comparison 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. 3A), rather than using a t-test we adopted a Mann-Whitney u-test (55).

For example, Fig. S5A 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.

Model Fitting

Let $\hat{\pi}$ be the naked transition matrix (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 y and, for any set S of serial positions, let $\pi(S|x) = \sum_{y \in S} \pi(y|x)$.

Further naming N the length of the presented lists, T the number of trials in the sample, m_α the performance of trial α, x_k^α the serial position of the k-th recall in the trial, q the impatience parameter, and using the notation where $x_{i:j}$ is the set of serial position recalled in between output positions i and j, both included, the normalized loglikelihood of the bouncing model is

$$\begin{aligned} L\big[\hat{\pi}, q\big] \quad &= \quad \ln q + \frac{1}{T}\sum_{\alpha=1}^{T} \ln \pi\big(x_{1:m_\alpha-1}|x_{m_\alpha}\big) + \\ &+ \frac{1}{T}\sum_{\alpha=1}^{T}\sum_{n=1}^{m_\alpha-1} \ln \pi(x_{n+1}|x_n) - \frac{1}{T}\sum_{\alpha=1}^{T}\sum_{n=1}^{m_\alpha} \ln\big[1-(1-q)\pi(x_{1:n-1}|x_n)\big] \end{aligned}$$

For the derivation, Supplementary Note 1.

We write down the gradient (Supplementary Note 1) and search for the minimum from a random initial condition using the L-BFGS-B algorithm (56). This converges in $< 10^2$ iterations to the matrix shown in Fig. 5A and to the optimal impatience $q = 0.1$.

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 non-strategic sequentiality.

Simulations

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

Fitting and simulations based on an alternative "skipping model" are described in Supplementary Note 2.

References

(1) Scott D Gronlund and Richard M Shiffrin. "Retrieval strategies in recall of natural categories and categorized lists." In: Journal of Experimental Psychology: Learning, Memory, and Cognition 12.4 (1986), p. 550.

(2) Sean M Polyn, Kenneth A Norman, and Michael J Kahana. "A context maintenance and retrieval model of organizational processes in free recall." In: Psychological review 116.1 (2009), p. 129.

(3) Jeroen G Raaijmakers and Richard M Shiffrin. "Search of associative memory." In: Psychological review 88.2 (1981), p. 93.

(4) Suzanna Becker and Jean Lim. "A computational model of prefrontal control in free recall: strategic memory use in the California Verbal Learning Task". In: Journal of Cognitive Neuroscience 15.6 (2003), pp. 821–832.

(5) Yitzhak Norman et al. "Neuronal baseline shifts underlying boundary setting during free recall". In: Nature communications 8.1 (2017), pp. 1–17.

(6) Hiroki R Hayama, Jeffrey D Johnson, and Michael D Rugg. "The relationship between the right frontal old/new ERP effect and post-retrieval monitoring: specific or non-specific?" In: Neuropsychologia 46.5 (2008), pp. 1211–1223.

(7) Hiroki R Hayama and Michael D Rugg. "Right dorsolateral prefrontal cortex is engaged during post-retrieval processing of both episodic and semantic information". In: Neuropsychologia 47.12 (2009), pp. 2409–2416.

(8) Henry L Roediger III and Endel Tulving. "Exclusion of learned material from recall as a postretrieval operation". In: Journal of Verbal Learning and Verbal Behavior 18.5 (1979), pp. 601–615.

(9) Michael D Rugg, Jeffrey D Johnson, and Melina R Uncapher. "Encoding and retrieval in episodic memory". In: The Wiley handbook on the cognitive neuroscience of memory (2015), pp. 84–107.

(10) David Badre et al. "Dissociable controlled retrieval and generalized selection mechanisms in ventrolateral prefrontal cortex". In: Neuron 47.6 (2005), pp. 907–918.

(11) Nicole M Long, Ilke Öztekin, and David Badre. "Separable prefrontal cortex contributions to free recall". In: Journal of Neuroscience 30.33 (2010), pp. 10967–10976.

(12) Cary R Savage et al. "Prefrontal regions supporting spontaneous and directed application of verbal learning strategies: evidence from PET". In: Brain 124.1 (2001), pp. 219–231.

(13) Donald T Stuss et al. "Organizational strategies with unilateral or bilateral frontal lobe injury in word learning tasks." In: Neuropsychology 8.3 (1994), p. 355.

(14) Roberto Cabeza et al. "The parietal cortex and episodic memory: an attentional account". In: Nature Reviews Neuroscience 9.8 (2008), pp. 613–625.

(15) Hermann Ebbinghaus. "Memory: a contribution to experimental psychology. 1885". In: New York: Teachers College, Columbia University (1913).

(16) Bennet B Murdock Jr. "The serial position effect of free recall." In: Journal of experimental psychology 64.5 (1962), p. 482.

(17) Endel Tulving. "Subjective organization in free recall of" unrelated" words." In: Psychological review 69.4 (1962), p. 344.

(18) Leo Postman and Laura W Phillips. "Short-term temporal changes in free recall". In: Quarterly journal of experimental psychology 17.2 (1965), pp. 132–138.

(19) Aldridge K Bousfield and Weston A Bousfield. "Measurement of clustering and of sequential constancies in repeated free recall". In: Psychological Reports 19.3 (1966), pp. 935–942.

(20) Alfred Binet and Victor Henri. "Mémoire des mots". In: L'Année psychologique 1.1 (1894), pp. 1–23.

(21) Edwin A Kirkpatrick. "An experimental study of memory." In: Psychological Review 1.6 (1894), p. 602.

(22) Michael Jacob Kahana. Foundations of human memory. OUP USA, 2012.

(23) Michael J Kahana. "Associative retrieval processes in free recall". In: Memory & cognition 24.1 (1996), pp. 103–109.

(24) Marc W Howard and Michael J Kahana. "Contextual variability and serial position effects in free recall." In: Journal of Experimental Psychology: Learning, Memory, and Cognition 25.4 (1999), p. 923.

(25) M Karl Healey, Nicole M Long, and Michael J Kahana. "Contiguity in episodic memory". In: Psychonomic bulletin & review 26.3 (2019), pp. 699–720.

(26) Michael J Kahana and Jeremy B Caplan. "Associative asymmetry in probed recall of serial lists". In: Memory & Cognition 30.6 (2002), pp. 841–849.

(27) Howard R Pollio, Richard A Kasschau, and Harry E Denise. "Associative structure and the temporal characteristics of free recall." In: Journal of Experimental Psychology 76.2p1 (1968), p. 190.

(28) Howard R Pollio, Steven Richards, and Richard Lucas. "Temporal properties of category recall". In: Journal of Verbal Learning and Verbal Behavior 8.4 (1969), pp. 529–536.

(29) Karalyn E Patterson, Richard H Meltzer, and George Mandler. "Inter-response times in categorized free recall". In: Journal of Verbal Learning and Verbal Behavior 10.4 (1971), pp. 417–426.

(30) M Karl Healey, Patrick Crutchley, and Michael J Kahana. "Individual differences in memory search and their relation to intelligence." In: Journal of Experimental Psychology: General 143.4 (2014), p. 1553.

(31) Bennet B Murdock and Ronald Okada. "Interresponse times in single-trial free recall." In: Journal of Experimental Psychology 86.2 (1970), p. 263.

(32) Doug Rohrer and John T Wixted. "An analysis of latency and interresponse time in free recall". In: Memory & Cognition 22.5 (1994), pp. 511–524.

(33) William J McGill. "Stochastic latency mechanisms". In: Handbook of mathematical psychology 1 (1963), pp. 309–360.

(34) Dirk Vorberg and Rolf Ulrich. "Random search with unequal search rates: Serial and parallel generalizations of McGill's model". In: Journal of Mathematical Psychology 31.1 (1987), pp. 1–23.

(35) Sandro Romani et al. "Scaling laws of associative memory retrieval". In: Neural computation 25.10 (2013), pp. 2523–2544.

(36) Mikhail Katkov, Sandro Romani, and Michail Tsodyks. "Memory retrieval from first principles". In: Neuron 94.5 (2017), pp. 1027–1032.

(37) Michelangelo Naim et al. "Emergence of hierarchical organization in memory for random material". In: Scientific reports 9.1 (2019), pp. 1–10.

(38) Bennet B Murdock Jr. "The immediate retention of unrelated words." In: Journal of Experimental Psychology 60.4 (1960), p. 222.

(39) DJ Murray, Carol Pye, and WE Hockley. "Standing's power function in long-term memory". In: Psychological Research 38.4 (1976), pp. 319–331.

(40) Michelangelo Naim et al. "Fundamental law of memory recall". In: Physical Review Letters 124.1 (2020), p. 018101.

(41) Sandro Romani, Mikhail Katkov, and Misha Tsodyks. "Practice makes perfect in memory recall". In: Learning & Memory 23.4 (2016), pp. 169–173.

(42) Rosamond Gianutsos. "The order of recall and the recall of order". In: Memory & cognition 4.5 (1976), pp. 627–636.

(43) Geoffrey Keppel, Leo Postman, and Bonnie Zavortink. "Studies of learning to learn: VIII: the influence of massive amounts of training upon the learning and retention of paired-associate lists". In: Journal of Verbal Learning and Verbal Behavior 7.4 (1968), pp. 790–796.

(44) Jeroen GW Raaijmakers, Richard M Shiffrin, et al. "SAM: A theory of probabilistic search of associative memory". In: The psychology of learning and motivation: Advances in research and theory 14 (1980), pp. 207–262.

(45) Marc W Howard, Michael J Kahana, et al. "A distributed representation of temporal context". In: Journal of Mathematical Psychology 46.3 (2002), pp. 269–299.

(46) Stefanie Brassen et al. "Hippocampal–prefrontal encoding activation predicts whether words can be successfully recalled or only recognized". In: Behavioural brain research 171.2 (2006), pp. 271–278.

(47) Bernhard P Staresina and Lila Davachi. "Differential encoding mechanisms for subsequent associative recognition and free recall". In: Journal of Neuroscience 26.36 (2006), pp. 9162–9172.

(48) Brad C Dickerson et al. "Prefrontal-hippocampal-fusiform activity during encoding predicts intraindividual differences in free recall ability: An event-related functional-anatomic MRI study". In: Hippocampus 17.11 (2007), pp. 1060–1070.

(49) Ethan A Solomon et al. "Hippocampal theta codes for distances in semantic and temporal spaces". In: Proceedings of the National Academy of Sciences 116.48 (2019), pp. 24343–24352.

(50) Hagar Gelbard-Sagiv et al. "Internally generated reactivation of single neurons in human hippocampus during free recall". In: Science 322.5898 (2008), pp. 96–101.

(51) Susan Pockett. "The concept of free will: Philosophy, neuroscience and the law". In: Behavioral Sciences & the Law 25.2 (2007), pp. 281–293.

(52) Aaron Schurger, Jacobo D Sitt, and Stanislas Dehaene. "An accumulator model for spontaneous neural activity prior to self-initiated movement". In: Proceedings of the National Academy of Sciences 109.42 (2012), E2904–E2913.

(53) Benjamin Libet et al. "Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential)". In: Neurophysiology of consciousness. Springer, 1993, pp. 249–268.

(54) JD Runyan. "Libet-Style Experiments and Volitions". In: Human Agency and Neural Causes. Springer, 2014, pp. 16–44.

(55) Thomas P Hettmansperger and Joseph W McKean. Robust nonparametric statistical methods. CRC Press, 2010.

(56) Richard H Byrd et al. "A limited memory algorithm for bound constrained optimization". In: SIAM Journal on scientific computing 16.5 (1995), pp. 1190–1208.

Acknowledgments

Funding: The study was supported by RIKEN Center for Brain Science, Brain/MINDS from AMED under Grant No. JP20dm020700, and JSPS KAKENHI Grant No. JP18H05432. **Author contributions:** The authors worked as a team; credit is not individualizable. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data are publicly available at the address `http://memory.psych.upenn.edu/files/pubs/HealEtal14.data.tgz`. The Python library we developed for free-recall datasets is available upon reasonable request.

Figures

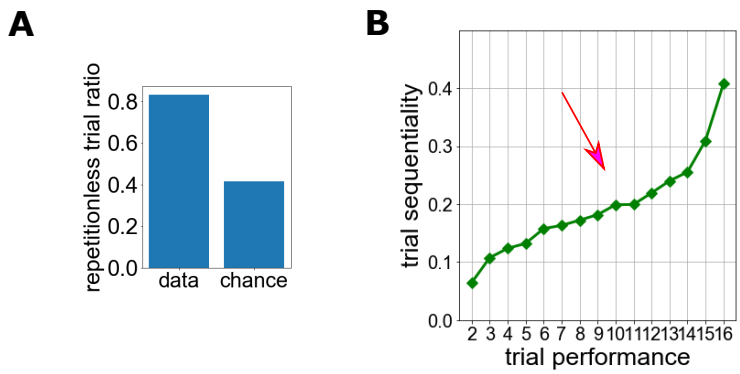

Fig. 1. Two unexplained features of free recall. **(A)** Paucity of observed output repetitions as compared to chance level (for details of chance level estimation see Fig. S2). **(B)** Fraction of sequential recalls in a trial as a function of trial performance. Mean curve is shown, while the standard error of the mean is too small to plot. Red arrow highlights approximate location of an inflection point, where the slope starts to increase.

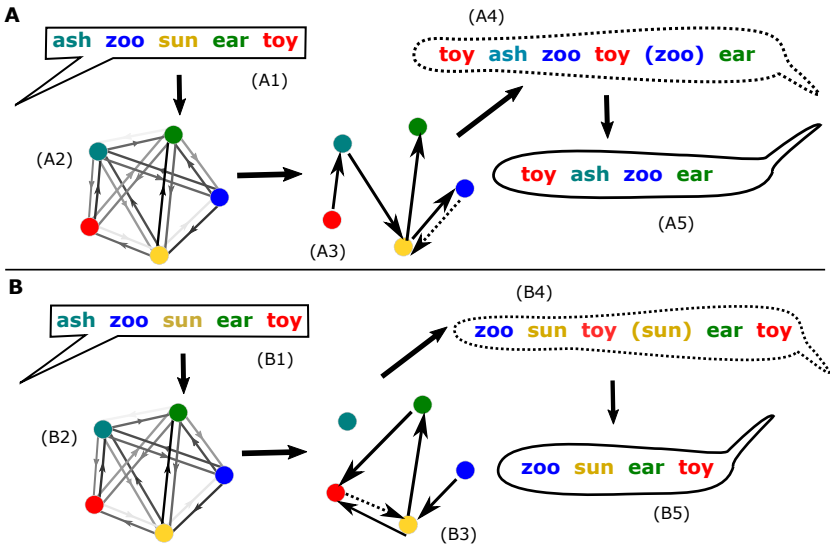

Fig. 2. Schematic depiction of bouncing. Intervention over the free association process by "bouncing" simultaneously explains paucity of observed repetitions and higher sequentiality of high-performance trials. Here, a five-item list is presented in two separate trials leading respectively to an instance of *repetition bouncing* (**A**) or one of *strategic bouncing* (**B**). During the presentation stage of free recall (A1,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,B2) by the shades of gray of the arrows. During the recall stage (A3), the subject instantiates a mental trajectory through those memories(A3,B3). Nothing prevents the sequence of retrieved memories from containing an unlimited number of repetitions (as for example in A3,A4), but the output is devoid of repetitions (A5), posing the problem whether some retrievals are discarded. In addition, a tendency toward sequentiality in high-performing trials is also observed (B5), posing the problem whether non-sequential transitions have been actively avoided (as when stepping back from "toy" to "sun" in B3,B4). Bouncing transitions are displayed as dotted arrows.

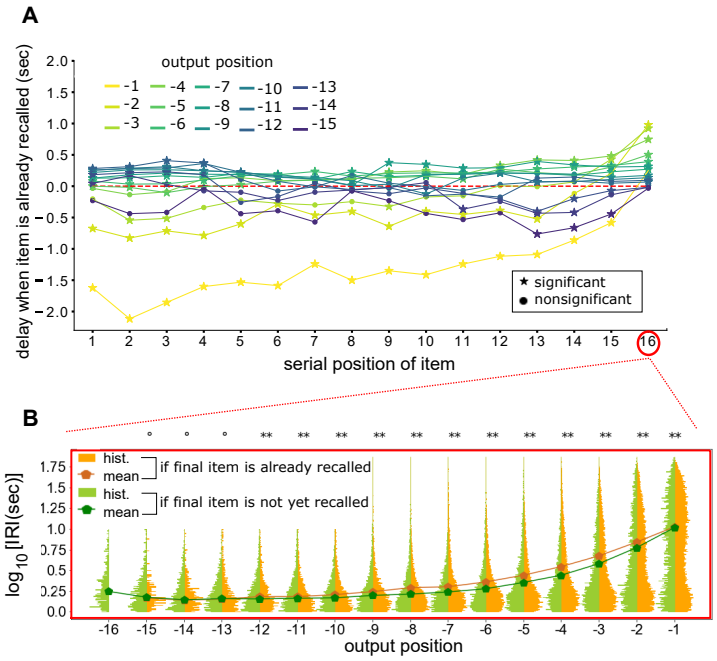

Fig. 3. Empirical evidence for a higher-order "silent recency effect" in free recall. **(A)** Difference between mean IRI when subsampling on the basis of whether the item with a given serial position on the list has already been recalled or not. IRI were histogrammed excluding recalls of the item with 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 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 final one, plotted separately for each output position and for the two conditions were the final item has already been recalled or not. The scale of the IRIs is logarithmic. 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, $10^{-4} > p$. The previously occurred recall of the final item is statistically associated to a delay in the recall of other items.

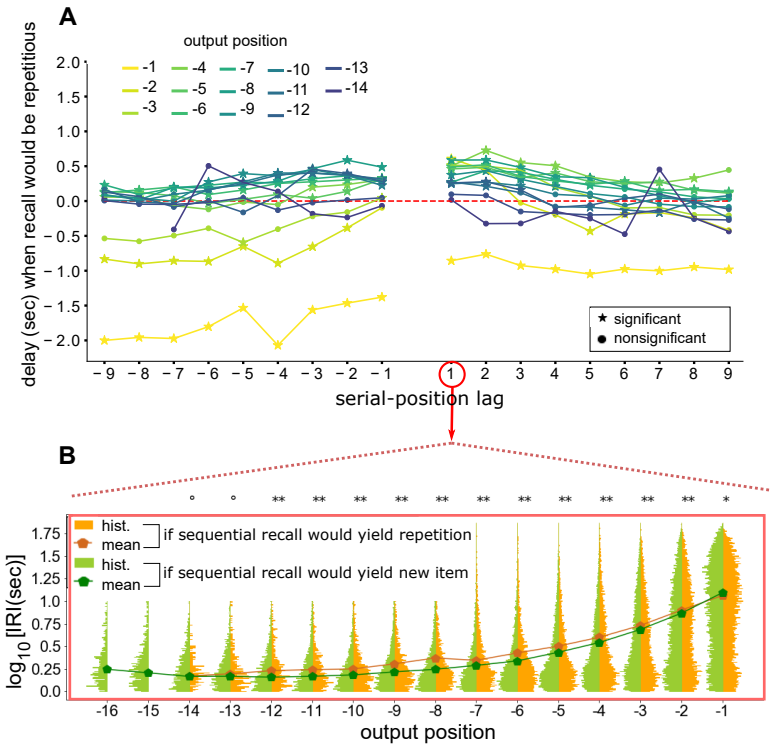

Fig. 4. Empirical evidence for a "silent contiguity effect". **(A)**: Difference between mean IRI when subsampling on the basis of whether the item situated at a certain fixed lag from the last recall has already been recalled or not. The lag position concerned is shown on the x-axis, and plot markers 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 were the sequential item has already been recalled or not. The scale of the IRIs is logarithmic. 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, $10^{-4} > p$. The previously occurred recall of the sequential item is statistically associated to a delay in the recall of other items.

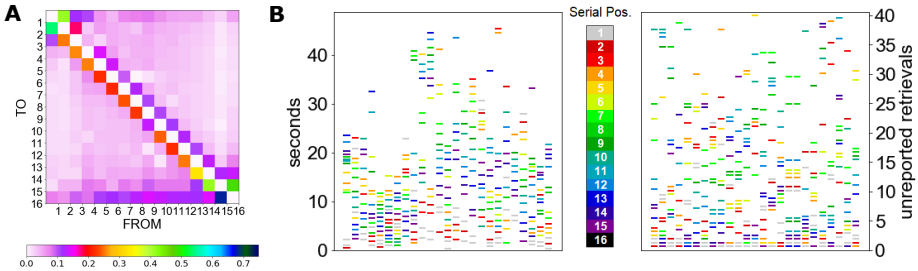

Fig. 5. Illustration of typical features of the bouncing model. **(A)** The optimal transition matrix obtained from fitting the data with the optimal impatience $q = 0.1$. **(B)** Example trials from the data (Left) and from simulations of the bouncing model (Right). Color encodes serial position of recalled items. The y-coordinate in the data panel is the timing of the recall event. The y-coordinate in the model panel is the number of unreported retrievals observed up to a given recall in simulations of the model, which has been jittered to prevent overlaps for visual aid.

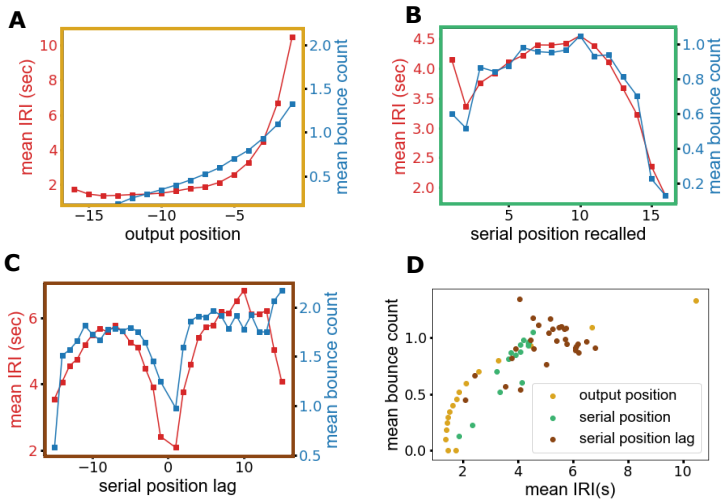

Fig. 6. Comparison between mean IRIs and the mean number of unreported retrievals of the model. Result of averaging is shown for fixed values of the recall event's output position **(A)**, the serial position of the recalled item **(B)**, and the serial position lag from the preceding recall **(C)**. In **(D)**, values shown in the previous panels have been scatter-plotted together.

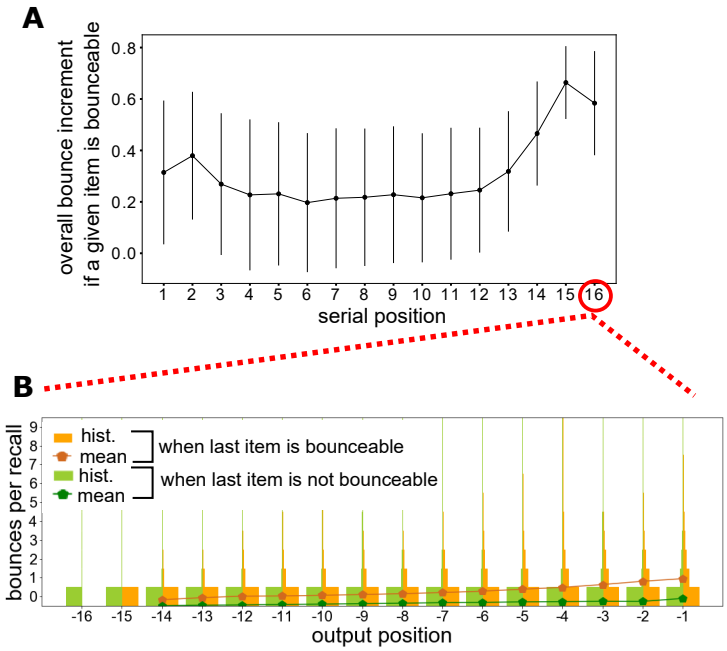

Fig. 7. Silent recency effect emerging from the bouncing model. (A) Selective variations in the mean bouncing counts estimated from model simulations. The mean number of bounces within a recall event has been averaged in the two subsamples of recalls where the list item in a given serial position is 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.

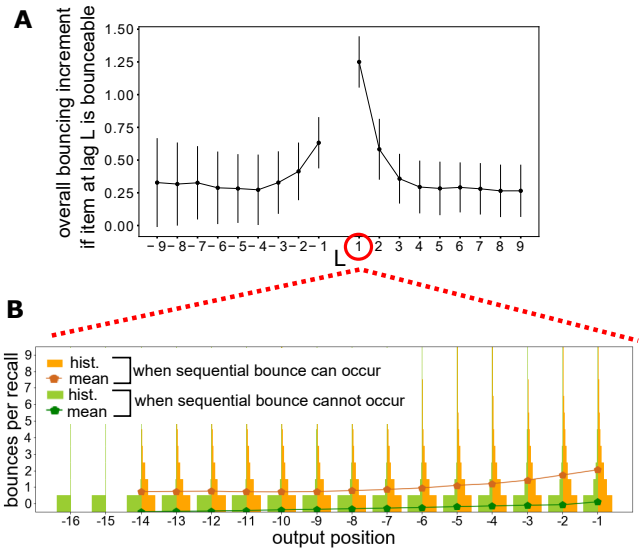

Fig. 8. Silent contiguity effect emerging from bouncing model. (A) Selective variations in the mean bouncing count per recall estimated from model simulations. The mean number of bounces within a recall event has been averaged in the two subsamples of recall events where a given lag does or does not lead to a bounceable item, and the difference between the two means is plotted as a function of the given lag. **(B)** 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.

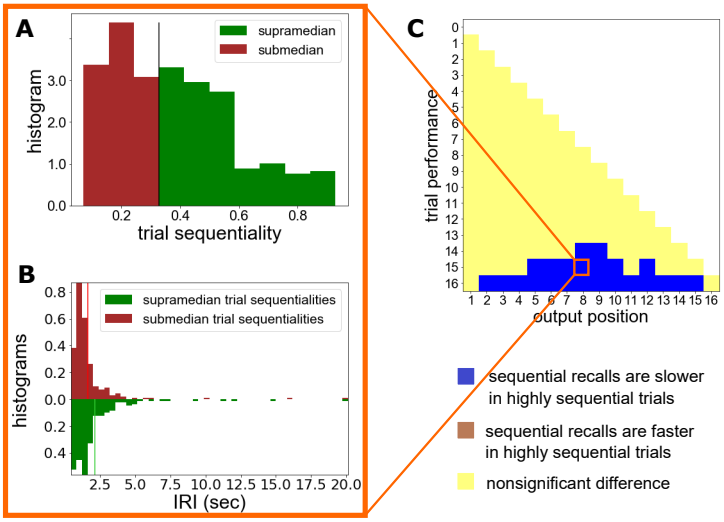

Fig. 9. "Sequential slowdown" in free-recall data. For each performance level, we list all trials that perform a sequential transition at a given output position and study how its speed changes in highly sequential trials. The sequentiality value associated to each trial is simply the fraction of sequential recalls in the trial. **(A)** The histogram of this quantity for a particular subsample 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 indicating the respective mean values. **(C)** Sign of the difference in the mean values whenever it 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; the opposite occurrence would be submedian trials performing significantly slower on sequential recalls but, strikingly, this is never observed.

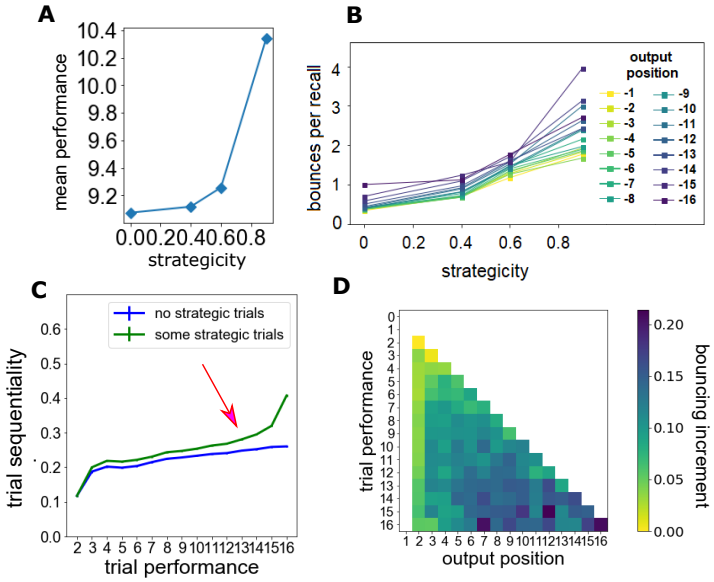

Fig. 10. Simulations of the bouncing model with strategizing. We simulated the model by adding in a given probability of strategic bouncing (i.e. the probability of rejecting a non-sequential retrieval), which we termed "strategicity". **(A)** Increasing strategicity also increases performance, coherently 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. 1B 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. 9D).

List of Supplementary Materials

- Supplementary Note 1. Loglikelihood of the bouncing model
- Supplementary Note 2. The "skipping" model
- Fig. S1. Overview of recall by serial position of recalled items.
- Fig. S2. Comparison to chance level of the amount of repetitions in the dataset.
- Fig. S3. Terminal Alignment of mean IRIs.
- Fig. S4. Bimodality of the final item's output-position in high performance trials.
- Fig. S5. Dependency of recall speed on sequentiality.
- Fig. S6. Schematic representation of the skipping model.
- Fig. S7. Deep network architecture for fitting the skipping model.
- Fig. S8. Silent serial-position effects in the skipping model.
- Table S1. Main definitions used in the text.

Supplementary Note 1: Loglikelihood of the 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 loglikelihood 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, what we refer to as a "bounce". In other word 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

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

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

In other words, the transition matrix is defined with the target items in the bounceable set masked away, and every other element 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 to 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

$$\Pi(y|x_n; J_n) = \Pi_0(y|x_n; J_n)[1 - s_q(x_n; J_n)] \tag{2}$$

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

$$s_q(x, J) + b_q(x, J) + r_q(x, J) = 1$$

For a single retrieval step, their elemental values are

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

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

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

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

$$s_q(x; J) = \sum_{n=0}^{\infty} \left(b_q^{(1)}(x, J)\right)^n s^{(1)}(x, J)q = \sum_{n=0}^{\infty} \left[\pi(J|x)(1-q)\right]^n \pi(J|x)q \tag{3}$$

Merging Eqs. (2) and (3), we arrive at

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

We can now merge in Eq. 1 to eliminate Π_0, obtaining

$$\Pi(y|x; J) = \frac{\pi(y|x)\ \mathbf{1}\big[y \notin J\big]}{1 - (1-q)\pi(J|x)} \qquad \forall y \neq 0 \tag{4}$$

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 (4) for the case of transitions to the sink state. Transitions from the sink state have probability one of staying there and are therefore not problematic:

$$\Pi(y|0; J) = \mathbf{1}\left[y = 0\right] \tag{5}$$

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

$$\Pi(0|x; J) = \frac{\pi(J|x)q}{1 - (1-q)\pi(J|x)} \tag{6}$$

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

$$\mathcal{L}\big[\hat{\pi}, q, \vec{\lambda}\big] = \frac{1}{T} \sum_{\alpha=1}^{T} L\left[x_{1:N}^{\alpha}; \hat{\pi}, q\right] \tag{7}$$

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

$$x_{1:N}^{\alpha} = \{x_1^{\alpha}, \ldots, x_N^{\alpha}\}$$

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

The one-trial loglikelihood appearing in Eq. (7) is defined as

$$L\left[x_{1:N};\hat{\pi},q\right]=\sum_{n=1}^{N-1}\ln\Pi\Big(x_{n+1}|x_n;x_{1:n-1}\Big) \tag{8}$$

with Π and computed according to Eqs. (4,5,6) for for the given point in the $(\hat{\pi},q)$ parameter space.

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

$$L\big[x_{1:N};\hat{\pi},q\big]=\sum_{n=1}^{m_\alpha-1}\ln\frac{\pi(x_{n+1}|x_n)}{1-(1-q)\pi(x_{1:n-1}|x_n)}+\ln\frac{\pi(x_{1:m_\alpha-1}|x_{m_\alpha})q}{1-(1-q)\pi(x_{1:m_\alpha-1}|x_{m_\alpha})} \tag{9}$$

from which, separating the arguments of the logarithms,

$$\begin{aligned}L\big[x_{1:N};\hat{\pi},q\big] &= \ln q+\ln\pi\big(x_{1:m_\alpha-1}|x_{m_\alpha}\big)+\\ &\quad+\sum_{n=1}^{m_\alpha-1}\ln\pi(x_{n+1}|x_n)-\sum_{n=1}^{m_\alpha}\ln\big[1-(1-q)\pi(x_{1:n-1}|x_n)\big]\end{aligned}$$

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

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

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

In terms of the matrix $\hat{u}$, the one-trial loglikelihood (8) becomes

$$\begin{aligned}L\big[x_{1:N};\hat{u},q\big] &= \ln q+\ln u\big(x_{1:m_\alpha-1}|x_{m_\alpha}\big)+\sum_{n=1}^{m_\alpha-1}\ln u(x_{n+1}|x_n)\\ &\quad-\sum_{n=1}^{m_\alpha}\ln\big[u(\text{all}|x_n)-(1-q)u(x_{1:n-1}|x_n)\big]\end{aligned}$$

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

$$\mathcal{D}^q\big[\hat{u},q\big]=\frac{\partial}{\partial q}\mathcal{L}=\frac{1}{q}-\frac{1}{T}\sum_{\alpha=1}^{T}\sum_{n=1}^{m_\alpha}\frac{u(x^\alpha_{1:n-1}|x^\alpha_n)}{u(\text{all}|x^\alpha_n)-(1-q)u(x^\alpha_{1:n-1}|x^\alpha_n)}.$$

As for the u-derivatives, we have

$$\mathcal{D}^u_{yx}\big[\hat{u},q\big] = \frac{\partial}{\partial u_{yx}}\mathcal{L} = -\frac{1}{T}\sum_{\substack{(\alpha,k):\\ x^\alpha_k=x}} \frac{1}{u(\text{all}|x)-(1-q)u(x^\alpha_{1:k-1}|x)} +$$

$$+ \quad \frac{1}{T}\sum_{\substack{(\alpha,k,l):\\ x^\alpha_k=x, x^\alpha_l=y}} \left[\frac{\mathbf{1}\big(k=m_\alpha\big)}{u(x^\alpha_{1:m_\alpha-1}|x)} + \frac{\mathbf{1}\big(l=k+1\big)}{u(y|x)} + \frac{(1-q)\mathbf{1}\big(k>l\big)}{u(\text{all}|x)-(1-q)u(x^\alpha_{1:k-1}|x)}\right]$$

As a sanity check for this last formula, one can isolate the contribution from trials where $x^\alpha_N = x$ and find that it is exactly zero, as it should 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

$$\begin{aligned} s &= \log\left(\frac{1}{q}-1\right) \\ v &= \log(u) \end{aligned}$$

so that the gradient becomes

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

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

or in terms of the physical variables:

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

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

Supplementary Note 2: 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 (what will be described as a "skip" rather than a "bounce", see Fig. S6).

As in the bouncing model, a $N \times N$ 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 on 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 n-th 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. S7A).

Since the dataset samples transitions driven by the T_S matrix, the π matrix can be obtained via maximizing the likelihood of T_S Matrix given the dataset. Unfortunately, even if infinite skips are allowed in a single recall event and no sink state included, a closed formula like Eq. (1) is not available for T_S. Although a Dyson-like summation can be performed for this propagator, when computing the loglikelihood on data implementing it leads to a combinatorial blow-up 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 an RNN-style computational graph as illustrated in Fig. S7B.

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:

$$T_S = \sum_{n=0}^{n_{\max}} D_{\bar{S}} \; \pi \; (D_{S+x} \; \pi)^n \tag{10}$$

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

$\pi D_{\bar{S}}$ is here a skip-connect indicating 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 π matrix. Even without masking, however, the diagonal elements of π matrix vanishes after fitting the data.

We implement the computational graph with PyTorch. Negative log-likelihood loss function between calculated source-to-target log-probability with 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 form 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 m-th 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, $1 \leq \alpha_k \geq N_T$ (N_T being the number of trials) and $1 \leq m_k \geq L$ (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 average is used instead of 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. S7C, 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. S8).

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. 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).

Supplementary Figures and Tables

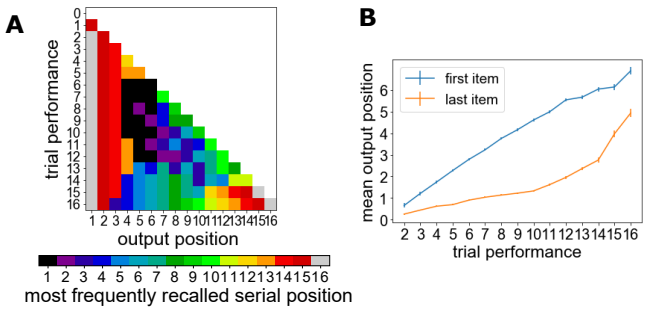

Fig. S1. Overview of recall by serial position of recalled items. (A) Most commonly recalled serial item at any given output position in trials with given performance. Notice in particular the bold recency effect, the subsequent primacy cluster, and the anomalous termination of the high performance trials. **(B)** Mean output position of first and last serial item, highlighting that primacy effects intervene after recency effects on average (error bars show the standard error of the mean).

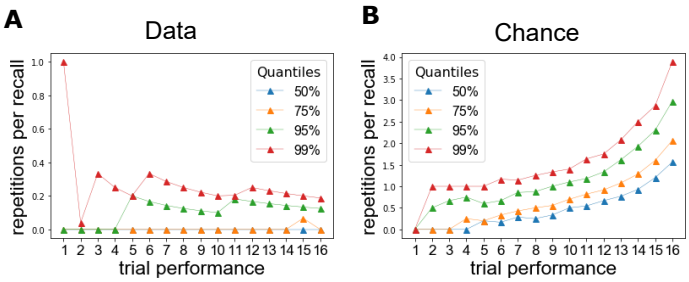

Fig. S2. Comparison to chance level of the amount of repetitions in the dataset. **(A)** Statistics of repetitions per trial. If we subgroup trials by their performance, we find that the number of repeated recalls increases as a function of performance but even for large performance over three quarters of trials contain no repetition, and the others little more than one each. **(B)** To provide a quantitative baseline, we simulated a purely associative model, which consists of Markovian transition in between serial positions. Notice that the model is complex enough to accommodate heterogeneous transition probabilities. Since this model does not incorporate bouncings, termination cannot be induced through the "impatience" mechanism discussed in the main text. Thus, we counted termination as transition to a sink state so that the transition matrix driving the process is of size $(N+1)\times(N+1)$. We extracted this transition matrix from a fit of the dataset and ran it starting from an initial condition (distribution of the first recall over serial positions) directly extracted from the dataset. The number of resulting repetitions is vastly larger than in the data.

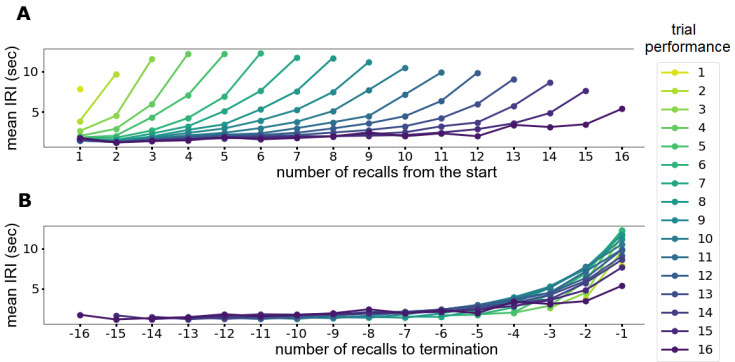

Fig. S3. Terminal Alignment of mean IRIs. **(A)** Mean IRIs computed over trials with fixed performance. Aligning the curves by the first recall of the trials does not reduce observed variance. **(B)** Alignment at the end, on the other hand, makes the mean IRIs collapse (and explains 20% of variance in the IRIs).

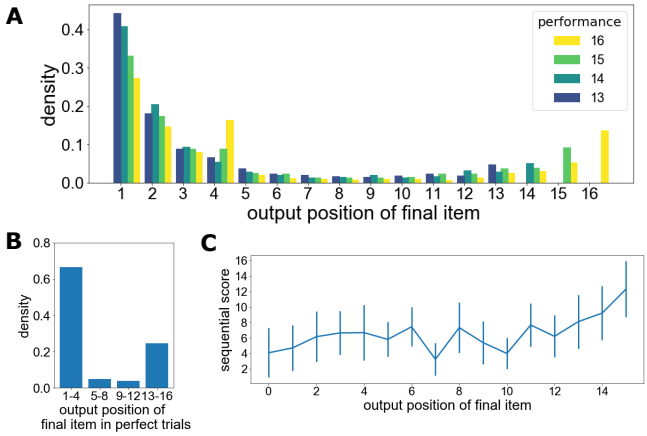

Fig. S4. Bimodality of the final item's output position in high performance trials. (A) Full distribution of the final item's output position over high performance trials. **(B)** Distribution in perfect trials over grouping of four consecutive serial positions. **(C)** Relationships to the sequential score in perfect trials (error bars display standard deviation).

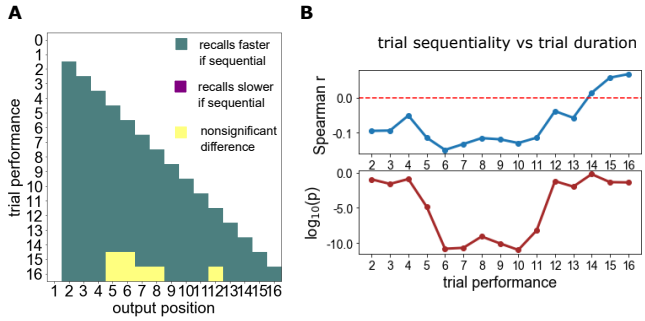

Fig. S5. Dependency of recall speed on sequentiality. The dependency is shown here on a recall-by-recall (panel A) and trial-by-trial (panel B) basis. **(A)** Sign of the difference between the mean IRI of sequential and non-sequential recall events, broken down for fixed output positions and trial performances. Color coding: *white* if data are not available from both the sequential and non-sequential conditions; *yellow* if the Mann-Whitney p-value between the two corresponding subsamples is lower than $p* = 10^{-3}$; *green* if sequential recalls are significant faster on average; no case is found where they are slower on average. **(B)** Spearman's correlation between overall duration of a trial and its sequentiality, computed over individual performance groups. Significance is found only for mid-performances, with the more sequential trials achieving faster termination.

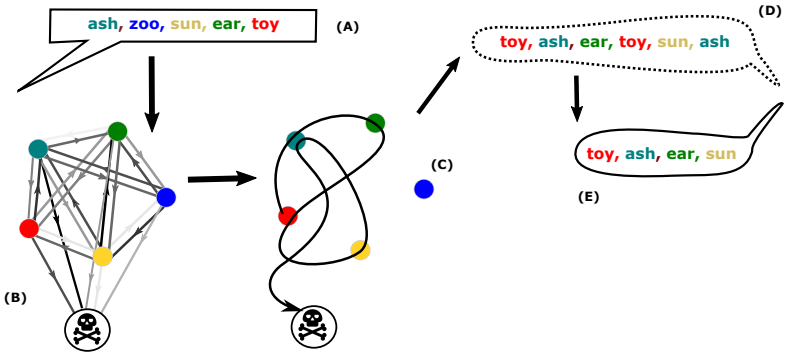

Fig. S6. Schematic representation of the skipping model. During the presentation stage **(A)** the subject is exposed (visually or orally) with a list of L items (in this example, $L = 5$ words). Memories of exposure to each of the items are stored in such a way that mental transitions among those memories have unequal probability, represented in **(B)** by the shade of gray of the arrows. During the recall stage **(C)**, the subject instantiates a mental trajectory through those memories. Nothing prevents the sequence of retrieved memories from containing an unlimited number of repetitions **(D)**, but the output is devoid of repetitions **(E)**, a fact possibly explained by self-censoring.

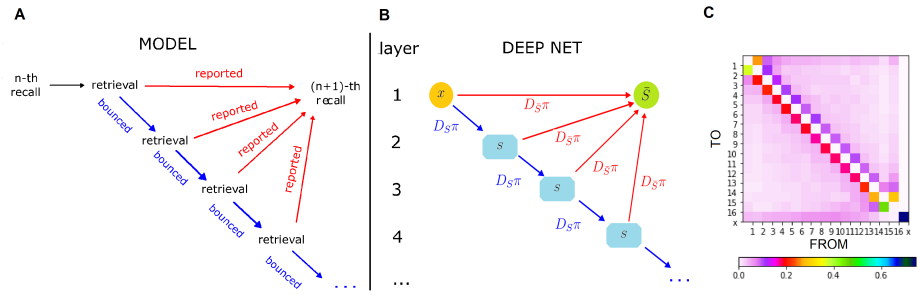

Fig. S7. Deep network architecture for fitting the skipping model. The two panels show the correspondence between the"skipping model and its implementation on a deep RNN-style neural network. **(A)** The process occurring between two consecutive recalls is mediated by an arbitrary number of retrieval that go unreported unless they provide a novel item. **(B)** Inter-layer weights in terms of the transition matrix π (and in terms of the other matrix operators defined in Supplementary Note 2) represent the operation of bouncing or reporting each retrieval, with each bounce corresponding to a new layer of the network. **(C)** Resulting transition matrix, with the sink state labeled as x.

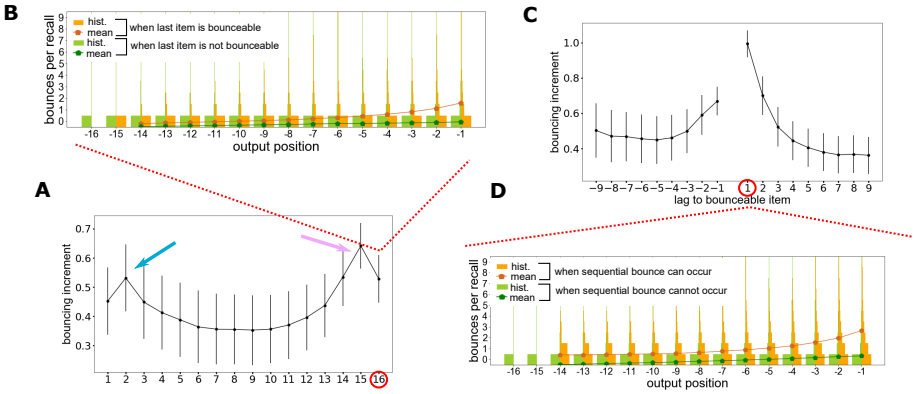

Fig. S8. Silent serial-position effects in the skipping model. Panels **(A-B)** correspond, for the skipping model, to panels A-B of Fig. 7 for the bouncing model – and panels **(C-D)** to panels A-B of Fig. 8. Predictions from the skipping model are substantially similar to those of the bouncing model in regard to lag-dependent silent effects. Silent effects depending on serial position depart from bouncing by the stronger emergence of silent primacy (blue arrow), here close to being on the same footing as silent recency.

Term	**Definition**
Contiguity effect	Tendency to recall contiguous items contiguously
Forward asymmetry	Tendency to recall items in forward order
Impatience	Probability of quitting the trial per unit bounce
Inter-Recall Interval	Time elapsing between two consecutive recalls
Lag	Difference between serial positions of the current and latest recall within a trial
Negative output position	Number of extant recalls to the end of a trial
Output position	Ordered position of an item in the sequence of reported recalls
Performance	Number of words explicitly recalled in a trial
Primacy effect	Strong recallability of initial items in the list
Recency effect	Strong recallability of final items in the list
Repetition bouncing	Discarding of a repetitious retrieval d
Sequential retrieval	A retrieval occurring with lag $L = +1$
Sequential slowdown	Sequential recalls in a given output position being performed slower in more sequential trial
Serial position	Ordered position of an item in the presented list
Silent contiguity	Recall delay associated to discarded retrievals with lag $L = +1$
Silent recency	Recall delay associated to discarded retrievals of the final item in the list
Strategic bouncing	Discarding of non-sequential retrievals to pursue a sequential recall strategy
Strategicity	Probability of strategically discarding a non-sequential item if retrieved
Trial sequentiality	Fraction of sequential recalls in a given trial

Table S1. Main definitions used in the text.