Eye Movements Reflect Causal Inference During Episodic Memory Retrieval

Yul HR Kang (yul.kang.on@gmail.com)[1*], Johannes Mahr (mahr_johannes@phd.ceu.edu)[2*]
Márton Nagy[3], Krisztina Andrási[3]
Gergely Csibra[2,4†], Máté Lengyel[1,2†]

[1]Computational and Biological Learning Lab, Department of Engineering, Cambridge University, Cambridge, United Kingdom
[2]Department of Cognitive Science, Central European University, Budapest, Hungary
[3]Department of Cognitive Psychology & MTA-ELTE Social Minds Research Group, Eötvös Loránd Universiy, Budapest, Hungary
[4]Department of Psychological Sciences, Birkbeck, University of London, London, United Kingdom
[*†]equal contributions

Abstract

During episodic memory retrieval, eye movements tend to distinguish between studied and unstudied items, a tendency known as "retrieval-dependent eye movements" (RDEs). However, what cognitive processes drive RDEs, and especially whether they are different from those that drive explicit choices, remains unknown. Here we dissect the cognitive processes underlying RDEs by model-based analyses of a false memory paradigm. Participants first memorized object-location pairs on a circular array ("learning"). They then saw object-location pairings allegedly produced by another participant in the upcoming memory test, and judged their correctness ("suggestion"). Finally, participants indicated the location of each object themselves ("retrieval"). A Bayesian cue-combination model that performed causal inference to assess whether the noisy memories of the learned and suggested object-location pairs (the two "cues") were from the same sources, and combined the memories accordingly, fit participants' explicit responses well. We also found that eye movements reflected the learned and the suggested stimulus even after controlling for the effects of explicit responses. Thus, RDEs contain information beyond that present in explicit responses, and they reflect the dynamics of the causal inference process underlying memory retrieval.

Keywords: episodic memory; eye movements; Bayesian ideal observer; false memory; causal inference

Introduction

It is well known that episodic memory retrieval involves "retrieval-dependent eye movements" (RDEs; Johansson, Holsanova, Dewhurst, & Holmqvist, 2012; Johansson & Johansson, 2013; Richardson & Spivey, 2000; Staudte & Altmann, 2017): in a recognition task, participants' eye movement patterns distinguish between studied and unstudied items. However, it has been debated whether RDEs show such differentiation when participants fail at making correct explicit responses, and relatedly, whether RDEs are driven by explicit or implicit forms of memory (Hannula, Baym, Warren, & Cohen, 2011; Smith, Hopkins, & Squire, 2006; Nickel, Henke, & Hannula, 2015; Smith & Squire, 2017; Urgolites, Smith, & Squire, 2018). Here we used a location memory task to compare how RDEs are affected by a memorized location when it is recognized versus forgotten, separately from the effect of mere exposure. We did so by asking participants to indicate a studied (or "learned") location as well as correctness of an unreliable "suggestion": we expected participants to ignore the suggested location when they deemed it wrong, although they were exposed to it. We then cast memory retrieval as Bayesian causal inference, whereby participants try to infer whether the suggestion on a given trial was correct and—based on that—what the learned location could have been, from unreliable representations of both the learned and the suggested locations. We show (1) that such a model fits participants' responses well, (2) that we can decode the learned and the suggested stimulus from participants' responses, and (3) that participants' gazes are attracted to the learned and suggested locations, even when they differ from each other and from the responded location, and that these effects depend on whether the suggestion is deemed correct.

Methods

Participants

A total of 17 participants performed the task (9 females; age 19–27). The experimental protocol was approved by the Institutional Review Board of Central European University. Participants gave written informed consent before starting the experiment.

Task

Participants were presented with a series of object-location pairs on a circular array and were asked to remember the location of each object for a later memory test ("learning phase"; Figure 1). Next, participants were presented with (50% correct) object-location pairs allegedly produced by another participant in the upcoming memory test and were asked to judge the correctness of each of these pairs in light of the learning phase ("suggestion phase"). Finally, participants were asked to indicate the location of each object themselves ("retrieval phase"). They were asked to wait while a target image (2 s), a blank screen (0.5 s), and an empty array of locations (2 s) were presented, after which a mouse cursor appeared at the center of the screen, prompting them to respond by clicking one of the 12 locations that they thought was paired with the target image in the learning phase.

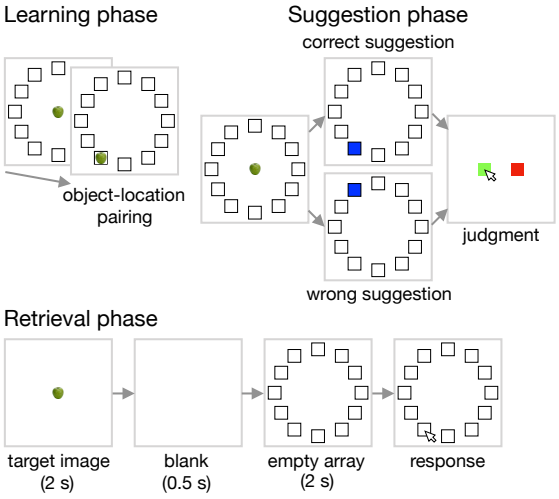

Figure 1: Behavioral task. See text for details.

Analysis of Explicit Responses

Our ultimate goal is to examine how much information gaze locations have about the learned and suggested locations (L and S), over and beyond what the explicit responses have (remembered location R and subjective, or "deemed", correctness of the suggestion D). As a baseline, we examined how much information the explicit responses have about the learned and suggested locations using a Bayesian decoding approach (the performance of a Bayesian decoder measured by cross-entropy is equivalent to the mutual information up to a change in sign and an additive constant). We will include gaze locations for decoding as a next step.

We constructed a Bayesian ideal observer model (the "encoder"), $\hat{\mathrm{P}}(R,D|z_\mathrm{L},z_\mathrm{S};\theta_\mathrm{en})$, that generated the explicit responses about the remembered location (R) and the subjective correctness of the suggestion (D) based on noisy memory representations (z_L and z_S) of the learned and suggested locations (L and S; Figure 2**A**). We first fit the parameters of this encoder model (θ_en) by maximizing the likelihood $\prod_j \hat{\mathrm{P}}(R_j,D_j|L_j,S_j;\theta_\mathrm{en})$, where j indexes trials, and the predictive distribution for each trial is obtained by integrating out the noisy memory representations of the ideal observer that are unknown to the experimenter $\hat{\mathrm{P}}(R,D|L,S;\theta_\mathrm{en}) = \int \hat{\mathrm{P}}(R,D|z_\mathrm{L},z_\mathrm{S};\theta_\mathrm{en})\,\mathrm{P}(z_\mathrm{L},z_\mathrm{S}|L,S,\theta_\mathrm{en})\,dz_\mathrm{L}\,dz_\mathrm{S}$. We then fixed the parameters θ_en, and examined how much information the explicit responses had about the stimuli, by decoding L_j or S_j using R_j and D_j. We measured decoding performance by cross entropy:

$$\mathrm{CE}(L|R,D;\theta_\mathrm{en}) = -\frac{1}{N}\sum_j \log_{12}\hat{\mathrm{P}}(L_j|R_j,D_j;\theta_\mathrm{en}) \quad (1)$$

where

$$\hat{\mathrm{P}}(L_j|R_j,D_j;\theta_\mathrm{en}) \propto \sum_S \mathrm{P}(L_j,S)\,\hat{\mathrm{P}}(R_j,D_j|L_j,S;\theta_\mathrm{en}) \quad (2)$$

is the "decoding" distribution obtained by the Bayesian inversion of the predictive distribution of the encoding model (see above), and N is the number of trials.

To ensure that we don't lose information by using the Bayesian ideal observer model, we also fit (1) parametric decoders where the parameters (called θ_de) are directly optimized to maximize the decoding performance, and (2) nonparametric decoders that do not assume Bayesian inference. To prevent overfitting, we used 10-fold cross-validation for every model in evaluation.

Analysis of Eye Movements

We analyzed eye movements during the retrieval phase from the target image onset until the onset of the mouse cursor (4.5 seconds after the target image onset), in order to determine how much information about the learned and suggested object-location pairings can be recovered from retrieval-dependent eye movements. To dissociate the effects of the learned, suggested, and responded locations, we constructed a multiple regression model of the following form for the (two-dimensional) gaze location y_t in a given time bin indexed by t:

$$y_t = \beta_{0,t} + \beta_{\mathrm{R},t}\,x_\mathrm{R} + \sum_i \sum_D \beta_{i,D,t}\,x_{i,D} + \varepsilon_t \quad (3)$$

where $\beta_{0,t}$ is an overall bias, x_R is the responded location on trials when the response differed from the learned and suggested locations and 0 otherwise, $\beta_{\mathrm{R},t}$ is x_R's effect on the gaze, $x_{i,D}$ are the location of the learned, suggested, and/or responded locations when some of them are the same ($i \in \{L=S, L=R, S=R, L=S=R\}$) or different from the other two ($i \in \{L,S\}$), when the suggested location is deemed correct or not ($D=1$ or 0) with $\beta_{i,D,t}$ being their effects on the gaze, and ε_t is (spatially and temporally i.i.d.) Gaussian noise.

For visualization, we used every 100 ms time bin in the first 4.5 s from the target image onset in the retrieval phase as t (Figure 3). For statistical tests, we used the difference of the average gaze position between 0.5–1 s after the array onset (which is of interest: Figure 3, gray bar on the time axis) and the 0.5-second period prior to the target image onset (as a baseline).

Results

Task Performance

Participants' performance in responding with the learned location was higher when it matched the suggested location (i.e., $\mathrm{P}(R=L|L=S) > \mathrm{P}(R=L|L\neq S)$, 65% vs. 47%; $p<0.001$, sign test across participants; for reference, chance performance would be ~8.3%), indicating that they used the suggestion adaptively. They also deemed the suggested location correct more when it matched the learned location (i.e., $\mathrm{P}(D=1|S=L) > \mathrm{P}(D=1|S\neq L)$, 82% vs. 22%; $p<0.001$, sign test across participants), and responded to the suggested location more when it was deemed correct (i.e., $\mathrm{P}(R=S|D=1) > \mathrm{P}(R=S|D=0)$, 70% vs. 9%; $p<0.001$, sign test

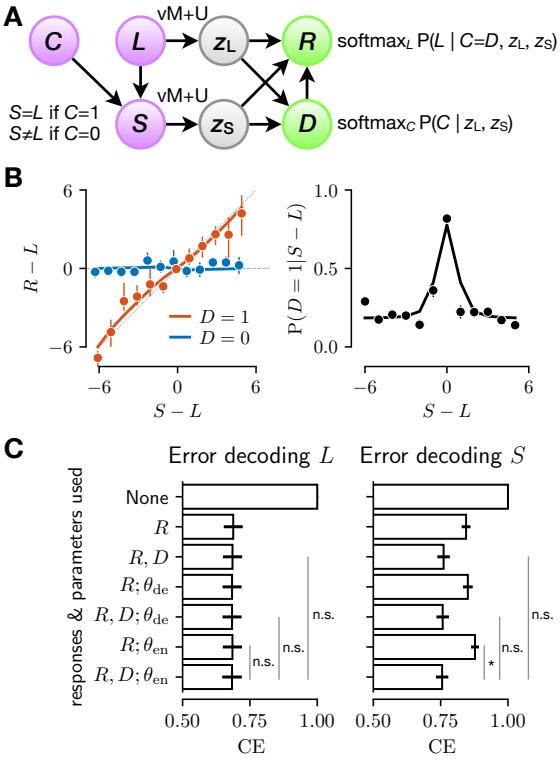

Figure 2: Analysis of explicit responses. **A.** Graphical model of the Bayesian ideal observer (see text for details). vM + U means a mixture of von Mises and uniform distributions. Purple: known to the experimenter; gray: known to the participant; green: known to both. **B.** Prediction of the Bayesian ideal observer model (curves) and the data (markers; circular mean across participants with bootstrapped 90% CI). The model is fit to each participant's data individually; it is pooled across participants only for visualization. *Left.* The deviation of the response from the learned location as a function of the deviation of the suggested location from the learned location. Red/blue indicates when the suggestion is judged correct/wrong, as in Figure 3. *Right.* The probability of judging the suggestion correct as a function of the deviation of the suggested location from the learned location. **C.** Decoding performance (for L or S) from explicit responses. *Left.* Each row shows $\mathrm{CE}(L|\cdot)$ (mean across participants $\pm$ SEM), where $\cdot$ is indicated in the row label. This would be 0 with a perfect prediction, and 1 with a uniform prediction (top row). Other rows (from top to bottom): performance of nonparametric models (R and R, D), parametric models directly optimized for decoding ($R; \theta_{de}$ and $R, D; \theta_{de}$), and the Bayesian ideal observer model ($R; \theta_{en}$ and $R, D; \theta_{en}$). *Right.* Each row shows $\mathrm{CE}(S|\cdot)$ (mean across participants $\pm$ SEM). n.s.: $p > 0.1$; *: $p < 10^{-6}$

across participants), showing that they made informed judgments about the correctness of the suggestion.

Causal Inference Explains Explicit Responses

The Bayesian ideal observer model performing causal inference was able to fit participants' response patterns. (1) The response (R) followed the suggestion (S) more when it was deemed correct (Figure 2**B**, *left*, red curve and markers), compared to when it was deemed wrong (blue). (2) The response interpolated between the learned and suggested locations when the suggestion was deemed correct (the red curve and markers are between the horizontal line and the diagonal line; Shams & Beierholm, 2010). (3) The suggestion was deemed correct more often when it was close to the learned location (Figure 2**B**, *right*).

Decoding Stimuli From Explicit Responses

The decoder using the Bayesian ideal observer model (whose parameters were optimized to predict R and D given L and S) predicted L successfully from responses R and D. Its performance was on par with the decoder directly optimized to predict the stimuli, and with the nonparametric decoder, suggesting that using the Bayesian ideal observer model did not lose information (Figure 2**C**, *left*: $\mathrm{CE}(L|R,D;\theta_{en}) - \mathrm{CE}(L|R,D) = -0.005 \pm 0.003$ ($p = 0.13$) and $\mathrm{CE}(L|R,D;\theta_{en}) - \mathrm{CE}(L|R,D;\theta_{de}) = -0.002 \pm 0.001$ ($p = 0.14$)). The same held for decoding S from R and D (Figure 2**C**, *right*: $\mathrm{CE}(S|R,D;\theta_{en}) - \mathrm{CE}(S|R,D) = -0.005 \pm 0.003$ ($p = 0.14$) and $\mathrm{CE}(S|R,D;\theta_{en}) - \mathrm{CE}(S|R,D;\theta_{de}) = -0.002 \pm 0.001$ ($p = 0.14$)). Note that the Bayesian ideal observer model's performance was slightly, although not significantly, better than the models directly optimized for decoding in the above comparisons which are done in the test set (but, reassuringly, not in the training set; data not shown), indicating that it was a faithful model of participants' behavior.

Analysis of Eye Movements

As expected, we found that the learned location attracted gaze during the retrieval phase. This was the case even when it was different from the suggested and the responded locations when the suggested location was deemed wrong ($\beta_{L,D=0,t} = 0.069 \pm 0.024$, $p = 0.01$; Figure 3, *top*, shaded interval). Note that any contribution to gaze locations from suggested or responded location is regressed out by the regressors for them ($\beta_{S,D=0,t}$, $\beta_{R,D=0,t}$, and $\beta_{S=R,D=0,t}$), so $\beta_{L,D=0,t}$ represents the "pure" effect of the learned location.

Surprisingly, we found that the suggested location, too, attracted gaze during the retrieval phase. This was the case even when the suggested location differed from the learned or responded locations, but only when it was deemed correct ($\beta_{S,D=1,t} = 0.20 \pm 0.08$, $p = 0.01$; Figure 3, *bottom*, shaded interval). Again, note that any contribution to gaze locations from learned or responded location is regressed out by the regressors for them ($\beta_{L,D=1,t}$, $\beta_{R,D=1,t}$, and $\beta_{L=R,D=1,t}$).

Discussion

We found that the learned location attracted gazes independent of the suggested or responded locations, and that its ef-

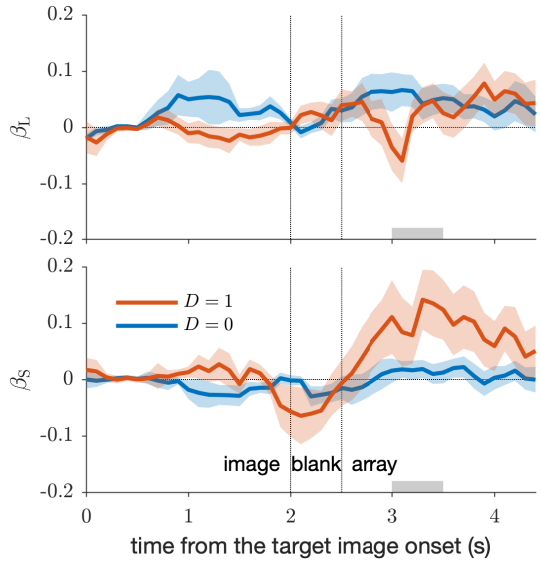

Figure 3: Analysis of eye movements. *Top*. Effect ($\beta_{L,D,t}$) of the learned location (L) when it differs from the suggested and responded locations (S and R), when the suggestion was deemed correct or wrong ($D = 1$ and $D = 0$, red and blue.) *Bottom*. Effect ($\beta_{S,D,t}$) of the suggested location (S) when it differs from the learned and responded locations (L and R), when the suggestion was deemed correct or wrong ($D = 1$ and $D = 0$, red and blue.) The average gaze position within the shaded interval was used for statistical tests.

fect depended on whether the suggestion was deemed correct. Conversely, we also found that the suggested location attracted gazes independent of the learned or responded locations, only when it was deemed correct. These results suggest that RDEs are not merely a reflection of an exposure or a rehearsal of a response to be made; instead they suggest that RDEs reflect a causal inference process where the relevance of the exposure (learned and suggested locations) is inferred based on the similarity between the unreliable memories of the learned and suggested locations (cf. Shams & Beierholm, 2010).

To clarify how memories affect RDEs, we plan to decode the learned and suggested locations from gazes, in order to compare the information contained in gazes on an equal footing with that contained in explicit responses. While we do not have direct access to the internal memory representations of the participants, we can infer them using the Bayesian ideal observer model (which fits participants' inference process well, as it could decode the stimuli from explicit responses.) This will help reconcile the debate about whether RDEs depend on explicit or implicit forms of memory, by telling us whether RDEs are generated after explicit responses are determined (and hence contains no further information about the stimuli), or RDEs derive from the internal representation of the memory at least in part separate from the explicit responses. Our regression analyses already suggest that eye movements indeed contain information separate from the explicit responses, although the information is not measured yet in the same units.

Acknowledgments

This work was supported by a Wellcome Trust Investigator Award in Science and ERC Consolidator Grant to ML.

References

Hannula, D. E., Baym, C. L., Warren, D. E., & Cohen, N. J. (2011). The Eyes Know. *Psychological Science*, *23*, 278–287. doi: 10.1177/0956797611429799

Johansson, R., Holsanova, J., Dewhurst, R., & Holmqvist, K. (2012). Eye movements during scene recollection have a functional role, but they are not reinstatements of those produced during encoding. *Journal of Experimental Psychology: Human Perception and Performance*, *38*, 1289. doi: 10.1037/a0026585

Johansson, R., & Johansson, M. (2013). Look Here, Eye Movements Play a Functional Role in Memory Retrieval. *Psychological Science*, *25*, 236–242. doi: 10.1177/0956797613498260

Nickel, A. E., Henke, K., & Hannula, D. E. (2015). Relational Memory Is Evident in Eye Movement Behavior despite the Use of Subliminal Testing Methods. *PLOS ONE*, *10*, e0141677. doi: 10.1371/journal.pone.0141677

Richardson, D. C., & Spivey, M. J. (2000). Representation, space and Hollywood Squares: looking at things that aren't there anymore. *Cognition*, *76*, 269–295. doi: 10.1016/s0010-0277(00)00084-6

Shams, L., & Beierholm, U. R. (2010, 9). Causal inference in perception. *Trends in Cognitive Sciences*, *14*(9), 425–432. doi: 10.1016/j.tics.2010.07.001

Smith, C. N., Hopkins, R. O., & Squire, L. R. (2006). Experience-Dependent Eye Movements, Awareness, and Hippocampus-Dependent Memory. *The Journal of Neuroscience*, *26*, 11304–11312. doi: 10.1523/jneurosci.3071-06.2006

Smith, C. N., & Squire, L. R. (2017). When eye movements express memory for old and new scenes in the absence of awareness and independent of hippocampus. *Learning & Memory*, *24*, 95–103. doi: 10.1101/lm.043851.116

Staudte, M., & Altmann, G. T. (2017). Recalling what was where when seeing nothing there. *Psychonomic Bulletin & Review*, *24*, 400–407. doi: 10.3758/s13423-016-1104-8

Urgolites, Z. J., Smith, C. N., & Squire, L. R. (2018). Eye movements support the link between conscious memory and medial temporal lobe function. *Proceedings of the National Academy of Sciences*, *115*, 201803791. doi: 10.1073/pnas.1803791115