A Model of Memory and Emotion Mechanisms Underlying the Continued Influence Effect*

Alexander R. Hough[1] and Othalia Larue[2]

[1] Air Force Research Laboratory, Wright-Patterson AFB, OH 45433, USA
[2] Parallax Advanced Research, Beavercreek, OH 45431, USA

Abstract. Human cognition is efficient, but vulnerable to misinformation and influence. An example is the continued influence effect (CIE), where misinformation has a lasting effect even after presentation of corrections/retractions. Experiments addressing the CIE produced memory-based (i.e., episodic and mental model updating) explanations, however, models are scarce. We recently developed a cognitive model of the CIE focusing on encoding and navigating memory, and were able to approximate and fit human behavior. Here, we extended the model by including basic affect mechanisms (whether positive/negative and its intensity) that allow affect to be associated with words to influence the strength (i.e., activation) of specific memories. The extended model provides a better fit and demonstrates how the CIE emerges from memory and emotion processes. We discuss relevant literature, present results with model comparisons, and discuss challenges to address in future research.

Keywords: Continued influence effect · cognitive modeling · ACT-R · misinformation · knowledge representation · core affect.

1 Introduction

Humans can successfully exploit regularities in the environment using heuristics [14], but they can also lead to biases and systematic errors [18]. Heuristics are developed in "benign" or stable/predictable environments and are vulnerable in "hostile" environments where cues are misleading [19]. They could be exploited to mislead or influence [27]. One example is the continued influence effect (CIE), where misinformation has a lasting effect on decisions after corrections/retractions [16, 20]. We start by presenting the first-ever computational cognitive model of the CIE that focused on memory [15]. Then we extend it by adding basic affect (positive/negative and intensity) capabilities so affect can be associated with words and information sources to influence the strength (i.e., activation) of memories. This allows testing of previous explanations and leads to a better understanding how memory and emotion interact in the CIE.

* This research was supported by the U. S. Air Force Research Laboratory's Cognition and Modeling Branch. Contents were reviewed and approved for public release: AFRL-2024-3054. The views expressed are those of the authors and do not reflect the official guidance or position of the United States Government, U.S. Department of Defense, U.S. Air Force, or any subsidiaries or employees.

1.1 The Continued Influence Effect

CIE research often presents two articles about a narrative: the first contains misinformation and the second has a correction. Corrections reduce, but do not eliminate behavior consistent with misinformation [6, 8, 16]. The CIE is robust and can only be reduced by 50% through mitigation [20]. This is concerning as experiments involve ideal situations where corrections or retractions reach everyone and closely follow misinformation.

Explanations of the CIE focus on episodic memory, where memory cannot be erased but rather re-activated or associated with other information [31]. This may lead to errors from competing memory activations [2, 12], recency effects [9], or familiarity-based fluency [11, 29]. People also rely on easier-to-access information [18] and corrections may reactivate some elements of misinformation through repetition, making it more "available". Research also suggests that initial presentation of misinformation leads to a coherent mental model that does not accommodate later corrections [31, 16, 20].

Another important consideration is emotion, which can influence information processing and memory. Correcting misinformation may create feelings of discomfort contributing to the CIE [28] and may be exacerbated when corrections are less coherent [20]. Higher emotional responses are associated with increased belief in fake news [21] and spread of misinformation [3]. Emotional experiences are more accessible [7] and remembered better [33], but may be less accurate [23]. People often rely on emotion associated with information [13], which may carry more weight than experience in certain contexts [24]. People attend more to negative information and losses [4], which is better remembered [30] and more easily recalled [32]. A recent article [25] mentioned we need to better understand how emotion influences memory related to misinformation.

Our understanding of the CIE is limited. Research produced some mixed results and no computational cognitive models exist to thoroughly test hypotheses. Here, we focus on two needs recently expressed by experts in the field [10]: 1) understanding the interplay between cognition, social, and emotional factors, and 2) an overarching theory and model including these factors, and spanning from individuals to groups. We start addressing these needs by extending the first ever cognitive model of the CIE [15] with basic affect mechanisms. We focus on general mechanisms, which can be used to predict and simulate human behavior, and assist in identifying mitigation strategies.

2 Cognitive Model

The CIE model was implemented in ACT-R [1], which is a hybrid cognitive architecture with symbolic and sub-symbolic structures. There are perceptual-motor and memory modules representing systems of the mind. Perceptual-motor modules enable perception of stimuli, actions like pressing buttons, and goal directed behavior. The declarative memory module represents facts as chunks in long-term memory, and a sub-symbolic component determines their availability. The

procedural module represents knowledge about how to do things, represented as condition-action rules. Behavior is represented as a series of rule firings that change the state of the model. In the following section, we describe how ACT-R declarative memory and affect mechanisms can simulate the CIE.

2.1 Declarative Memory and Core Affect Mechanisms

The literature suggests the CIE involves competition between misinformation and corrections. In ACT-R, declarative memory captures this competition through chunk activation. A chunk is comprised of slot/value pairs and is the basic unit of declarative memory. Each chunk has an activation value corresponding to the probability and speed of its retrieval in a given situation [1]. A chunk's activation, A_i, is a function of the: 1) base level term, B_i, representing recency and frequency of chunk use, 2) spreading activation, S_i, dividing activation among related chunks, 3) partial matching, P_i, allowing retrievals of imperfect matches, and 4) noise term, ϵ_i, representing variability in memory. Here, we only use B_i, ϵ_i, and add two terms (V_i and Ar_i as well as, aw and vw their respective weights), which we explain later. The base level term, B_i, is important for opposing dynamics of learning with experience and forgetting across time: n_i, is the number of times chunk i has been used or retrieved, t_{ij} is elapsed time in seconds since the j^{th} retrieval, and $d \in [0, 1]$ is a decay parameter:

$$A_i = \boldsymbol{B_i} + S_i + P_i + \boldsymbol{\epsilon_i} + (vw * \boldsymbol{V_i}) + (aw * \boldsymbol{Ar_i}) \qquad \boldsymbol{B_i} = \log\left(\sum_{j=1}^{n_i} t_{ij}^{-d}\right) \quad (1)$$

Chunks are retrieved based on retrieval cues (i.e., a slot or value) and can compete when several chunks possess the same cue. The chunk with the highest activation (i.e., was created with a higher base level activation and/or was used frequently) would be more likely to be retrieved. Base level activation at creation can represent the weight or value of information. However, it does not include emotion, which can make emotional memories more accessible [7] and/or easier to recall [33]. Recent research inspired by the core affect theory of emotion [26], laid the initial groundwork for incorporating affect into the ACT-R declarative memory system [17]. Core affect focuses on feelings underlying emotion using two dimensions: valuation (positive or negative) and arousal (magnitude). A module was developed [17] to compute valuation, V_i, and arousal, Ar_i, which affect chunk activations through the activation equation. The current valuation of chunk i at the j^{th} use is based on its previous valuation $V_i(j-1)$ and the difference between the previous valuation and current reward $R_i(j)$ multiplied by a valuation learning rate av. Arousal is the absolute magnitude of valuation and represents the importance of a chunk:

$$V_i(j) = V_i(j-1) + \text{av}[R_i(j) - V_i(j-1)] \qquad Ar_i(j) = abs(V_i(j) \quad (2)$$

Valuation and arousal are updated each time a chunk is referenced within a time window over which to update the valuations. This differentiates core-affect learning from utility learning where rewards are used as boundaries. Valuation and

arousal can also be used as retrieval cues. They directly affect chunk activations and affect memory dynamics. For instance, a chunk with associated negative affect could have greater activation and this effect could persist over time and affect decision-making despite the accumulation of conflicting evidence.

2.2 Model Description and Processes

The CIE model used six declarative memory parameters. The first four were based on default or recommended values and the last two were set higher than recommended to prevent an endless loop of retrievals (See [15] for more detail): 1) Retrieval threshold (changed from 1 to default of 0), that restricts which chunks can be retrieved based on activation, 2) base-level activation constant (lowered from 10 to 2.5), β_i, 3) base-level decay (.5), d, 4) activation noise (.25), ϵ_i, 5) declarative finsts that sets number of items marked as retrieved, and 6) declarative finst span that sets the time items remain marked. We did not include partial matching, P_i, or spreading activation, S_i, terms from the activation equation (equation 1). The initial CIE model (Model1) focused on retrievals and changes in chunk activation based on declarative memory dynamics (e.g., decay and frequency of use) to capture the competition between misinformation and corrections. The extended CIE model (Model2) included core affect mechanisms influencing activation through the activation equation. Model2 used five valuation parameters: 1) valuation weight (2), 2) valuation alpha or learning rate (1), 3) valuation time window (.5), 4) arousal weight (1), and 5) initial chunk valuation (1) (see [17] for additional details). In addition, there were two additional features added during the formation of chunks and during recall to capture the differential weighting of information based on affect associated with words and source trust/credibility. For affect associated with words, we leveraged an existing dataset that included valuation and arousal values for 20,000 words [22]. The current implementation of core affect uses rewards to update both valuation and arousal terms in the activation equation. Our initial reward value emphasizes negative affect based on the research with emotion and memory. The reward for each word is calculated by subtracting valuation (0-1) from 1 and multiplying by arousal. This gives a higher number for negative information and lower for positive. For source trust/credibility, we leveraged the trust/credibility ratings included with the human data [5, 8] and scaled them for activations.

We developed the model with both simplicity and generality in mind so that we could explore datasets and address challenges with modeling the CIE. Most CIE experiments involve participants reading narratives about event scenarios that include misinformation and corrections, then they answer questions that often include a combination of open-ended recall, inferences, and ratings of beliefs in misinformation and corrections. The model has two main processing stages: it "reads" narrative information for scenarios that were parsed into word pairs (i.e., predicates) via a separate language model and manual generation (see [15] for more details), then it gives a summary of each scenario after navigating the relevant knowledge representation. Figure 1 shows processes (blue rectangles),

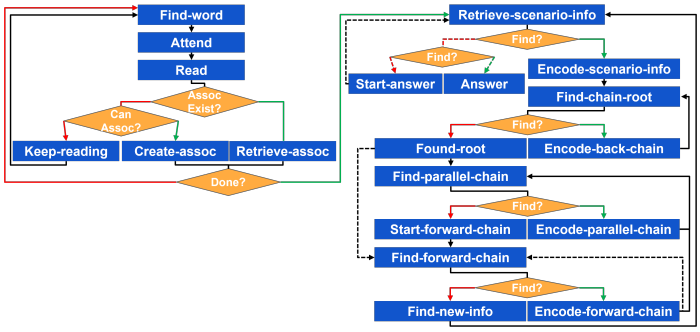


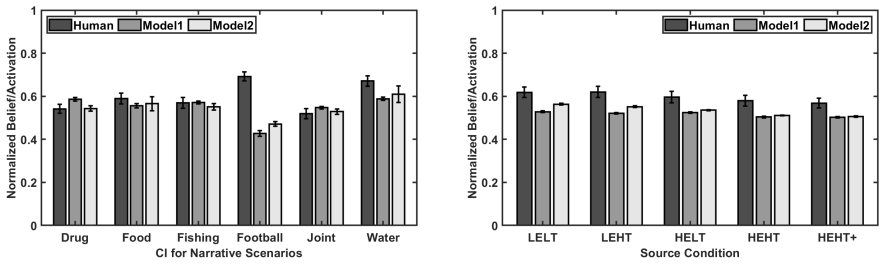
Fig. 1: Depiction of the CIE model processes

conditions (yellow diamonds), and flow of behavior (arrows). We use a combination of this answer and the activation of chunks in memory to approximate specific questions in experiments.

For the reading stage, the model directs visual attention to find, attend, and read (i.e., encode) word pairs. The model attempts to retrieve word associations in memory and if not possible, it creates a new chunk containing the word pair (and associated affect for Model2). After reading all scenarios, the model prepares a summary for each scenario. The model navigates its mental representation through a chaining process. It completes an open recall for a chunk (retrieve-scenario-info), then finds all associated chunks through backward (find-chain-root), parallel (find-parallel-chain), and forward (start-forward-chain) chaining and encodes them (Model2 reactivates affect here). Once all related chunks are recalled, it repeats the process until all chunks for the scenario have been retrieved. The model then starts the answer process (dotted line arrows) by completing an open recall for the most active chunk and finding the most active chain it belongs to (skips parallel chaining). This chain is then given as the summary answer (see additional details in previous paper [15]).

3 Initial CIE Task [8]

We used a CIE task from the literature [8] to develop the initial model [15]. The CIE task [8] used one narrative that included both misinformation and a correction. Corrections included source information and were used to manipulate credibility (low and high) and trustworthiness (low and high). Participants completed: 1) a recall question and inferential reasoning questions relating to each scenario, and 2) belief ratings for misinformation, and correction on a 10-point scale (not at all-very strong). We previously approximated the recall question and belief rating results across six scenarios [15]. We focused only on memory mechanisms and did not include source manipulations. Here, we extend the model by including emotion mechanisms and attempt to capture source manipulations.



(a) (b)

Fig. 2: Misinformation belief ratings and chunk activations for scenarios (a) and source conditions (b) for human and model data. Error bars are SEM.

3.1 Results

We simulated 50 participants for the memory only (Model1) and extended model with core affect (Model2). Common model fit statistics (e.g., likelihood function) are not possible for models implemented in the ACT-R architecture. Instead, correlations are used for fit to trends in the data and root mean squared error ($RMSE$) for the average difference. We place more weight on $RMSE$ for fit, because data has to be grouped (sometimes arbitrarily) for correlations. We faced challenges approximating misinformation scores [15], which are a composite of open-ended and inference questions that require some language understanding and composition. They remain a challenge for cognitively plausible models. Here, we only present belief score approximations and model comparisons across scenarios and source conditions.

Belief scores were given on a 1-10 scale and were approximated by averaging chunk activations for misinformation and corrections after memory navigation. For comparisons, we normalized belief ratings and activations. We first fit scenario data (Fig 2a) by collapsing source manipulations and organizing the human data into scenarios. This first comparison does not include source information, so differences between models are due to the addition of core affect mechanisms and affect associated with words. Model2 with core affect had a lower average difference with the human data, $RMSE$, $r(10) = -0.07, p = 0.89, RMSE = 0.09$, compared to Model1 with only memory, $r(10) = -0.53, p = 0.28, RMSE = 0.12$.

Next, we report belief rating model fits across source conditions: 1) low expertise/trust (LELT), 2) low expertise/high trust (LEHT), 3) high expertise/low trust (HELT), 4) high expertise/high trust (HEHT), and 5) highest trust/expertise (HEHT+). Note that Model1 was included as a baseline, but it had no mechanisms to deferentially weight source information. Only Model2 could interpret source information by associating affect based on trust and credibility ratings from [8]. For this comparison (Fig 2b), we included affect associated with source information only. Model 2 had a lower average difference with the human data, $r(8) = 0.98, p = 0.004, RMSE = 0.06$, compared to Model1, $r(8) = 0.88, p = 0.052, RMSE = 0.08$.

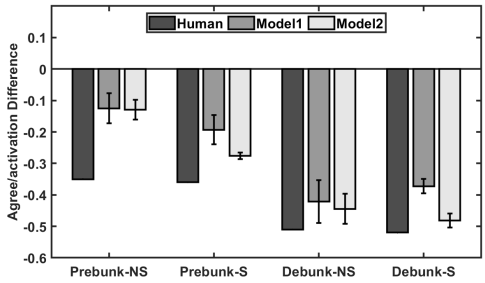


Fig. 3: Difference from control in misinformation claim agreement for prebunks, debunks, and source (NS or S). Error bars are SEM

4 Second CIE Task [5]

We used a second task to assess whether the CIE model(s) could fit additional datasets and conditions. The task [5] includes presenting a correction article before (i.e., prebunk) or after (i.e., debunk) the misinformation article. Corrections either did or did not include source information. There was also a control without corrections. Participants answered questions about agreement with claims (misinformation), credibility of corrections, and intention to engage in discussion. Here, we focus on the agreement with claims and use their method of correction effectiveness. They used the control as a baseline and calculated the agreement difference for prebunk and debunks with and without source information.

4.1 Preliminary Results

We present preliminary results with 50 simulated participants and affect associated with source information only (Fig 3). We used the average chunk activation for misinformation chunks to represent agreement with claims. Just like [5], we subtracted the agreement for the control from the agreement of prebunk and debunk conditions (with and without sources). Model2 had a lower average difference, $r(6) = 0.94, p = 0.06, RMSE = 0.12$, compared to Model 1, $r(6) = 0.97, p = 0.03, RMSE = 0.16$.

5 Discussion

We presented a previous CIE model (Model1) that explored competition between misinformation and corrections based on activation of chunks in memory [15]. Here, we changed values of three basic parameters (i.e., retrieval threshold, base-level activation constant, and activation noise) to better align with default values and improve fit. We then extended the model to include emotion mechanisms to associate valuation and arousal with words and information sources.

The emotion mechanisms were based on core affect [26] implemented as a valuation module [17] that directly affected chunk activation. These mechanisms align with research findings where emotion can increase memory [7, 30, 32, 33], belief in misinformation [3, 21], and attention to information [4]. We used a word list with associated valuation and arousal [22] to calculate affect associated with words and leveraged trust/credibility ratings for affect associated with source information [8, 5]. The model with emotion (Model2) had a lower average difference from the human data across datasets, demonstrating it captured human behavior better than the memory only model (Model1). We note that adding emotion increased model complexity and only slightly reduced the average difference with human data. Increasing complexity often increases fitting power, but ACT-R already has significant constraints and introducing emotion mechanisms added additional constraints that further reduced the fitting power.

Adding affective mechanisms and language processing (i.e. transformation of articles into ACT-R readable inputs) introduced several challenges. The word list with associated valuation and arousal did not include all emotionally charged words from the experimental materials, and there were issues with matching. For instance, hyphenated words (e.g., industrial-pollutants) and words in different tenses (e.g., contaminants, contamination, and contaminated) could not match. Our method using source information had some unexpected consequences. For instance, increasing activation of chunks encoded or retrieved directly afterward. Lastly, we did not include a version of Model2 that included both affect associated with words and source information. There were unexpected interactions that averaged out some effects of words or source information, which we need to better understand. We plan to continue to develop a better and more appropriate method to include affect associated with words and sources to better assess the benefits of adding emotion mechanisms, if any, to memory in this context.

We note several other limitations: 1) We manually generated predicates and focused on declarative memory with chunk chaining, 2) we selected questions congruent to the reasoning and expression capacities of our model (i.e while inferential reasoning was too complex for our model, belief rating was possible). We will address these limitations in future work. We will improve our language parsing method, utilize additional declarative memory components, and include similarity information to reduce the need for direct memory matching during chaining. We plan to leverage previous work with analogical reasoning to extend the question answering capabilities of the model.

Overall, we demonstrated a CIE using general components of declarative memory and core affect, which can be used for any information presented in word pairs. The model could fit two datasets without modification and provides a good base for future research to explore social factors, group interactions, and theoretical explanations across experiments and datasets.

References

1. Anderson, J.R.: How can the human mind occur in the physical universe? Oxford University Press, New York, NY (2007)

2. Ayers, M.S., Reder, L.M.: A theoretical review of the misinformation effect: Predictions from an activation-based memory model. Psychonomic Bulletin and Review **5**, 1–21 (1998)
3. Baum, J., Abdel Rahman, R.: Emotional news affects social judgments independent of perceived media credibility. Social Cognitive and Affective Neuroscience **16**(3), 280–291 (2021)
4. Baumeister, R.F., Bratslavsky, E., Finkenauer, C., Vohs, K.D.: Bad is stronger than good. Review of general psychology **5**(4), 323–370 (2001)
5. Bruns, H., Lewandowsky, S., Pennycook, G., Pantazi, M., Schmid, P., Krawczyk, M.W., Dessart, F.J., Smillie, L.: The role of (trust in) the source of prebunks and debunks of misinformation. evidence from online experiments in four eu countries. https://doi.org/10.31219/osf.io/vd5qt (2023)
6. Brydges, C.R., Gignac, G.E., Ecker, U.K.: Working memory capacity, short-term memory capacity, and the continued influence effect: A latent-variable analysis. Intelligence **69**, 117–122 (2018)
7. Buchanan, T.W.: Retrieval of emotional memories. Psychological bulletin **133**(5), 761 (2007)
8. Ecker, U.K., Antonio, L.M.: Can you believe it? An investigation into the impact of retraction source credibility on the continued influence effect. Memory & Cognition **49**, 631–644 (2021)
9. Ecker, U.K., Lewandowsky, S., Cheung, C.S., Maybery, M.T.: He did it! she did it! no, she did not! multiple causal explanations and the continued influence of misinformation. Journal of memory and language **85**, 101–115 (2015)
10. Ecker, U.K., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L.K., Brashier, N., Kendeou, P., Vraga, E.K., Amazeen, M.A.: The psychological drivers of misinformation belief and its resistance to correction. Nature Reviews Psychology **1**(1), 13–29 (2022)
11. Ecker, U.K., Lewandowsky, S., Swire, B., Chang, D.: Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. Psychonomic Bulletin & Review **18**, 570–578 (2011)
12. Ecker, U.K., Lewandowsky, S., Tang, D.T.: Explicit warnings reduce but do not eliminate the continued influence of misinformation. Memory & Cognition **38**, 1087–1100 (2010)
13. Finucane, M.L., Alhakami, A., Slovic, P., Johnson, S.M.: The affect heuristic in judgments of risks and benefits. Journal of behavioral decision making **13**(1), 1–17 (2000)
14. Gigerenzer, G., Gaissmaier, W.: Heuristic decision making. Annual review of psychology **62**, 451–482 (2011)
15. Hough, A.R., Larue, O.: Exploring memory mechanisms underlying the continued influence effect. In Proceedings of the 22nd International Conference on Cognitive Modeling. Via mathpsych.org/presentation/1605 (2024)
16. Johnson, H.M., Seifert, C.M.: Sources of the continued influence effect: When misinformation in memory affects later inferences. Journal of Experimental Psychology: Learning, Memory, and Cognition **20**(6), 1420–1436 (1994)
17. Juvina, I., Larue, O., Hough, A.: Modeling valuation and core affect in a cognitive architecture: The impact of valence and arousal on memory and decision-making. Cognitive Systems Research **48**, 4–24 (2018)
18. Kahneman, D.: Thinking fast and slow. New York: Farrar, Straus and Giroux (2011)
19. Kahneman, D., Klein, G.: Conditions for intuitive expertise: a failure to disagree. American psychologist **64**(6), 515 (2009)

20. Lewandowsky, S., Ecker, U.K., Seifert, C.M., Schwarz, N., Cook, J.: Misinformation and its correction: Continued influence and successful debiasing. Psychological Science in the Public Interest **13**(3), 106–131 (2012)
21. Martel, C., Mosleh, M., Rand, D.G.: You're definitely wrong, maybe: Correction style has minimal effect on corrections of misinformation online. Media and Communication **9**(1), 120–133 (2021)
22. Mohammad, S.: Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers). pp. 174–184 (2018)
23. Neisser, U., Harsch, N.: Phantom flashbulbs: False recollections of hearing the news about challenger. Affect and accuracy in recall: Studies of "flashbulb" memories **4**, 9–31 (1992)
24. Pachur, T., Hertwig, R., Steinmann, F.: How do people judge risks: Availability heuristic, affect heuristic, or both? Journal of Experimental Psychology: Applied **18**(3), 314 (2012)
25. Phillips, S., Wang, S.Y.N., Carley, K.M., Rand, D., Pennycook, G.: Emotional language reduces belief in false claims (2024)
26. Russell, J.A., Barrett, L.F.: Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. Journal of personality and social psychology **76**(5), 805 (1999)
27. Stanovich, K.E.: Miserliness in human cognition: The interaction of detection, override and mindware. Thinking & Reasoning **24**(4), 423–444 (2018)
28. Susmann, M.W., Wegener, D.T.: The role of discomfort in the continued influence effect of misinformation. Memory & Cognition **50**(2), 435–448 (2022)
29. Swire, B., Ecker, U.K., Lewandowsky, S.: The role of familiarity in correcting inaccurate information. Journal of Experimental Psychology: Learning, Memory, and Cognition **43**(12), 1948–1961 (2017)
30. Vaish, A., Grossmann, T., Woodward, A.: Not all emotions are created equal: the negativity bias in social-emotional development. Psychological bulletin **134**(3), 383 (2008)
31. Wilkes, A., Leatherbarrow, M.: Editing episodic memory following the identification of error. The Quarterly Journal of Experimental Psychology **40**(2), 361–387 (1988)
32. Williamson, J.B., Drago, V., Harciarek, M., Falchook, A.D., Wargovich, B.A., Heilman, K.M.: Chronological effects of emotional valence on the self-selected retrieval of autobiographical memories. Cognitive and Behavioral Neurology **32**(1), 11–15 (2019)
33. Yonelinas, A.P., Ritchey, M.: The slow forgetting of emotional episodic memories: An emotional binding account. Trends in cognitive sciences **19**(5), 259–267 (2015)