Using Cognitive Models to Test Hypotheses for a Misinformation-related Effect

[1]Alexander R. Hough (alexander.hough.1@us.af.mil)
[2]Othalia Larue (othalia.larue@parallaxresearch.org)

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

Abstract

Human's systematic cognitive processes are vulnerable to misinformation-related effects. For example, misleading information can have a lasting influence even after it has been corrected. This continued influence effect (CIE) was found to be resistant to mitigation in experiments. Leading explanations are memory-based, but some include emotion and/or reasoning. However, mixed findings have hindered our understanding of the phenomenon. We argue that cognitive models are uniquely suited to help clarify these mixed findings and theories through specification of underlying mechanisms, testing hypotheses, and identifying why and when mitigations are effective. We start by discussing relevant experimental findings, then present an updated cognitive model of the CIE, compare model fits to two experiments across three model variations, and discuss the results along with recent exploratory analyses with previous experiments.

Keywords: ACT-R; Cognitive modeling; misinformation; continued influence effect; knowledge representation; affect

Introduction

Humans often leverage heuristics that exploit statistical regularities (i.e., cues) in the environment, which can lead to successes (Gigerenzer & Gaissmaier, 2011) and systematic failures (Kahneman, 2011). Failures are usually related to applying a heuristic in the wrong context or when the environment is "hostile" and the diagnostic cues are misleading, limited, or manipulated (Stanovich, 2018). These heuristic vulnerabilities were successfully leveraged to spread misinformation in the past (Lewandowsky et al., 2013) and can now be spread more quickly (Vosoughi et al., 2018). Here, we focus on a specific vulnerability, the continued influence effect (CIE; (Johnson & Seifert, 1994; Lewandowsky et al., 2012)), where misinformation is presented and has a persistent effect on decisions even after it has been corrected.

The CIE

CIE experiments typically present two narratives about scenarios (i.e., events), where one contains misinformation and the second has a correction. In general, corrections reduce, but do not eliminate misinformation reliance (Ecker & Antonio, 2021; Ecker et al., 2017; Johnson & Seifert, 1994). Explanations suggest memory is permanent, but varies in strength due to re-activation or association with different information (Wilkes & Leatherbarrow, 1988). One account focuses on issues with memory retrieval, such as competing memory activations (Ayers & Reder, 1998; Ecker et al., 2010), recency effects (Ecker et al., 2015), or familiarity-based fluency (Ecker et al., 2011). These processes can lead to errors. However, if detected, errors can be avoided through directed retrievals and/or reasoning, leading to more thorough consideration of information (Pennycook & Rand, 2019; Stanovich, 2018). An alternative account involves the preference for mental models that are more complete or coherent, which may involve causal information, believability, or interconnected details (Gentner, 1983; Johnson & Seifert, 1994; Lewandowsky et al., 2012; Wilkes & Leatherbarrow, 1988). Misinformation may be preferred when it is part of a more complete mental model or the correction cannot be integrated, which may create a feeling of discomfort (Ecker et al., 2011; Susmann & Wegener, 2022). Feelings alone can also influence memories by making them easier to recall (Yonelinas & Ritchey, 2015), with a greater effect for negative information (Vaish et al., 2008; Williamson et al., 2019). However, the extent emotion influences memory related to misinformation is still unclear (Phillips et al., 2024).

Research covered a range of correction methods to mitigate and explain the CIE (Prike & Ecker, 2023; Walter & Tukachinsky, 2020), and some individual differences (e.g., Brydges et al. (2018)), however, findings are rather inconsistent. There are likely interactions between the content of events, sources of influence, individual differences (e.g., prior experience), and mitigations (Hough et al., Under review). We argue computational cognitive models could help provide a clearer understanding of the CIE by testing explanations in different contexts and help to determine how, when, and why context and sources of influence interact. Here, we extend a basic computational cognitive model framework (Hough & Larue, 2024) by including an additional model variant and experimental dataset to show differences resulting from degree of encoding, affect, and memory rumination.

Modeling the CIE

We implemented our CIE model within the ACT-R cognitive architecture (Anderson, 2007). ACT-R is a hybrid cognitive architecture with symbolic and sub-symbolic structures, and modules representing systems of the mind. The CIE model uses the goal module for the model's focus and storage of goal-relevant information. The vision module allows visual perception, and the imaginal module serves as temporary working memory. The declarative memory module stores in-

formation as chunks and captures memory dynamics. The procedural memory module uses condition-action rules (i.e., productions) to represent knowledge about how to do things and to drive behavior. Here we focus on declarative memory and affect mechanisms to capture the CIE and compare three model variants we used to fit data from two experiments.

Declarative Memory and Affect Mechanisms

Through chunk activation, ACT-R declarative memory can capture memory retrieval and mental model effects. Chunks are the basic unit of declarative memory. They include slots with values and have an activation value corresponding to the probability and speed of its retrieval in a given situation (Anderson, 2007). Chunks often compete because they match a retrieval request, but the chunk with the highest activation is more likely to be retrieved. Here, we reduced the terms of the standard activation equation to just base level activation, B_i, which represents recency and frequency of chunk use, and activation noise, ε_i, representing variability in memory. 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. If used, the base-level constant, β_i, is a constant offset to the equation.

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

We added valuation, V_i, and arousal, Ar_i, terms to the equation, with their accompanying weights (i.e., *vw* and *aw*). These terms approximate emotion dynamics based on the core affect theory of emotion (Russell & Barrett, 1999), which focuses on feelings underlying emotion using two dimensions: valuation (positive or negative) and arousal (magnitude). Previous work developed a valuation module (Juvina et al., 2018) to compute the valuation, V_i, and arousal, Ar_i, terms. 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. They affect activations and can be used as retrieval cues. This aligns with research suggesting emotional memories are more accessible (Buchanan, 2007) and/or easier to recall (Yonelinas & Ritchey, 2015). For example, if negative affect was associated with a chunk, its activation would increase. It could persist over time and affect decision-making despite the accumulation of conflicting, more neutral, evidence.

Model Descriptions and Processes

In this paper, we compare and contrast three CIE model variants: Memory, Memory-affect, and Updated-memory. Despite having features of both the memory error and mental model explanations above, we did not explicitly create any variant to directly represent them. Variants have features of both, but we argue the Memory and Memory-affect variants better align with the memory encoding/retrieval error explanation, and the Updated-memory with the mental model explanation. The Memory variant was used as a base for the other two model variants. Changes from the base model (i.e., Memory) included parameter values (Table 1), productions (Figure 1), and/or the activation equation (Equation 1). We start by describing the Memory variant and move on to the other two variants.

Table 1: Declarative and valuation parameters for the Ecker and Antonio (2021) experiment and Bruns et al. (2023) experiment (in parentheses if different).

	Memory	Memory-affect	Updated-memory
Retrieval threshold	0	0	0
Base-level decay	0.5	0.5	0.5
Activation noise	0.25	0.25	0.45 (0.25)
Base-level constant	2.5	2.5	0
Declarative finst num	1000	1000	1000
Declarative finst span	1000	1000	1000
Initial valuation	-	1	-
Valuation weight	-	2	-
Valuation learning	-	1	-
Valuation time window	-	0.5	-
Arousal weight	-	1	-
Word affect scaling	-	10 (1)	-
Source affect	-	0.4-1.4 (5)	-

Memory The Memory variant used six declarative memory parameters and the first three were previously discussed: 1) base-level constant, β_i, was set to 2.5 (0 is default), 2) base-level decay, d was set to the default of .5, 3) activation noise, ε_i, was set to .25 (recommended range is .2-.8), 4) Retrieval threshold restricts which chunks can be retrieved based on activation and was set at the default value of 0. The last two parameters: 5) declarative finsts that sets number of items marked as retrieved, and 6) declarative finst span that sets the time items remain marked, were both essentially shut off by setting higher than recommended (i.e., 1000) to prevent an endless loop of retrievals.

Table 2: Full football scenario (Ecker & Antonio, 2021) with misinformation in bold and correction italicized.

Football Scenario
Stockholm FC star player Emil Larsson will not be available for the opening match of the Swedish Superettan league season.
Larsson is believed to have tested positive to performance enhancing drugs.
The 27 year-old signed with Stockholm at the beginning of the 2012 season and has since become one of their strongest players.
Larsson scored 23 goals in his first season with Stockholm, and gave 11 assists.
Club president Asgeir Soerenssen, who recently refused several lucrative offers to sell Larsson, was not available for comments.
Recent acquisition Lucas Johansson is predicted to take Larsson's position in the opening round match against arch-rival Goteborg SK.
Oliver Lindgren, SOURCE, stated that "I do not believe that Larsson has engaged in drug use."
Under recently introduced rules, players suspended for drug-related offenses will not receive pay throughout the duration of their suspension.

Table 3: Football scenario parsed into word pairs/chunks and categorized as neutral, misinformation, and correction.

Type	Word Pairs/chunks
Neutral	(Stockholm star-player) (star-player Emil-Larsson) (Emil-Larsson not-available) (not-available opening-match) (club-president Asgeir-Soerenssen) (Asgeir-Soerenssen refused-sale) (refused-sale Emil-Larsson) (Stockhold aquired) (aquired Lucas-Johansson) (Lucas-Johansson replace) (replace Asgeir-Soerenssen) (performance-drugs suspended) (suspended no-pay)
Misinformation	(Emil-Larsson tested) (tested positive) (positive performance-drugs)
Correction	(Oliver-Lindgren source) (source statement) (statement I-do-not-believe-that-Larsson-has-engaged-in-drug-use) (Emil-Larsson not-engaged) (not-engaged performance-drugs)

The Memory variant focused on retrievals and chunk activation changes based on declarative memory dynamics to capture competition between misinformation and corrections. There are two main sets of processes that: 1) read text and create chunks, and 2) navigate chunks in memory by chaining them together based on associations and then giving a simulated summary of the content (Figure 1). Before information was presented to the model, we parsed the experimental materials (Table 2) into word pairs (Table 3). We provided an in-context learning example to ChatGPT (e.g., Romero et al. (2023)) and had to use the output as a guide to manually generate words pairs for most of the materials due to dramatic changes to ChatGPT (we are working on a better long-term solution). During the first set of processes, the model is presented with word pairs in Table 3 one at a time. The model finds, attends, and reads a word. If the model can't associate words (i.e., only one word was read), then searches for another word to read (i.e., keep-reading). Once two words are read, the model attempts to retrieve (i.e., retrieve-assoc) a matching chunk and if not possible, it creates a new chunk (i.e., create-assoc). Once all experimental materials are read and encoded into memory, the model starts the memory chaining processes. It starts by openly retrieving a chunk for a random scenario (i.e., retrieve-scenario-info) and encoding it (i.e., encode-scenario-info). Next, it attempts to find the root or start of chunk chain through back-chaining (i.e., find-chain-root) using the first word of the encoded chunk (e.g., **tested** positive) and trying to retrieve a chunk with a matching second word (e.g., Emil-Larsson **tested**).

If a chunk was retrieved, it is encoded (i.e., encode-back-chain), otherwise the root is considered found and back-chaining stops (i.e., found-root). Next, the model attempts to parallel chain using the first word of the root (e.g., **Emil-Larsson** not-available) to find a chunk with the same first word (e.g., **Emil-Larrson** tested). Parallel chaining prevents some chunks from being neglected and can link information from different chains. If such a chunk is retrieved, it is encoded (i.e., encode-parallel-chain) and if not, forward chaining is started (i.e., start-forward-chain). The model uses the second word of the currently encoded chunk (e.g., Emil-Larsson **tested**) to find a chunk with a matching first word (e.g., **tested** positive). If a chunk is retrieved, it is encoded

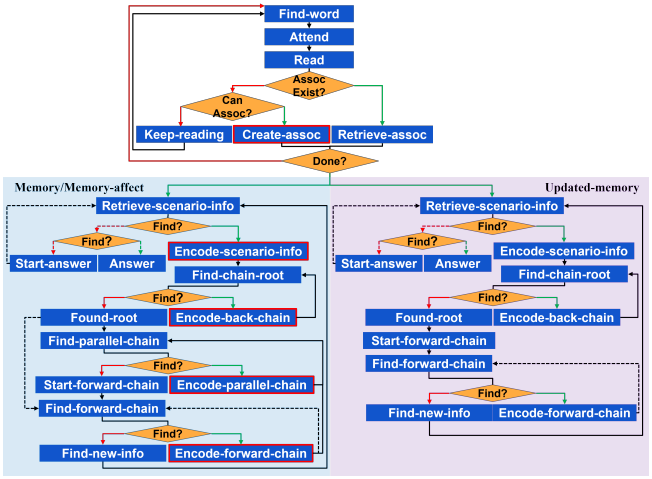

Figure 1: CIE Model Framework Processes.

(i.e., encode-forward-chain) and if not, the chain is considered complete and the model starts a new chain (i.e., find-new-info). If the open retrieval (i.e., retrieve-scenario-info) fails because chunks are marked as recently retrieved and/or their activations are below the retrieval threshold, memory navigation is complete and the model prepares a summary answer (i.e., start-answer). The model uses the same navigation processes, excluding parallel chaining (see dotted lines), to find the most active chunk and the chain it belongs to; which is given as the answer (i.e., answer).

Memory-affect The Memory-affect variant included valuation and arousal terms that were added to the activation equation (equation 1), and affect was associated with content when reading and reinforced during retrieval/encoding (red borders in Figure 1). We used valance and arousal values scaled from 0-1 (i.e., below .5 is negative and above is positive) from the 20,000 word NRC Emotion Lexicon Mohammad (2018) to specify affect associated with words. We collapsed valence and arousal values into a single positive reward (i.e., greater than 0) so that negative valence had a greater value, $arousal * (1 - valence)$. We scaled values (Table 1) so that they were high enough to influence chunk activations. We specified source affect based on source credibility and trustworthiness ratings collected in the two experiments we used for model fitting (Bruns et al., 2023; Ecker & Antonio, 2021). These values ranged from .4-1.4 across the five source conditions for one experiment (Ecker & Antonio, 2021) and were set to 5 for a single source for the other (Bruns et al., 2023) Based on previous work with the valuation module (Juvina et al., 2018), we used rewards to update both valuation and arousal terms in the activation equation. These rewards were scaled differently based on valance and arousal values for words (0-1 * scaling value) and ratings for information sources for the two experiments (Table 1). To align with previous research (Vaish et al., 2008; Williamson et al., 2019), words with negative valence increased activations more than positive. We 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).

Updated-memory We leveraged declarative finsts to restrict retrievals to chunks not recently retrieved for the open retrieval (i.e., retrieve-scenario-info) and during backward, parallel, and forward chunk chaining to prevent infinite chaining loops with additional datasets. The downside was that some chunk chains were unable to complete because some of their chunks had already been retrieved. Therefore, interconnected chunks (i.e., chunks that can chain or link to many other chunks) were not retrieved as much as originally intended and their activation was less affected by memory navigation. We argue that this results in a less coherent mental model. The Updated-memory variant was developed to address this issue. Our initial solution, which still had limitations that we address in the discussion, was to modify the open retrieval (i.e., retrieve-scenario-info) that started the process for each chain and to reset declarative finsts (i.e., chunks marked as retrieved) at the end of each chain. We created another list of non-retrieved chunks so that the model would retrieve each chunk for a given scenario and find the chain it belonged to. The result is that more interconnected chunks (i.e., are present in more chains) are retrieved significantly more. In addition, we eliminated the parallel chaining productions (Figure 1), set the base-level constant, β_i, to the default of 0, and increased activation noise to .45 (Table 1).

Ecker and Antonio (2021) Experiment 1

In experiment 1 from Ecker and Antonio (2021), 53 participants (62% female and mean age 18.6) from the University of Western Australia read narratives about six scenarios (i.e. anti-viral drug, fishing restrictions, food additives, football scandal, joint condition treatments, and water contamination) with embedded misinformation and corrections. Each scenario had a different correction source that varied in "quality" (i.e., no retraction [NoR], low expertise/trust [LELT], low expertise/high trust [LEHT], high expertise/low trust [HELT], high expertise/trust [HEHT], and highest expertise/trust [HEHT+]), and was counterbalanced across six versions. Participants answered 10 open-ended and interference questions to comprise a misinformation reliance score, and rated beliefs in misinformation and corrections. Here, we focused on belief scores and approximated them by calculating the average activation for misinformation and correction chunks. We normalized belief scores and activation for misinformation by dividing values for misinformation by the sum of misinformation and correction values. We simulated 60 participants (i.e., 10 for each counterbalancing version) for each model variant. We first compared model fits across narrative scenarios using a correlation and root mean squared error (i.e., *RMSE*), however, we emphasize *RMSE* as the main indicator of fit. The Updated-memory variant had the lowest *RMSE*, $r(8) = 0.11, p = 0.86, RMSE = 0.04$, followed by the Memory-affect variant, $r(8) = 0.65, p = 0.24, RMSE = 0.05$, and the Memory variant, $r(8) = -0.10, p = 0.87, RMSE = 0.10$ (Figure 2).

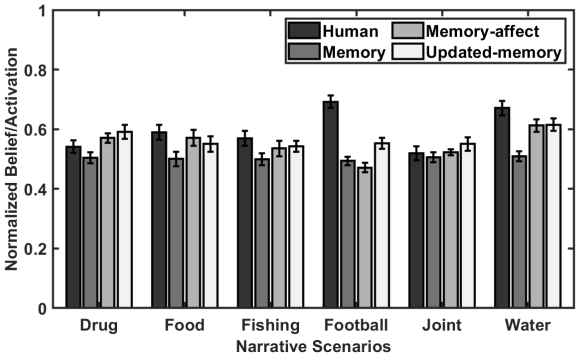

Figure 2: Misinformation belief/activation across scenarios for human and model data. Error bars are SEM.

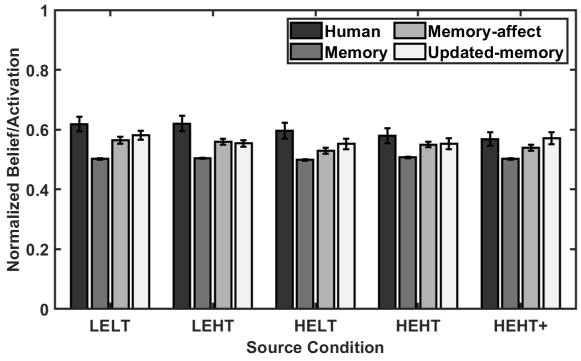

Figure 3: Misinformation belief/activation across source conditions for human and model data. Error bars are SEM

Next, we compared model variants for source conditions, which were arranged from least to highest quality. We note that only the Memory-affect model is sensitive to the quality of source information. The Updated-memory variant had the lowest *RMSE*, $r(10) = 0.29, p = 0.57, RMSE = 0.07$, followed the the Memory-affect variant, $r(10) = -0.09, p = 0.86, RMSE = 0.10$, and the Memory variant, $r(10) = -0.32, p = 0.53, RMSE = 0.12$ (Figure 3).

Bruns et al. (2023) Experiment

In Bruns et al. (2023), 2,614 participants were recruited from Europe and read one of three claims (i.e., misinformation) about climate change and a separate correction article that came before (i.e., prebunk) or after (i.e., debunk) the claim with or without a source. Similar to experiment 1, sources had associated affect. There was no associated affect for no source conditions. As there were three claims and five conditions (no correction control, prebunk no source, prebunk source, debunk no source, and debunk source), there were 15 separate conditions. 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, which is comparable to beliefs, and fit the models to reduction in agreement. Reduction in agreement was calculated by subtracting the control group misinformation agreement [activation] from misinformation agreement [activation] for all conditions (i.e., prebunk and debunks with and without source information). We simulated 150 participants (i.e., 10 for each of the 15 conditions) for all model variants. The Updated-memory model has the lowest *RMSE*, $r(6) = 0.93, p = 0.07, RMSE = 0.06$, followed by the Memory-affect variant, $r(6) = 0.70, p = 0.30, RMSE = 0.07$, and the Memory variant, $r(6) = 0.88, p = 0.12, RMSE = 0.24$ (Figure 4).

Discussion

We argued computational cognitive models could help resolve mixed findings and ambiguity with CIE explanations, and compared three CIE model variants. The Memory variant included creation of chunks and limited memory navigation that influenced chunk activation based on retrieval frequency. The Memory-affect variant was an extension that included affect associated with words and information sources at chunk creation, which was reinforced during retrieval. The Updated-memory variant eliminated some processes, better aligned with default parameter values, and increased the extent of memory navigation affecting activation of the more interconnected chunks within narratives.

We argue that the Memory and Memory-affect variants best aligned with the memory encoding/retrieval error explanation and the Updated-memory aligned best with the mental model explanation. Although the Updated-memory model was only slightly better than the Memory-affect variant, it was simpler. These results suggest misinformation beliefs in the two experiments may be best explained by the mental model explanation. These findings also raise a concern for CIE experiments, as some results might be due to the wording of the narrative, which should be considered when testing manipulations to reduce the CIE. However, models developed to directly test specific explanations or hypotheses would provide more clear conclusions. Despite successes, we discuss several limitations that may limit generalization of results.

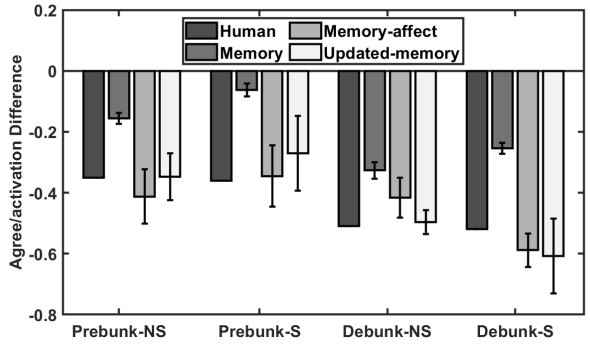

Figure 4: Misinformation agreement/activation difference from control (no correction) across correction type and source conditions for human and model data. Error bars are SEM.

Limitations and Future Work

Our current work has several limitations: 1) We manually generated predicates, 2) we focused on declarative memory with syntactic chunk chaining, 3) we focused on belief and agreement because inferential reasoning was too complex for our current model, 4) we simplified valuation and arousal terms to a single reward value that favored negative affect, and 5) we fit data for only two experiments, and 6) models were developed and compared to theoretical explanations post-hoc.

Misinformation is often in text format, which is a challenge for modeling. Our initial strategy was to parse the scenarios using ChatGPT, but we had to manually generate predicates for most scenarios after ChatGPT was updated. The parsing and subsequent representation of written content has a large impact on a model's behavior, and in our case, the model is dependent on the structure of representations to chain and connect information. We are currently finding a more stable language processing method and are currently improving chaining by: 1) treating words (i.e., rather than word pairs) as chunks, 2) using spreading activation to reduce memory retrievals, and 3) using semantic and affective information rather than word matching to connect information. This will enable comparing spreading activation and memory navigation mechanisms, more holistic knowledge representation, and extend question answering capabilities.

We focused on beliefs and agreement, and approximated them by the average activation of misinformation and correction chunks. Most experiments include several open-ended and inference questions. Our planned improvements to language parsing and knowledge representation will extend the model's behavior capabilities, but we also need to add some reasoning ability and/or causal knowledge.

Our emotion implementation used a previous valuation module to update chunk activation through rewards, which combined valuation and arousal values and favored negative affect. Rather than rewards, we plan to associate valuation and arousal values with words to allow for differential influence of both negative and positive affect.

We fit our model to only two experiments, which limits generalization. Our improvements mentioned above should increase the ability of the model to handle different content and provide different types of responses. In addition, this will enable more direct testing of explanations and hypotheses. Our goal is to capture fundamental cognitive processes and knowledge representations underlying the CIE and related phenomena.

Conclusion

Overall, we were able to simulate the CIE with and without emotion for two datasets. Planned future improvements

will enable exploration of social factors, group interactions, and theoretical explanations across experiments and datasets. Our current work, mixed findings, and interactions between sources of influence, content, and mitigations suggest we may benefit from a methodological re-framing. We suggest treating sources of influence and mitigations as cues could enable predictions similar to multiple-cue decision making (Lee et al., 2019; Weber & Johnson, 2009), and some literature has already suggested source information (Traberg et al., 2024) and emotion (Phillips et al., 2024) serve as cues.

Acknowledgments

This research was supported by the U. S. Air Force Research Laboratory's 711^{th} Human Performance Wing, Cognition and Modeling Branch. Contents were reviewed and approved for public release (AFRL-2025-1039). The views expressed are those of the authors and do not reflect the official guidance or position of the United States Government, the U.S. Department of Defense, the U.S. Air Force, or any of their subsidiaries or employees.

References

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

Ayers, M. S., & Reder, L. M. (1998). A theoretical review of the misinformation effect: Predictions from an activation-based memory model. *Psychonomic Bulletin and Review*, *5*, 1–21.

Bruns, H., Dessart, F. J., Krawczyk, M. W., Lewandowsky, S., Pennycook, G., Pantazi, M., ... Smillie, L. (2023). Investigating the role of source and source trust in prebunks and debunks of misinformation in online experiments across four eu countries. *Scientific Reports*, *20723*.

Brydges, C. R., Gignac, G. E., & Ecker, U. K. (2018). Working memory capacity, short-term memory capacity, and the continued influence effect: A latent-variable analysis. *Intelligence*, *69*, 117–122.

Buchanan, T. W. (2007). Retrieval of emotional memories. *Psychological Bulletin*, *133*(5), 761.

Ecker, U. K., & Antonio, L. M. (2021). Can you believe it? An investigation into the impact of retraction source credibility on the continued influence effect. *Memory & Cognition*, *49*, 631–644.

Ecker, U. K., Hogan, J. L., & Lewandowsky, S. (2017). Reminders and repetition of misinformation: Helping or hindering its retraction? *Journal of Applied Research in Memory and Cognition*, *6*(2), 185–192.

Ecker, U. K., Lewandowsky, S., Cheung, C. S., & Maybery, M. T. (2015). 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.

Ecker, U. K., Lewandowsky, S., Swire, B., & Chang, D. (2011). Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. *Psychonomic Bulletin & Review*, *18*, 570–578.

Ecker, U. K., Lewandowsky, S., & Tang, D. T. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition*, *38*, 1087–1100.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*(2), 155–170.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451–482.

Hough, A. R., Arakal, A., & Larue, O. (Under review). Scenarios impact the continued influence effect. *Computational and Mathematical Organization Theory*.

Hough, A. R., & Larue, O. (2024). Exploring memory mechanisms underlying the continued influence effect. *In Proceedings of the 22nd International Conference on Cognitive Modeling. Via mathpsych.org/presentation/1605*.

Johnson, H. M., & Seifert, C. M. (1994). 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.

Juvina, I., Larue, O., & Hough, A. (2018). 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.

Kahneman, D. (2011). *Thinking fast and slow*. New York: Farrar, Straus and Giroux.

Lee, M. D., Gluck, K. A., & Walsh, M. M. (2019). Understanding the complexity of simple decisions: Modeling multiple behaviors and switching strategies. *Decision*, *6*(4), 335.

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131.

Lewandowsky, S., Stritzke, W. G., Freund, A. M., Oberauer, K., & Krueger, J. I. (2013). Misinformation, disinformation, and violent conflict: From iraq and the "war on terror" to future threats to peace. *American Psychologist*, *68*(7), 487–501.

Mohammad, S. (2018). 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)*, 174–184.

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50.

Phillips, S., Wang, S. Y. N., Carley, K. M., Rand, D., & Pennycook, G. (2024). Emotional language reduces belief in false claims. *OSF Preprint*.

Prike, T., & Ecker, U. K. (2023). Effective correction of misinformation. *Current Opinion in Psychology*, 101712.

Romero, O. J., Zimmerman, J., Steinfeld, A., & Tomasic, A. (2023). Synergistic integration of large language models and cognitive architectures for robust ai: An exploratory analysis. *Proceedings of the AAAI Symposium Series*, *2*(1), 396–405.

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of Personality and Social Psychology*, *76*(5), 805.

Stanovich, K. E. (2018). Miserliness in human cognition: The interaction of detection, override and mindware. *Thinking & Reasoning*, *24*(4), 423–444.

Susmann, M. W., & Wegener, D. T. (2022). The role of discomfort in the continued influence effect of misinformation. *Memory & Cognition*, *50*(2), 435–448.

Traberg, C. S., Harjani, T., Roozenbeek, J., & van der Linden, S. (2024). The persuasive effects of social cues and source effects on misinformation susceptibility. *Scientific Reports*, *14*(1), 4205.

Vaish, A., Grossmann, T., & Woodward, A. (2008). Not all emotions are created equal: the negativity bias in social-emotional development. *Psychological Bulletin*, *134*(3), 383.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151.

Walter, N., & Tukachinsky, R. (2020). A meta-analytic examination of the continued influence of misinformation in the face of correction: How powerful is it, why does it happen, and how to stop it? *Communication Research*, *47*(2), 155–177.

Weber, E. U., & Johnson, E. J. (2009). Mindful judgment and decision making. *Annual Review of Psychology*, *60*(1), 53–85.

Wilkes, A., & Leatherbarrow, M. (1988). Editing episodic memory following the identification of error. *The Quarterly Journal of Experimental Psychology*, *40*(2), 361–387.

Williamson, J. B., Drago, V., Harciarek, M., Falchook, A. D., Wargovich, B. A., & Heilman, K. M. (2019). Chronological effects of emotional valence on the self-selected retrieval of autobiographical memories. *Cognitive and Behavioral Neurology*, *32*(1), 11–15.

Yonelinas, A. P., & Ritchey, M. (2015). The slow forgetting of emotional episodic memories: an emotional binding account. *Trends in Cognitive Sciences*, *19*(5), 259–267.