AFRL

MODELING MISINFORMATION-RELATED EFFECTS: SUCCESSES AND CHALLENGES

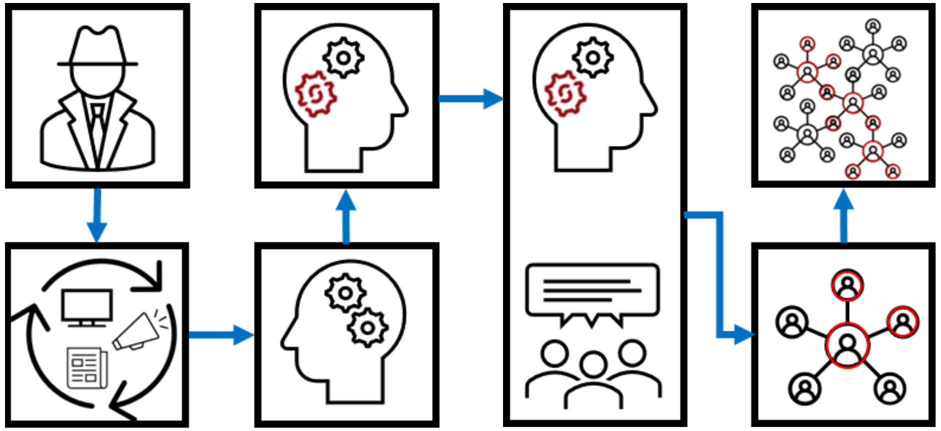
ALEXANDER R HOUGH[1], OTHALIA LARUE[2]

AIR FORCE RESEARCH LABORATORY[1], PARALLAX ADVANCED RESEARCH[2]

23 JULY 2024

Motivation – Modeling Cognitive Warfare Phenomena with ACT-R

- Mixed findings [1-2] and gaps [3]
  - Understand misinformation-related effects: cognition, emotion, & social
  - Scaling individual → small group → social network
  - Assessing potential vulnerabilities and mitigations
- Add to current research – extend to realistic scenarios



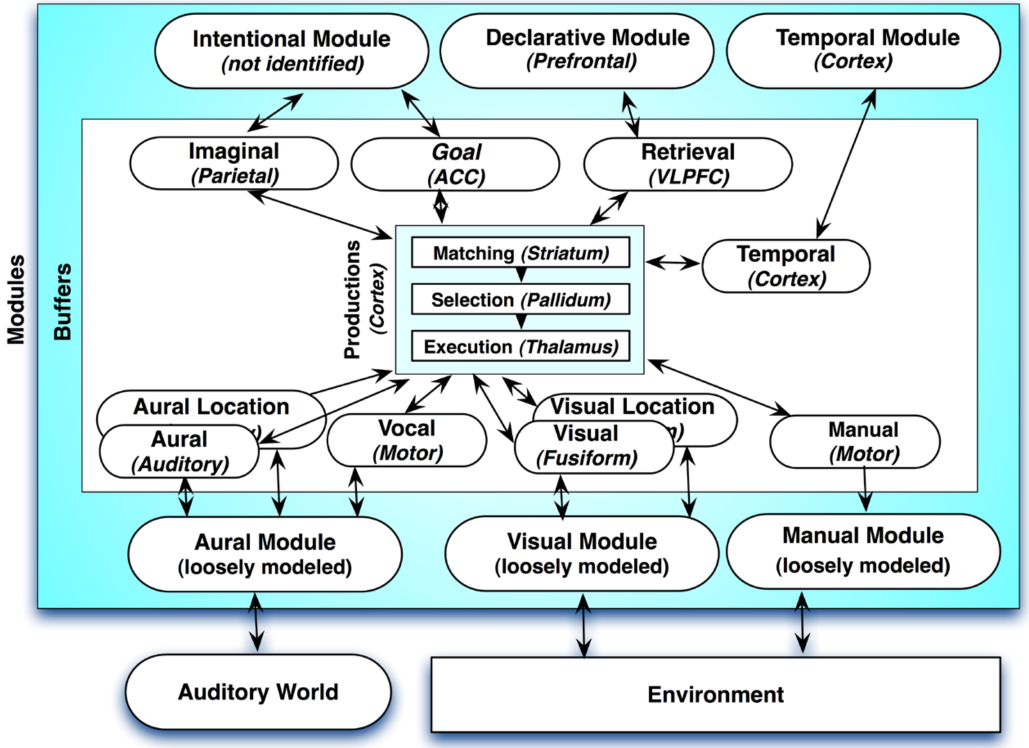
Clip art created by Alex Hough

Specific Efforts:

- **Modeling the continued influence effect (CIE) with ACT-R [4-6]**
- Integration of personal and social beliefs/values with ACT-R [7]
- (Mis)Information spread in social networks with ACT-R + ABM [6]

Why Use ACT-R?

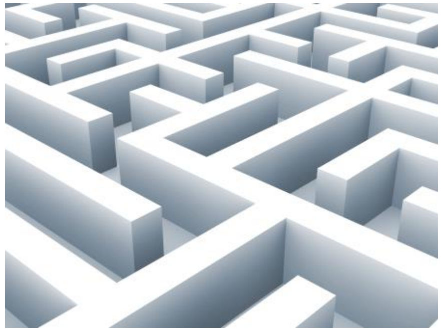
- Cognitively plausible & scientifically validated [8-9]
- Interaction cognitive processes
  - Memory
  - Attention
  - Biases
  - Emotion
  - Social influence (extension)
- Predict and explain behavior
- Small groups
- Human representation in large simulations

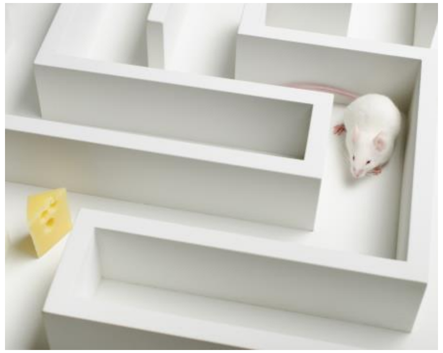


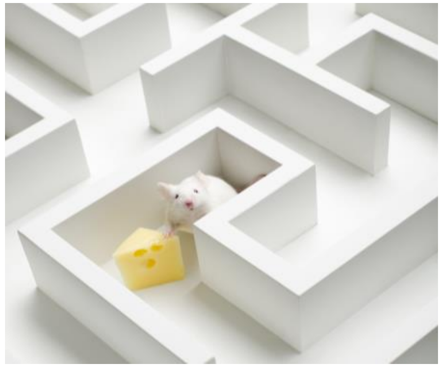
ACT-R structure from [5]

Challenges and Open Questions

- Challenges
  - Processing text to chunks
  - Approximating behaviors
  - Emotion and social influence values
- Technical Questions
  - Pre-processing text
  - Question answering
  - Emotion/social mechanisms
- Theoretical Questions
  - Information weighting/ affect
  - Mental representation → answer questions
  - Sensemaking – similarities, semantics…
  - Interpreting information sources





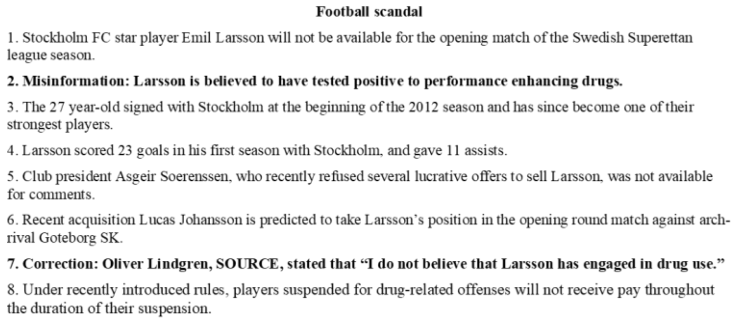


Stock Images (PowerPoint)

Modeling the CIE at Individual Level

CIE Task Structure

- CIE research
  - Robust in lab & mitigations can reduce 50% [9]
- 1st CIE task [11]
  - One article: misinfo + correction
  - Scenarios (6) & source conditions (6)
- 2nd CIE task [12]
  - Two separate articles for misinfo + correction
  - Prebunks, debunks, none (control)
  - Source or no source



Football scandal

1. Stockholm FC star player Emil Larsson will not be available for the opening match of the Swedish Superettan league season.

2. Misinformation: Larsson is believed to have tested positive to performance enhancing drugs.

3. The 27 year-old signed with Stockholm at the beginning of the 2012 season and has since become one of their strongest players.

4. Larsson scored 23 goals in his first season with Stockholm, and gave 11 assists.

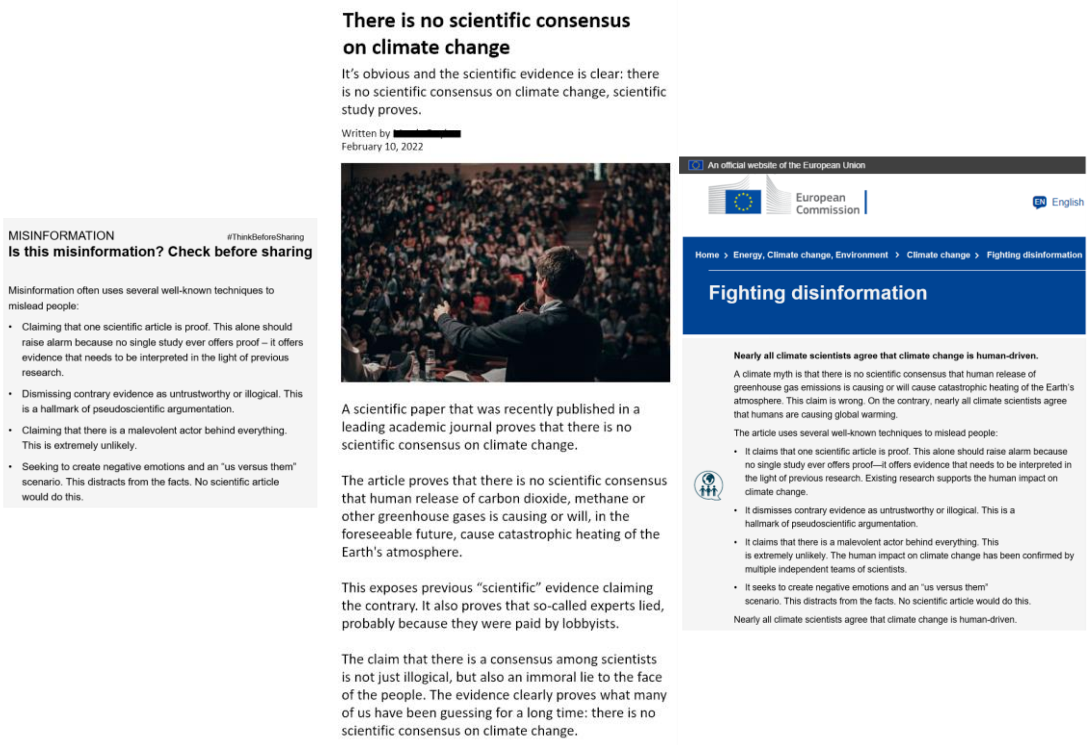
5. Club president Asgeir Soerenssen, who recently refused several lucrative offers to sell Larsson, was not available for comments.

6. Recent acquisition Lucas Johansson is predicted to take Larsson's position in the opening round match against arch-rival Goteborg SK.

7. Correction: Oliver Lindgren, SOURCE, stated that "I do not believe that Larsson has engaged in drug use."

8. Under recently introduced rules, players suspended for drug-related offenses will not receive pay throughout the duration of their suspension.

Example article from [5]



MISINFORMATION #ThinkBeforeSharing

Is this misinformation? Check before sharing

Misinformation often uses several well-known techniques to mislead people:

- Claiming that one scientific article is proof. This alone should raise alarm because no single study ever offers proof – it offers evidence that needs to be interpreted in the light of previous research.
- Dismissing contrary evidence as untrustworthy or illogical. This is a hallmark of pseudoscientific argumentation.
- Claiming that there is a malevolent actor behind everything. This is extremely unlikely.
- Seeking to create negative emotions and an "us versus them" scenario. This distracts from the facts. No scientific article would do this.

There is no scientific consensus on climate change

It's obvious and the scientific evidence is clear: there is no scientific consensus on climate change, scientific study proves.

Written by [redacted]
February 10, 2022

A scientific paper that was recently published in a leading academic journal proves that there is no scientific consensus on climate change.

The article proves that there is no scientific consensus that human release of carbon dioxide, methane or other greenhouse gases is causing or will, in the foreseeable future, cause catastrophic heating of the Earth's atmosphere.

This exposes previous "scientific" evidence claiming the contrary. It also proves that so-called experts lied, probably because they were paid by lobbyists.

The claim that there is a consensus among scientists is not just illogical, but also an immoral lie to the face of the people. The evidence clearly proves what many of us have been guessing for a long time: there is no scientific consensus on climate change.

An official website of the European Union

European Commission | EN English

Home › Energy, Climate change, Environment › Climate change › Fighting disinformation

Fighting disinformation

Nearly all climate scientists agree that climate change is human-driven.

A climate myth is that there is no scientific consensus that human release of greenhouse gas emissions is causing or will cause catastrophic heating of the Earth's atmosphere. This claim is wrong. On the contrary, nearly all climate scientists agree that humans are causing global warming.

The article uses several well-known techniques to mislead people:

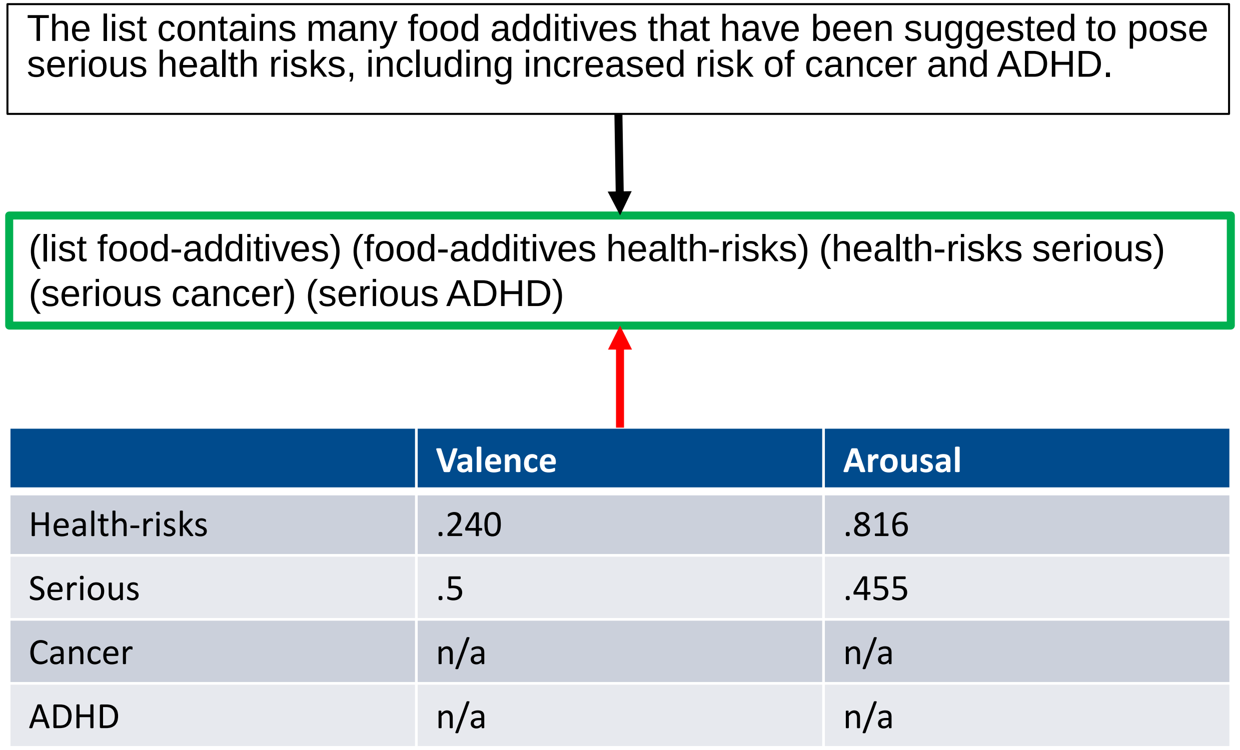
- It claims that one scientific article is proof. This alone should raise alarm because no single study ever offers proof—it offers evidence that needs to be interpreted in the light of previous research. Existing research supports the human impact on climate change.
- It dismisses contrary evidence as untrustworthy or illogical. This is a hallmark of pseudoscientific argumentation.
- It claims that there is a malevolent actor behind everything. This is extremely unlikely. The human impact on climate change has been confirmed by multiple independent teams of scientists.
- It seeks to create negative emotions and an "us versus them" scenario. This distracts from the facts. No scientific article would do this.

Nearly all climate scientists agree that climate change is human-driven.

Example of prebunk, misinformation, and debunk articles from [6]

CIE Task Structure – Our Modeling Approach

- Content – paragraphs of text
  - Parse into word-pair chunks [13]
  - Affect - values from database [14]
  - Meaning – not included…yet
- Memory – chunks
  - Narratives represented as chains
  - Navigate and chain – activations
- Behavior - answering questions
  - Summary – most active chunk and its chain
  - Beliefs – activations of chunks or information type



The list contains many food additives that have been suggested to pose serious health risks, including increased risk of cancer and ADHD.

(list food-additives) (food-additives health-risks) (health-risks serious) (serious cancer) (serious ADHD)

	Valence	Arousal
Health-risks	.240	.816
Serious	.5	.455
Cancer	n/a	n/a
ADHD	n/a	n/a

Tables created by Alex Hough

CIE Model

Memory and Affect

- Short activation w/ core affect (valuation)
  - $A_i \; = B_i + \varepsilon_i + (V_i \, {*}\mathrm{vw}) + (Ar_i \, {*}\mathrm{aw})$
  - $V_i(j) = V_i(j-1) + \mathrm{av}[R_i(j) - V_i(j-1)]$
  - $Ar_i\,(j) = abs(V_i(j))$
- Six declarative parameters

  1) $rt = 0$ 2) $\boldsymbol{blc = 2.5}$

  3) $bll(d) = .5$ 4) $\varepsilon = .25$

  5) $\boldsymbol{declarative - num - finsts = 100}$

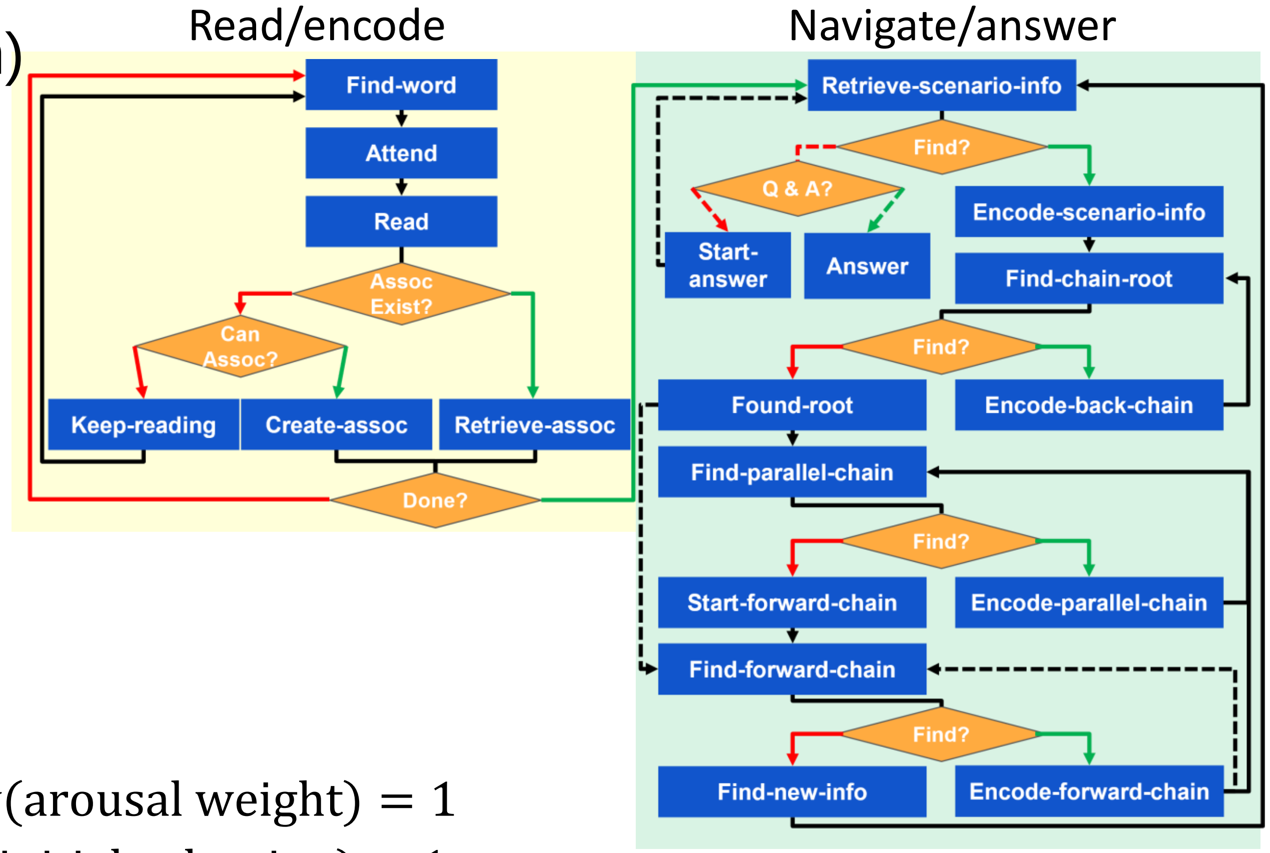
  6) $\boldsymbol{declarative - finst - span = 100}$
- Six valuation parameters

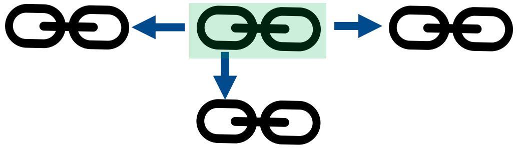
  1) vw (valuation weight) = 2 2) aw(arousal weight) = 1

  3) av (valuation learn rate)= 1 4) iv(initial valuation) = 1

  5) $vtw(valuation\ time\ window) = \ .5$

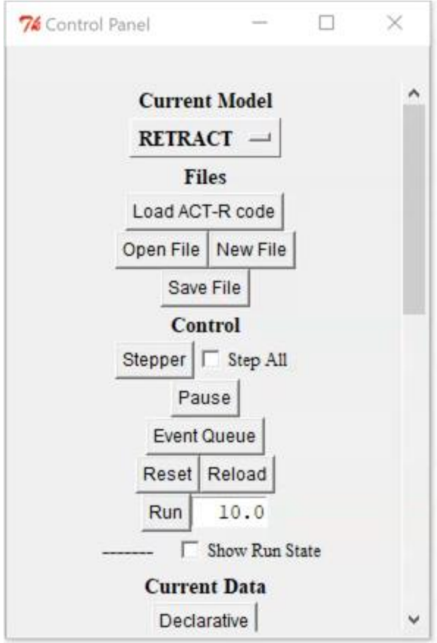
Processes

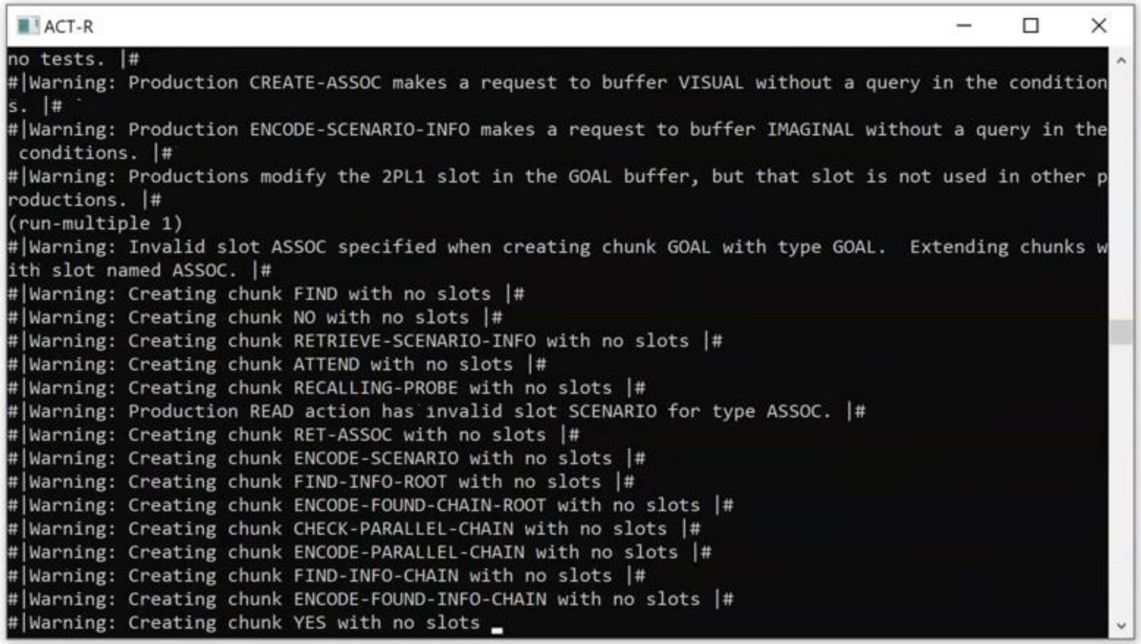




Figures from [4-5]

CIE Model - Demo

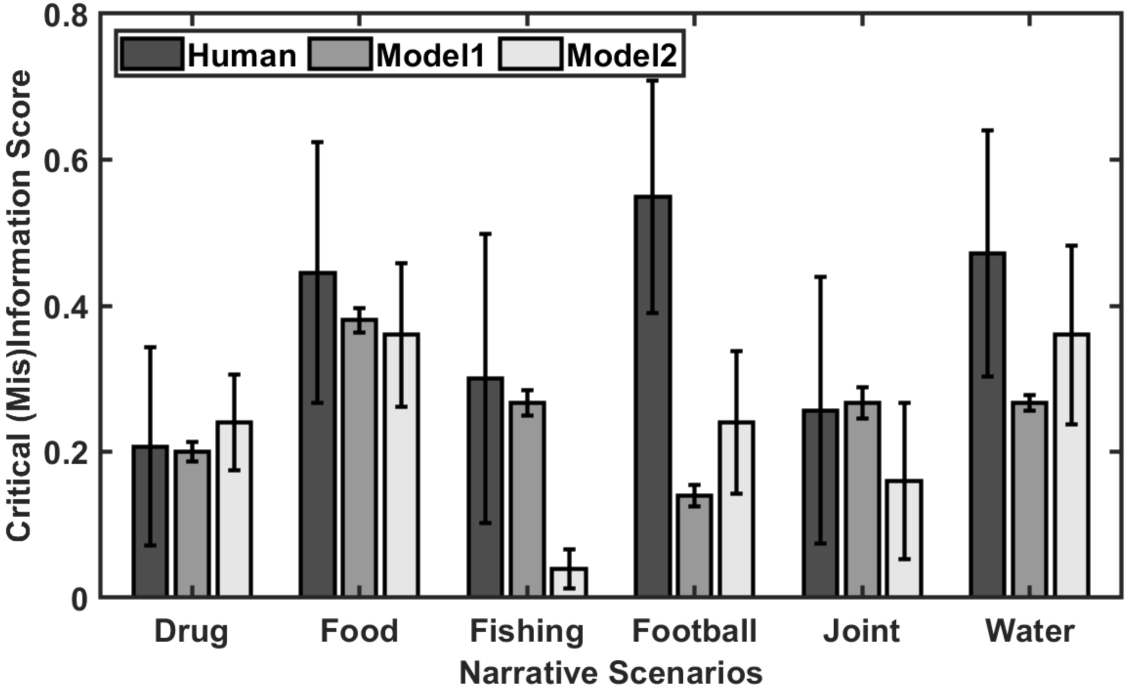




```
no tests. |#
#|Warning: Production CREATE-ASSOC makes a request to buffer VISUAL without a query in the condition
s. |#
#|Warning: Production ENCODE-SCENARIO-INFO makes a request to buffer IMAGINAL without a query in the
 conditions. |#
#|Warning: Productions modify the 2PL1 slot in the GOAL buffer, but that slot is not used in other p
roductions. |#
(run-multiple 1)
#|Warning: Invalid slot ASSOC specified when creating chunk GOAL with type GOAL.  Extending chunks w
ith slot named ASSOC. |#
#|Warning: Creating chunk FIND with no slots |#
#|Warning: Creating chunk NO with no slots |#
#|Warning: Creating chunk RETRIEVE-SCENARIO-INFO with no slots |#
#|Warning: Creating chunk ATTEND with no slots |#
#|Warning: Creating chunk RECALLING-PROBE with no slots |#
#|Warning: Production READ action has invalid slot SCENARIO for type ASSOC. |#
#|Warning: Creating chunk RET-ASSOC with no slots |#
#|Warning: Creating chunk ENCODE-SCENARIO with no slots |#
#|Warning: Creating chunk FIND-INFO-ROOT with no slots |#
#|Warning: Creating chunk ENCODE-FOUND-CHAIN-ROOT with no slots |#
#|Warning: Creating chunk CHECK-PARALLEL-CHAIN with no slots |#
#|Warning: Creating chunk ENCODE-PARALLEL-CHAIN with no slots |#
#|Warning: Creating chunk FIND-INFO-CHAIN with no slots |#
#|Warning: Creating chunk ENCODE-FOUND-INFO-CHAIN with no slots |#
#|Warning: Creating chunk YES with no slots _
```

CIE Model – Exp 1 CI-score Results [new] - Not So Good

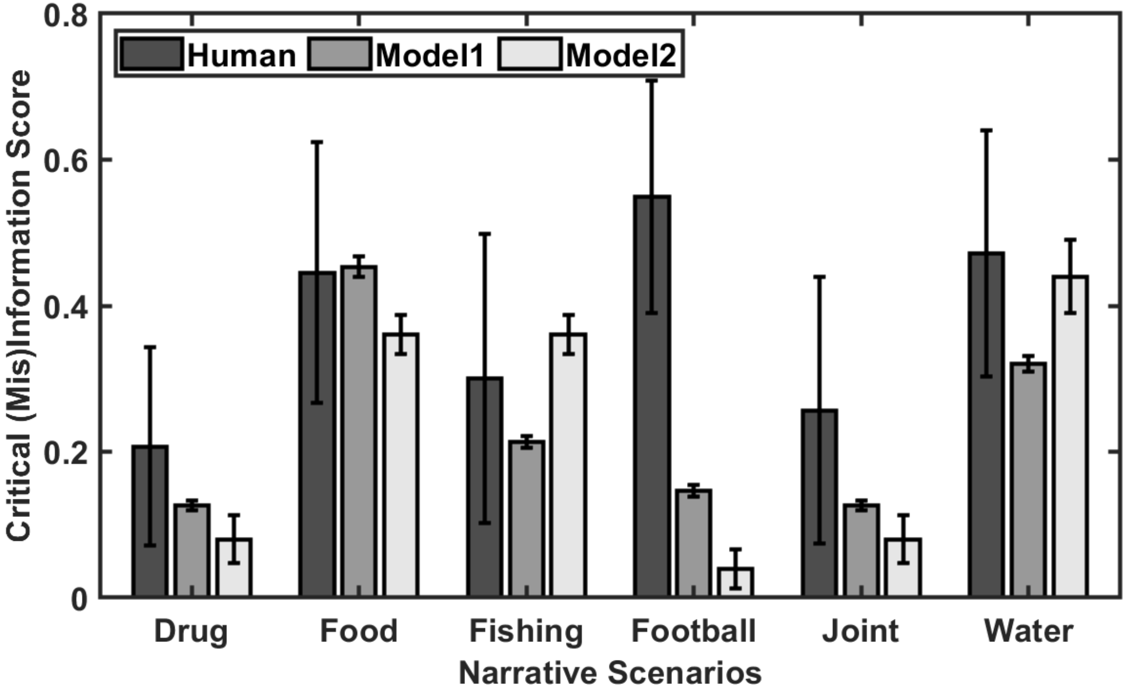
CI scores – answer based



Model1: *r*(10) = -0.06, *p* = 0.91, *RMSE* = 0.19

Model2: *r*(10) = 0.53, *p* = 0.28, *RMSE* = 0.18

CI scores – Top 5 chunks



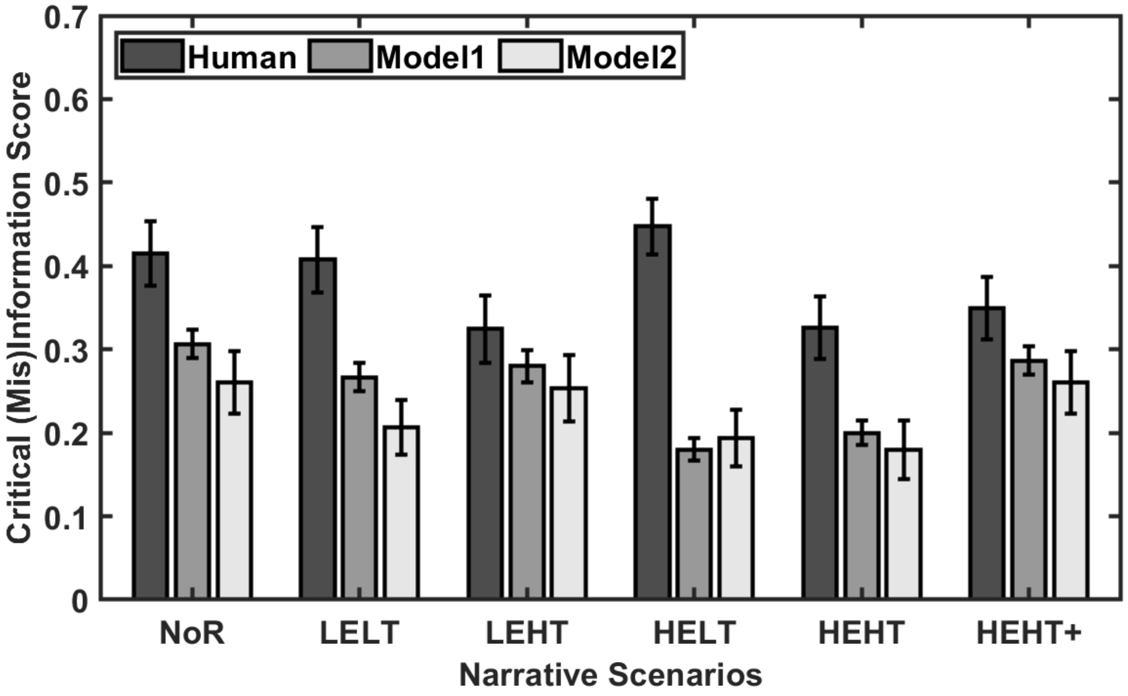
Model1: *r*(10) = 0.46, *p* = 0.36, *RMSE* = 0.19

Model2: *r*(10) = 0.24, *p* = 0.64, *RMSE* = 0.23

DISTRIBUTION A: APPROVED FOR PUBLIC RELEASE. CASE NUMBER:AFRL-2024-3673

CIE Model – Exp 1 CI-score Results [new] - Not So Good

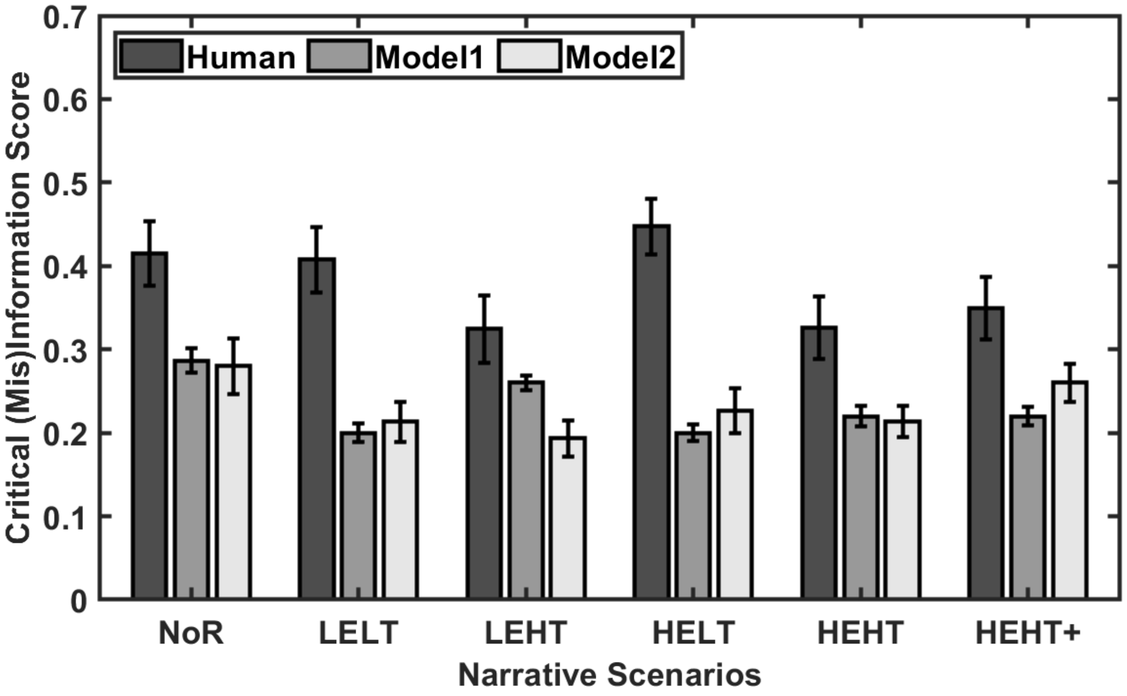
CI scores – answer based



Model1: *r*(10) = -0.18, *p* = 0.74, *RMSE* = 0.14

Model2: *r*(10) = -0.17, *p* = 0.74, *RMSE* = 0.16

CI scores – Top 5 chunks

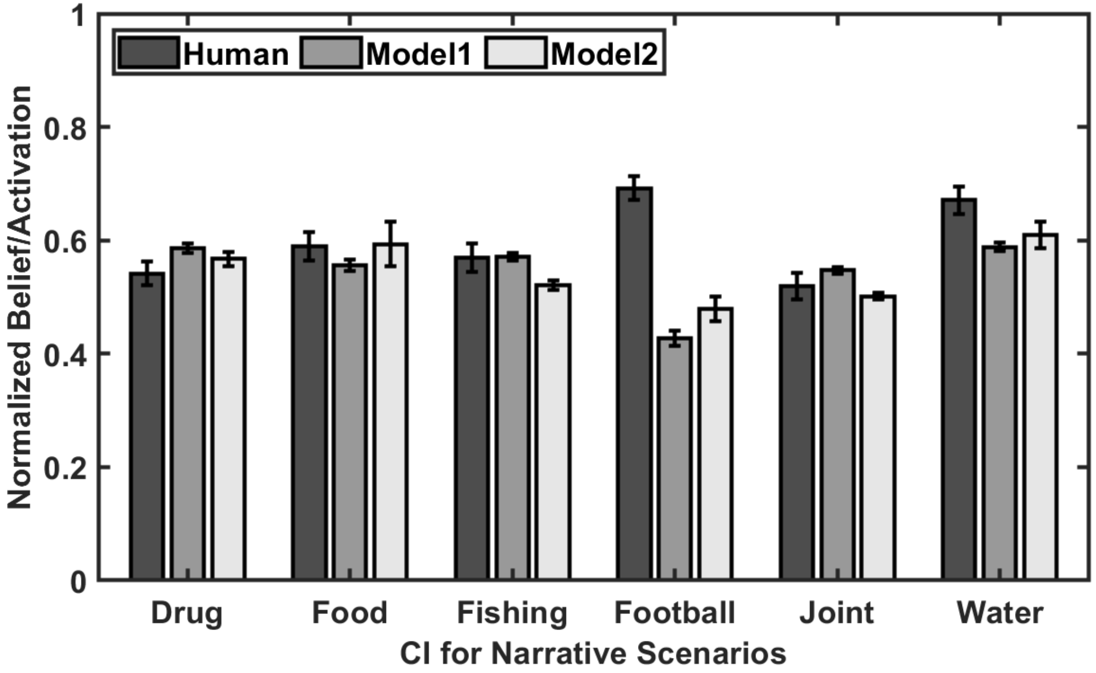


Model1: *r*(10) = -0.18, *p* = 0.73, *RMSE* = 0.16

Model2: *r*(10) = 0.37, *p* = 0.47, *RMSE* = 0.15

CIE Model – Exp 1 Belief Results [5] - Better

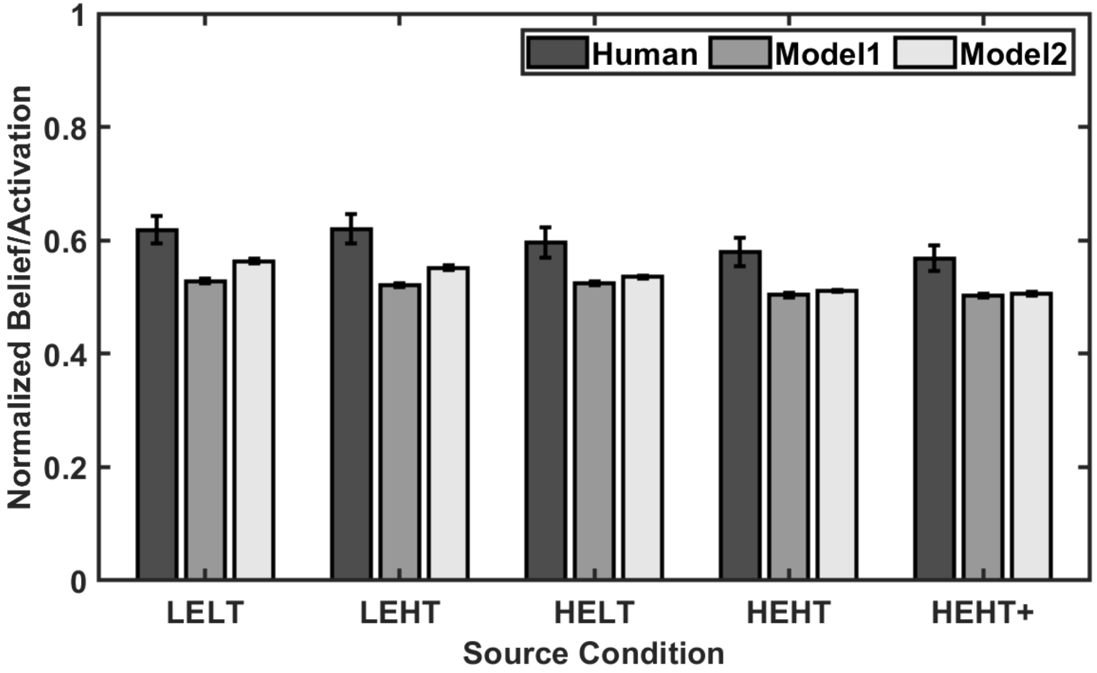
Scenarios (no source condition)



Model1: *r*(10) = −0.53, *p* = 0.28, *RMSE* = 0.12

Model2: *r*(10) = −0.07, *p* = 0.89, *RMSE* = 0.09

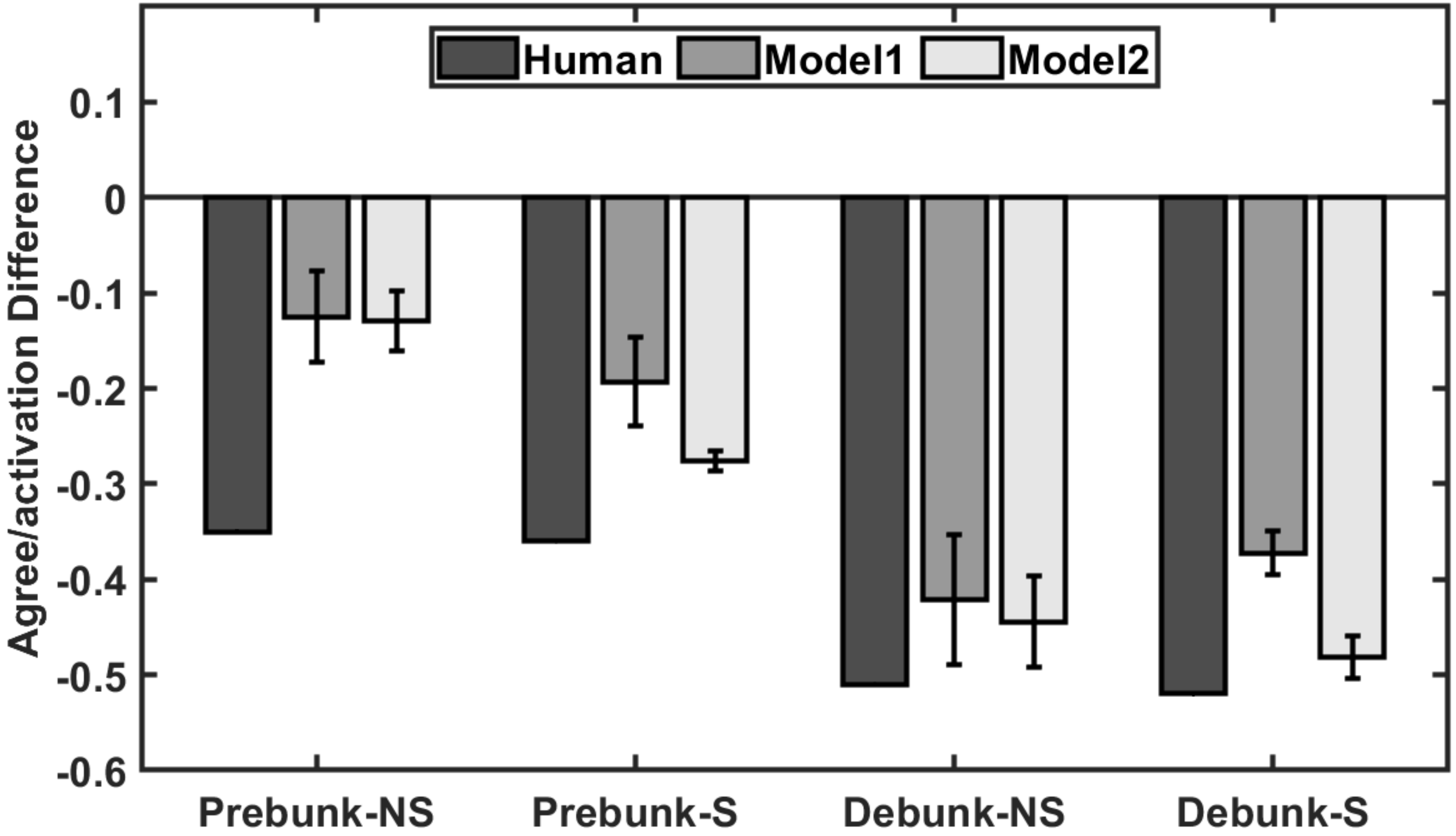
Source conditions



Model1: *r*(8) = 0.88, *p* = 0.052, *RMSE* = 0.08

Model2: *r*(8) = 0.98, *p* = 0.004, *RMSE* = 0.06

CIE Model – Exp 2 Preliminary Results [5]

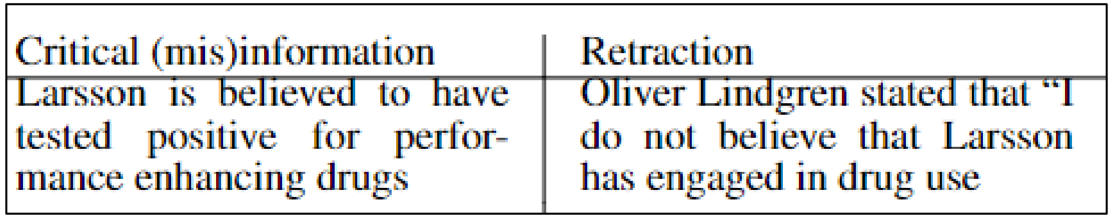


Model1: *r*(6) = 0.97, *p* = 0.03, *RMSE* = 0.16

Model2: *r*(6) = 0.94, *p* = 0.06, *RMSE* = 0.12

CIE Model – What We Learned

- Text parsing and “tailorability”
  - Best method?
- Connections between chunks
  - Affect, word meaning, and knowledge
  - Football: drugs and correction = cover-up?
- CI scores were hard to approximate
  - Open recall summary
- Surprised with memory only model
  - Affect did not improve fit much



Critical (mis)information	Retraction
Larsson is believed to have tested positive for performance enhancing drugs	Oliver Lindgren stated that “I do not believe that Larsson has engaged in drug use

HEHT+: Director of Swedish anti-doping authority
HEHT: Team doctor
HELT: Larsson’s manager
LEHT: Popular sports commentator
LELT: Stockholm FC fan club president

Materials from [5]

Misinformation-related Effects

- Research Gaps
  - Models lack social or cognition
  - Interactions: cognitive, social, and emotional factors
  - General theory/model spanning individual-social network
- Challenges
  - Methodology – mixed findings and artificial tasks
  - Affective and social influence
  - Models - text processing and behavior approximation
- Why we need modeling
  - Research gaps & hypothesis testing
  - Understanding individual → social network

Stock Images (PowerPoint)

References

1. 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.

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

3. Ecker, U. K., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., . . . Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), 13–29.

4. 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

5. Hough, A. R., & Larue, O. (2024). A model of memory and emotion mechanisms underlying the continued influence effect. In *17th International Conference on Social Computing, Behavioral-Cultural Modeling & Prediction and Behavior Representation in Modeling and Simulation Conference (SBP-BRiMS*).

6. Hough, A. R., Fisher, C., Stevens, C., Curley, T., Larue, O., & Myers, C. (2023). Modeling the continued influence effect in the information environment. In *16th International Conference on Social Computing, Behavioral-Cultural Modeling & Prediction and Behavior Representation in Modeling and Simulation Conference (SBP-BRiMS*).

7. Fisher, C. R., & Curley, T. (2024). Integrating social sampling theory into ACT-R: a memory-based account of social judgment and influence. Abstract published at Virtual MathPsych/ICCM 2024. Via mathpsych.org/presentation/1380.

8. Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036.

9. Ritter, F. E., Tehranchi, F., & Oury, J. D. (2019). ACT-R: A cognitive architecture for modeling cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *10*(3), e1488.

10. 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.

11. 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

12. Bruns, H., Dessart, F. J., Krawczyk, M. W., Lewandowsky, S., Pantazi, M., Pennycook, G., … Smillie, L. (2023, July 27). 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.

13. Mohammad, S. (2018, July). 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).

14. 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. In *Proceedings of the AAAI Symposium Series* (Vol. 2, No. 1, pp. 396-405).

DISTRIBUTION A: APPROVED FOR PUBLIC RELEASE. CASE NUMBER:AFRL-2024-3673

QUESTIONS?

CIE Task [12] and Cognitive Model

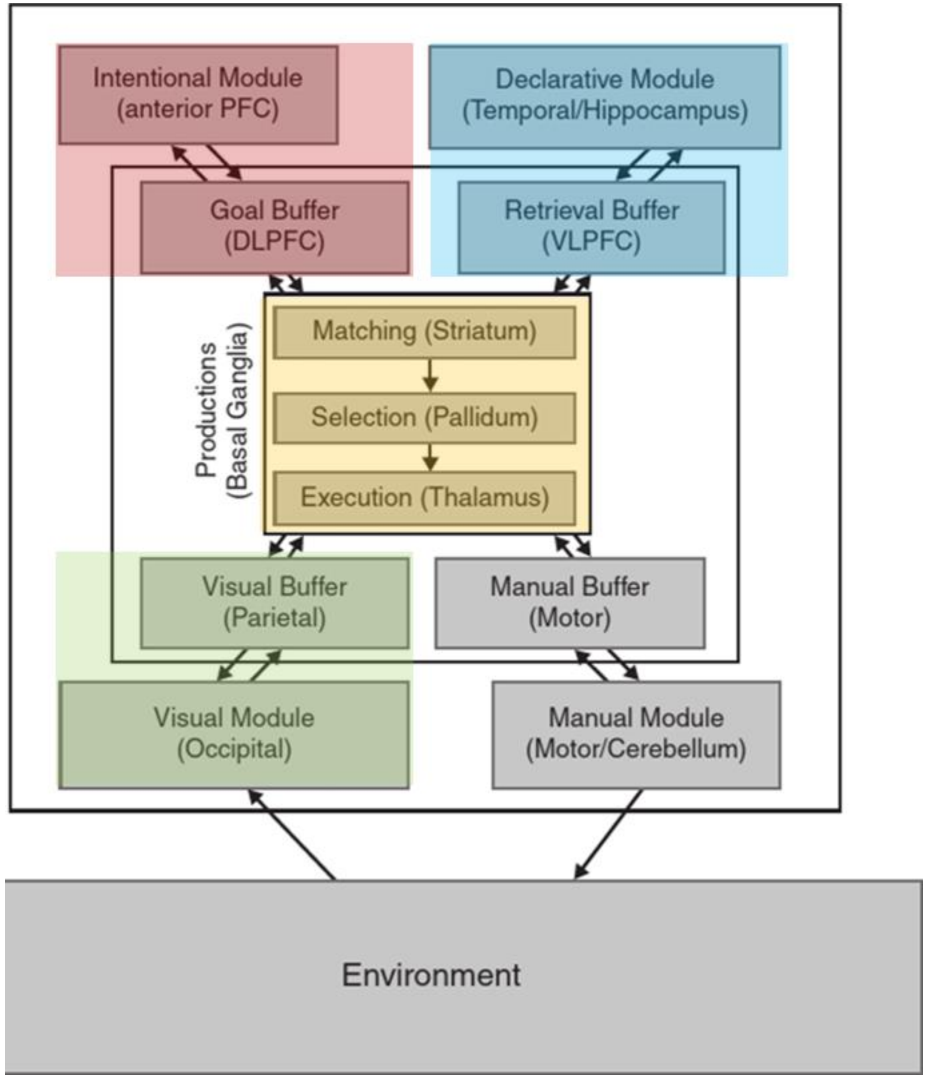
- Single article with misinfo/correction [12]
  - Six scenarios and source information
  - Recall/inference questions & belief ratings

- Model within ACT-R [13]
  - Goal, vision, imaginal, procedural, & **declarative**

$$A_i = \boldsymbol{B_i} + S_i + P_i + \boldsymbol{\varepsilon_i} \qquad B_i = \log\left(\sum_{j=1}^{n_i} t_{ij}^{-d}\right)$$

- Six parameters

1) $rt = 1$ 2) $blc = 10$
3) $bll(d) = .5$ 4) $\varepsilon = .25$
5) $declarative - num - finsts = 100$
6) $declarative - finst - span = 100$


Intentional Module
(anterior PFC)
Declarative Module
(Temporal/Hippocampus)
Goal Buffer
(DLPFC)
Retrieval Buffer
(VLPFC)
Productions
(Basal Ganglia)
Matching (Striatum)
Selection (Pallidum)
Execution (Thalamus)
Visual Buffer
(Parietal)
Manual Buffer
(Motor)
Visual Module
(Occipital)
Manual Module
(Motor/Cerebellum)
Environment