MODELING MISINFORMATION-RELATED EFFECTS: SUCCESSES AND CHALLENGES

ALEXANDER R HOUGH¹, OTHALIA LARUE²
AIR FORCE RESEARCH LABORATORY¹, PARALLAX ADVANCED RESEARCH²
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 \(\rightarrow\) small group \(\rightarrow\) social network
  • Assessing potential vulnerabilities and mitigations

• Add to current research – extend to realistic scenarios

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
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
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

Example article from [5]

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.

Byline: [13]
February 23, 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 beat, probably because they were paid by lobbyists.

The claim that there is a consensus among scientists is not just baseless, but also an enormous 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.

Example of prebunk, misinformation, and debunk articles from [6]
The list contains many food additives that have been suggested to pose serious health risks, including increased risk of cancer and ADHD.

### 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

### Table

<table>
<thead>
<tr>
<th></th>
<th>Valence</th>
<th>Arousal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health-risks</td>
<td>.240</td>
<td>.816</td>
</tr>
<tr>
<td>Serious</td>
<td>.5</td>
<td>.455</td>
</tr>
<tr>
<td>Cancer</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>ADHD</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Tables created by Alex Hough
CIE Model

Memory and Affect

• Short activation w/ core affect (valuation)
  • \( A_i = B_i + \varepsilon_i + (V_i \cdot vw) + (A_r_i \cdot aw) \)
  • \( V_i(j) = V_i(j - 1) + av[R_i(j) - V_i(j - 1)] \)
  • \( A_r_i(j) = abs(V_i(j)) \)

• Six declarative parameters
  1) \( r_t = 0 \)
  2) \( blc = 2.5 \)
  3) \( bll(d) = .5 \)
  4) \( \varepsilon = .25 \)
  5) \textit{declarative} – \textit{num} – \textit{finsts} = 100
  6) \textit{declarative} – \textit{finst} – \textit{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

Figures from [4-5]
CIE Model - Demo
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: $r(10) = 0.24, p = 0.64, RMSE = 0.23$
CIE Model – Exp 1 CI-score Results [new] - Not So Good

CI scores – answer based

CI scores – Top 5 chunks

Model1: $r(10) = -0.18$, $p = 0.73$, $RMSE = 0.16$
Model2: $r(10) = -0.17$, $p = 0.74$, $RMSE = 0.16$

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)

Source conditions

Model 1: \( r(8) = 0.88, p = 0.052, RMSE = 0.08 \)
Model 2: \( r(8) = 0.98, p = 0.004, RMSE = 0.06 \)

Model 1: \( r(10) = -0.53, p = 0.28, RMSE = 0.12 \)
Model 2: \( r(10) = -0.07, p = 0.89, RMSE = 0.09 \)
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

<table>
<thead>
<tr>
<th>Critical (mis)information</th>
<th>Retraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Larsson is believed to have tested positive for performance enhancing drugs</td>
<td>Oliver Lindgren stated that “I do not believe that Larsson has engaged in drug use”</td>
</tr>
</tbody>
</table>

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
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.


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

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

- Six parameters
  1) \( rt = 1 \)
  2) \( blc = 10 \)
  3) \( bl(d) = .5 \)
  4) \( \varepsilon = .25 \)
  5) \( \text{declarative} − \text{num} − \text{finsts} = 100 \)
  6) \( \text{declarative} − \text{finst} − \text{span} = 100 \)