Comparing Similarity and Homophily-Based Cognitive Models of Influence and Conformity

Robert Thomson

2024 ACT-R Workshop

*The opinions expressed are my personal opinions and do not reflect the official position of the United States Military Academy, United States Army, or United States Government.

Overview

- Challenging Assumptions
    - System 1 vs System 2
    - Cognitive Biases

- Topics for Discussion
    - Modeling Cognitive Biases
    - Cognitive Models of Trust Diffusion via Homophily and Similarity
    - Critical Thinking and the Role of Rehearsal Strategy

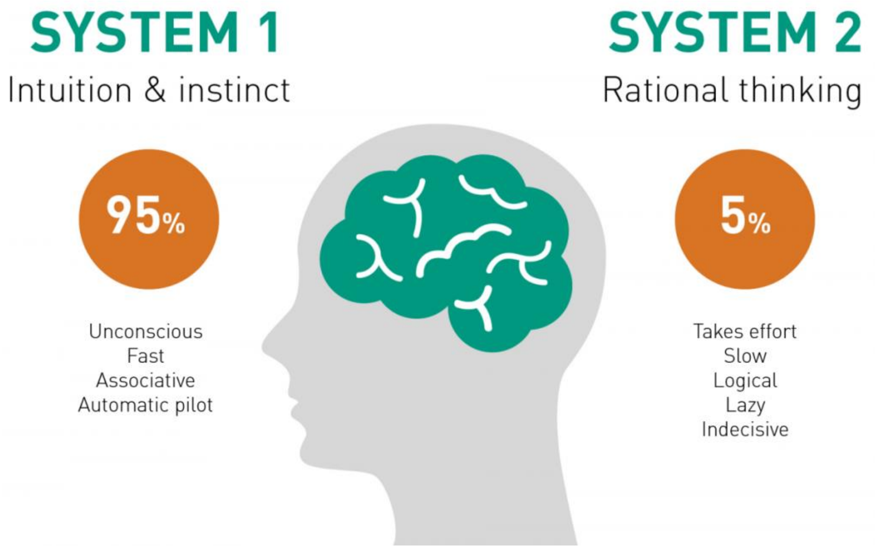
Challenging Assumptions: **False Dichotomies**

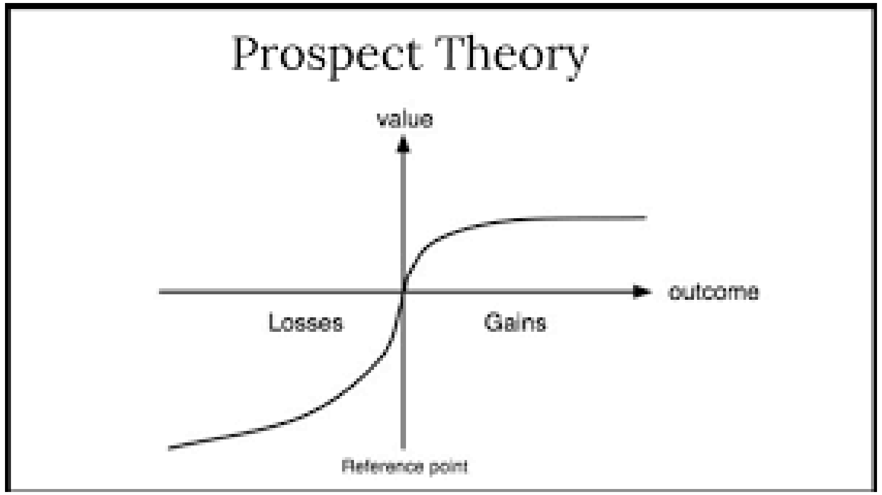
- There is **no physiological evidence** supporting this dichotomy

- Doesn’t **explain** how decisions are made

- Largely **unfalsifiable**

- We use multiple systems in parallel based on numerous factors (e.g., payoff, risk, effort)



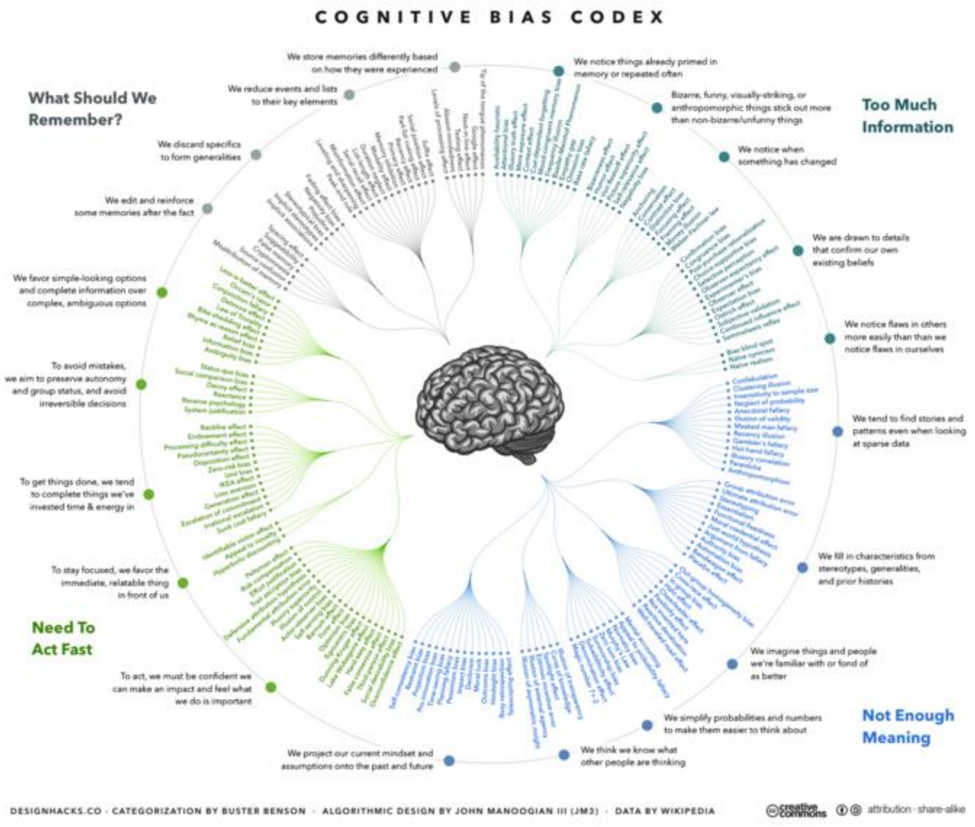


Challenging Assumption 2: Cognitive Biases

- Doesn’t **explain** how decisions are made

- Many seem to be specific cases of general principles

- Many biases can be explained by 3 factors: **recency**, **frequency**, and **order** of information.
  - Add in expectancy/effort and I think we have a 90% solution to model them


COGNITIVE BIAS CODEX
What Should We Remember?
Too Much Information
Need To Act Fast
Not Enough Meaning
We store memories differently based on how they were experienced
We notice things already primed in memory or repeated often
We reduce events and lists to their key elements
Bizarre, funny, visually-striking, or anthropomorphic things stick out more than non-bizarre/unfunny things
We discard specifics to form generalities
We notice when something has changed
We edit and reinforce some memories after the fact
We are drawn to details that confirm our own existing beliefs
We favor simple-looking options and complete information over complex, ambiguous options
We notice flaws in others more easily than than we notice flaws in ourselves
To avoid mistakes, we aim to preserve autonomy and group status, and avoid irreversible decisions
We tend to find stories and patterns even when looking at sparse data
To get things done, we tend to complete things we've invested time & energy in
We fill in characteristics from stereotypes, generalities, and prior histories
To stay focused, we favor the immediate, relatable thing in front of us
We imagine things and people we're familiar with or fond of as better
To act, we must be confident we can make an impact and feel what we do is important
We simplify probabilities and numbers to make them easier to think about
We project our current mindset and assumptions onto the past and future
We think we know what other people are thinking

DESIGNHACKS.CO · CATEGORIZATION BY BUSTER BENSON · ALGORITHMIC DESIGN BY JOHN MANOOGIAN III (JM3) · DATA BY WIKIPEDIA

creative commons attribution - share-alike

What is the Origin of Biases (from Cognition)

- Task structure
  - Environment/Task/Interface affordances
  - Time course of information (recency, frequency, & order)
- Cognitive Architecture
  - Mechanisms (e.g., spreading activation, blending)
  - Constraints (e.g., working memory)
  - Information flow
- Knowledge and strategies
  - Adaptive to interactions between task and cognition
  - Metacognitive determination of estimated effort

From Biases to Persuasive Communication

- **Challenge**: Understand how people consume information from sources they trust vs sources they don’t trust (reputational trust/reliability)
  - How does the reliability of the source further impact how an agent consumes the message, under various conditions?
  - Can this provide a cognitive explanation for social contagion, conformity, bias?

- **Assumption**: You increase trust with sources who communicate similar beliefs, and decrease for those who hold contrasting beliefs
  - Homophily-based belief updating (trust-weighted belief)
  - Similarity learning mechanism

Simulation Environment (Codename: **Othello**)

- Model receives messages and compares messenger’s opinions to the model’s own prior belief
  - Messages have a source (messenger), destination (recipient), topic, and stance (pro vs anti)
  - Assume that the model already supports a position on the topic
- Data consists of 1800 messages from 9 sources:
  - Each source has a different proportion of ‘pro’ stances
  - Iago 10%, Roderigo 20%, Cassio 20%, Montano 30%, Lodovico 40%, Brabantio 40%, Gratiano 50%, Bianca 50%, and Emilia 90%
  - Each 'trial', the model processes a single message from a single source
- Will the model ‘conform’ to the majority view or be resilient?
  - Compare homophily (algorithmic) vs similarity (structural) models

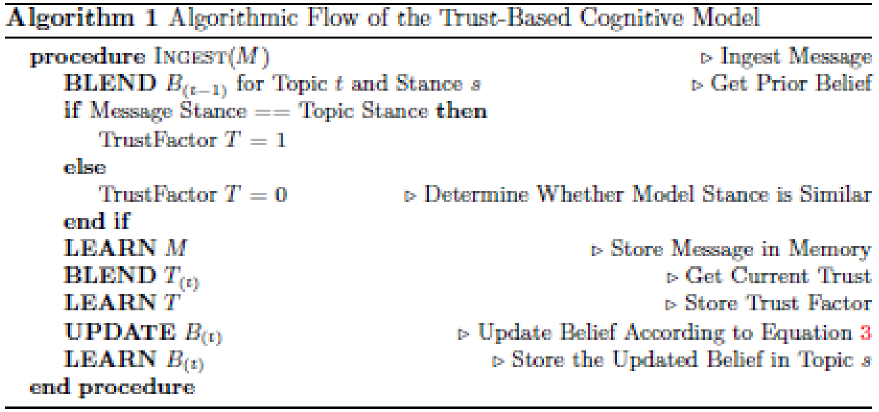
Biasing Belief Updating via Homophily

- **TRUST***: Stored as either 0 or 1 based on homophily
  - Blended retrieval for current strength (blends between the **0** and **1** instance)
    - Solely based on power law of learning and forgetting

- **BELIEF**: Stored as current belief (from 0 to 1)
  - Blended retrieval stores current belief, based on trust-weighted exponential:

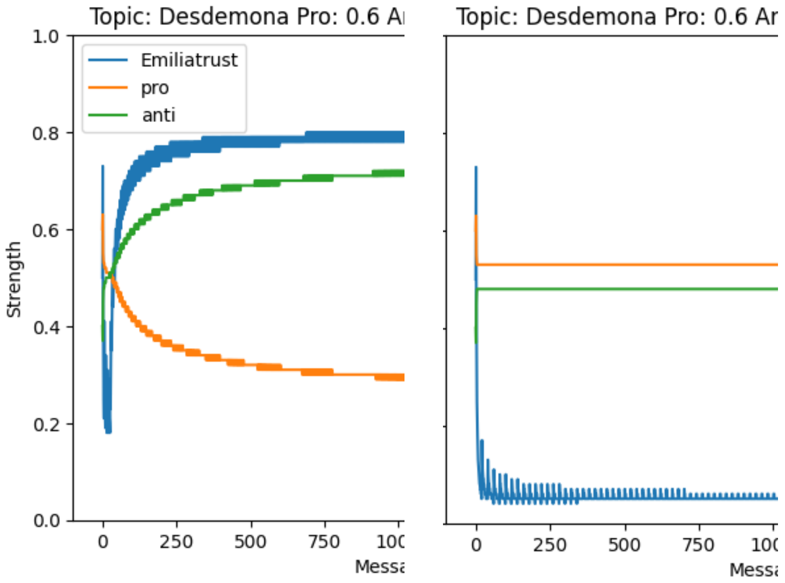
$$\text{New}Strength = \frac{priorStrength + updateFactor}{2}, \text{ where}$$

$$updateFactor = \ (messageStrength \ - prior) \cdot trust^2$$



Algorithm 1 Algorithmic Flow of the Trust-Based Cognitive Model

```
procedure INGEST($M$)                                          ▷ Ingest Message
    BLEND $B_{(t-1)}$ for Topic $t$ and Stance $s$                ▷ Get Prior Belief
    if Message Stance == Topic Stance then
        TrustFactor $T = 1$
    else
        TrustFactor $T = 0$                  ▷ Determine Whether Model Stance is Similar
    end if
    LEARN $M$                                           ▷ Store Message in Memory
    BLEND $T_{(t)}$                                          ▷ Get Current Trust
    LEARN $T$                                              ▷ Store Trust Factor
    UPDATE $B_{(t)}$                       ▷ Update Belief According to Equation 3
    LEARN $B_{(t)}$                           ▷ Store the Updated Belief in Topic $s$
end procedure
```



L: 20% Pro (flipped)

R: 5% Pro (resilient

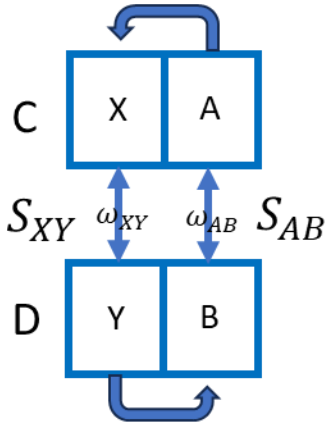
* Could be any latent factor(s) that bias how beliefs are updated.

Similarity Learning Mechanism

- Similarity prior at :MD (max-dif default to -1.0)
  - Bracket with :MS (max-sim default to 0.0) for self-similarities
  - Essential to get symmetrical push-backs
- Similarity weight of prior at 1
  - Multiply learning factor by product of frequencies
  - Can be generalized to product of chunk probabilities
- Alternating learning fields
  - First learn sources similarities from beliefs then belief similarities from sources
  - Alternative would be to learn both fields at the same time but potential stability issues

$$\omega'_{XY} = \omega_{XY} + p_C p_D$$

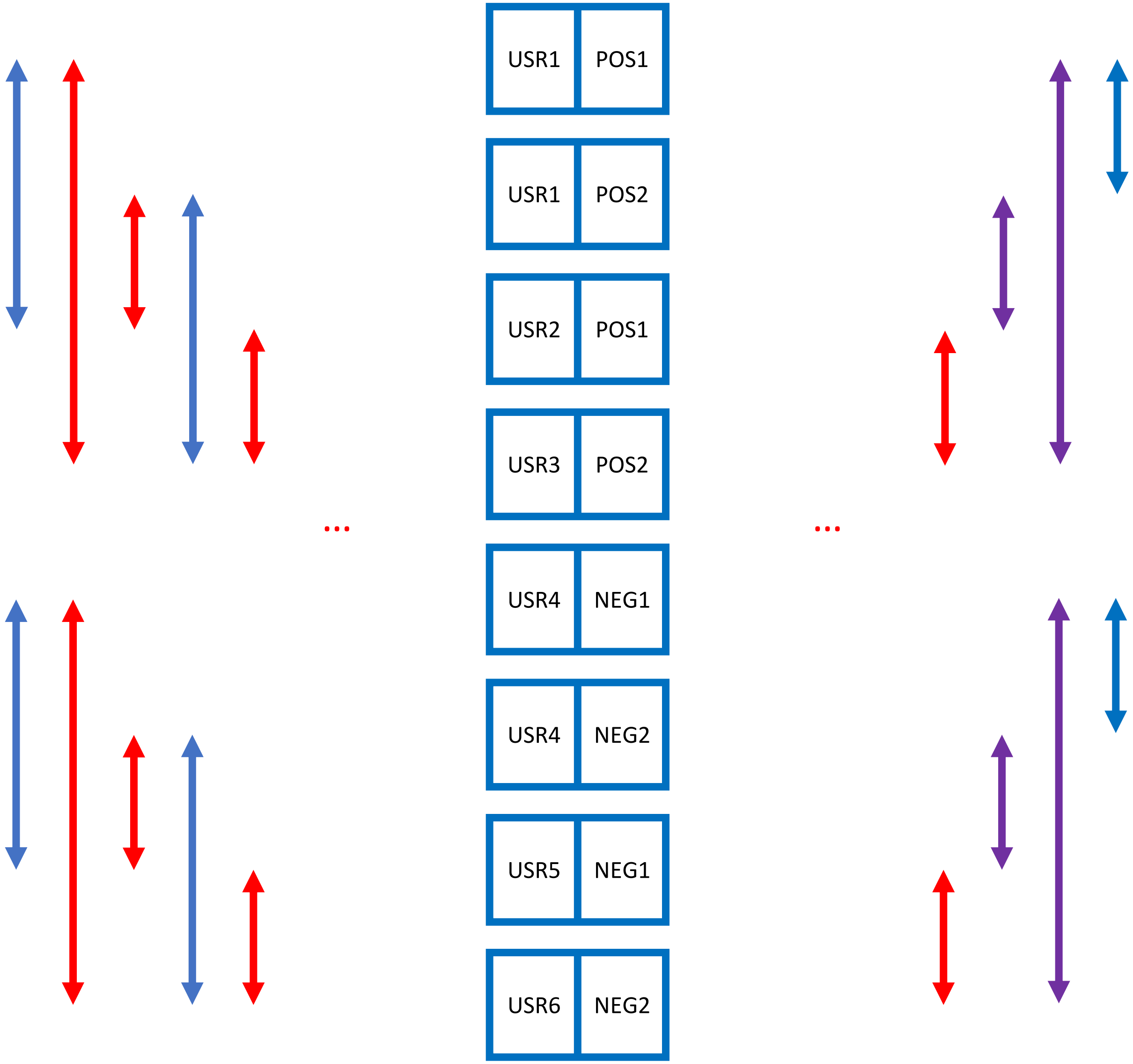
$$S'_{XY} = \frac{\omega_{XY} S_{XY} + p_C p_D S_{AB}}{\omega_{XY} + p_C p_D}$$



$$S'_{AB} = \frac{\omega_{AB} S_{AB} + p_C p_D S_{XY}}{\omega_{AB} + p_C p_D}$$

$$\omega'_{AB} = \omega_{AB} + p_C p_D$$

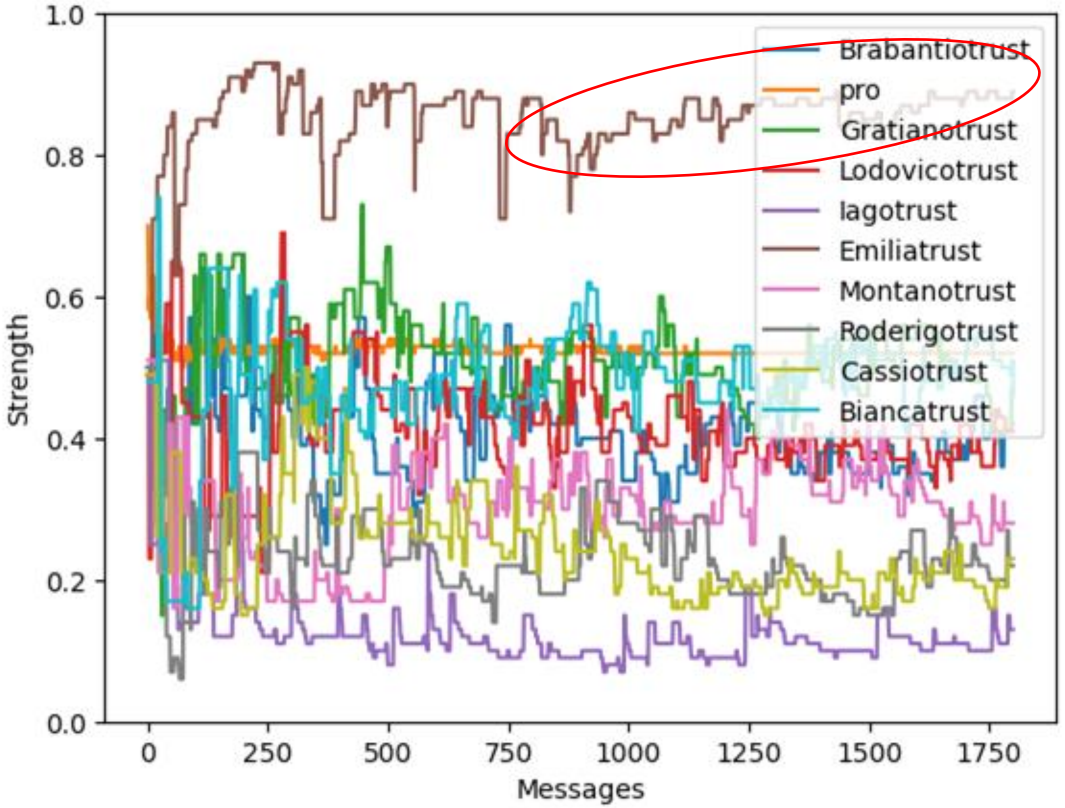
Similarities Between Users and Positions

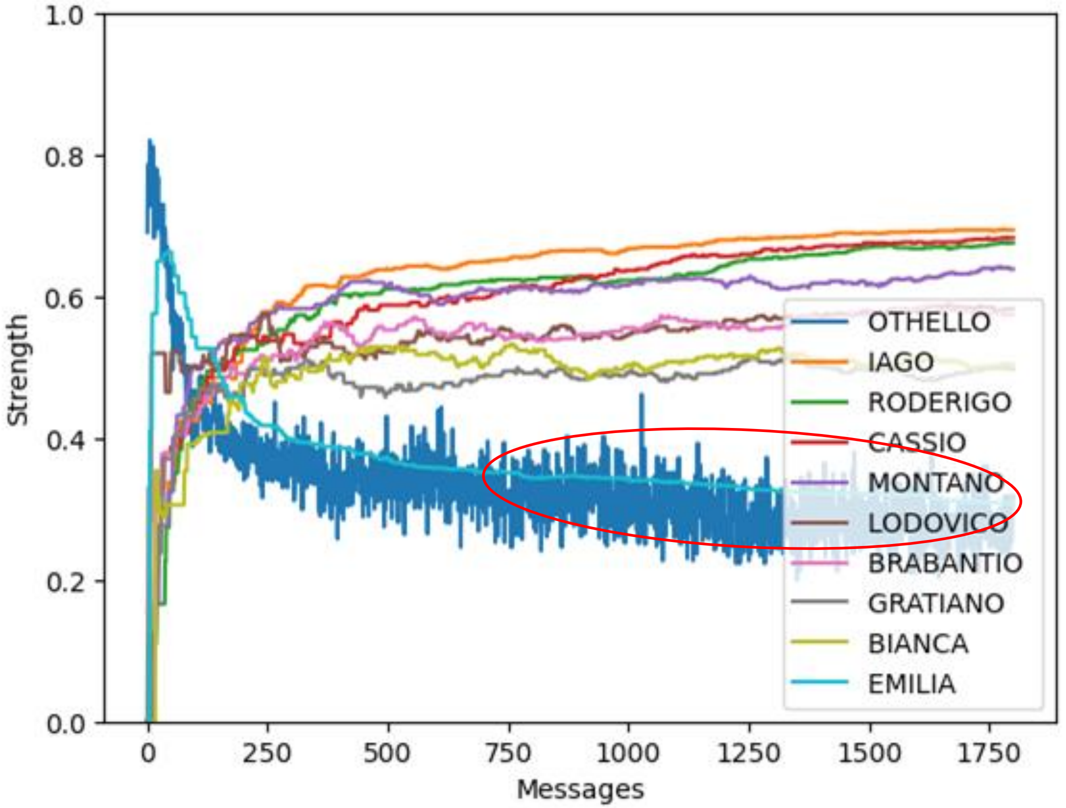


USR1	POS1
USR1	POS2
USR2	POS1
USR3	POS2
USR4	NEG1
USR4	NEG2
USR5	NEG1
USR6	NEG2

```
(defparameter *data*
 '((user1 1 pos1 pos2)
   (user2 1 pos1)
   (user3 1 pos2)
   (user4 1 neg1 neg2)
   (user5 1 neg1)
   (user6 1 neg2)))
```

Results (Measuring Trust in Emilia)





Homophily Model: Nonlinear trust dynamics mean a single confederate helps maintain trust
…crystallized but is relatively weak (.53 Pro/.47 Anti)

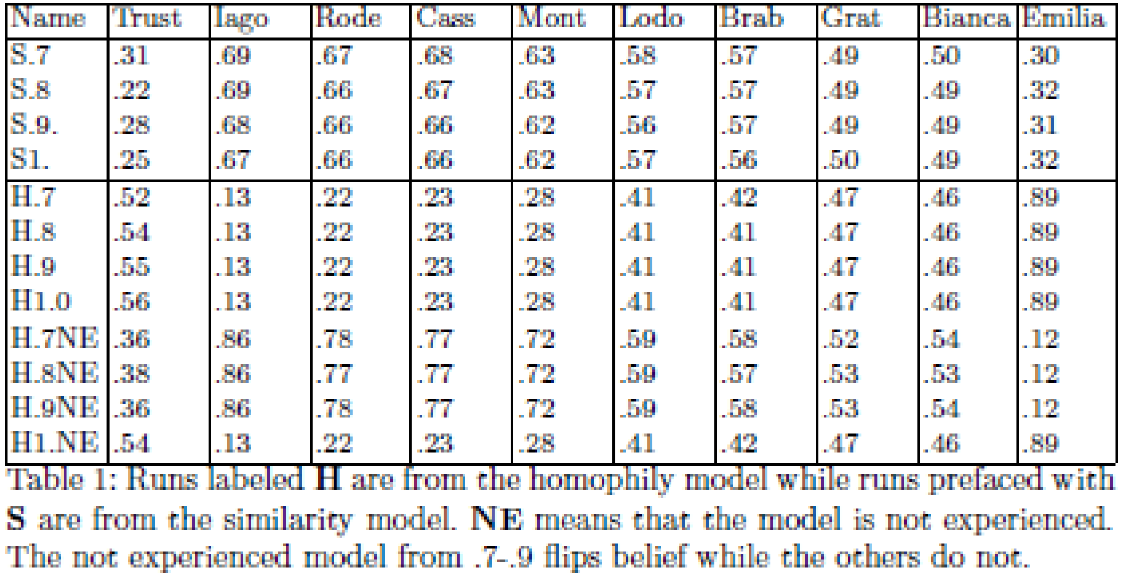
Similarity Model: Eventually asymptotes to average stance over time, with caveat

Long-Form Results

- Starting strength is less important than experience

- More experienced models are more robust against flipping stance

- With enough consistent counter-messaging, even an experienced model can be flipped



Name	Trust	Iago	Rode	Cass	Mont	Lodo	Brab	Grat	Bianca	Emilia
S.7	.31	.69	.67	.68	.63	.58	.57	.49	.50	.30
S.8	.22	.69	.66	.67	.63	.57	.57	.49	.49	.32
S.9.	.28	.68	.66	.66	.62	.56	.57	.49	.49	.31
S1.	.25	.67	.66	.66	.62	.57	.56	.50	.49	.32
H.7	.52	.13	.22	.23	.28	.41	.42	.47	.46	.89
H.8	.54	.13	.22	.23	.28	.41	.41	.47	.46	.89
H.9	.55	.13	.22	.23	.28	.41	.41	.47	.46	.89
H1.0	.56	.13	.22	.23	.28	.41	.41	.47	.46	.89
H.7NE	.36	.86	.78	.77	.72	.59	.58	.52	.54	.12
H.8NE	.38	.86	.77	.77	.72	.59	.57	.53	.53	.12
H.9NE	.36	.86	.78	.77	.72	.59	.58	.53	.54	.12
H1.NE	.54	.13	.22	.23	.28	.41	.42	.47	.46	.89

Table 1: Runs labeled **H** are from the homophily model while runs prefaced with **S** are from the similarity model. **NE** means that the model is not experienced. The not experienced model from .7-.9 flips belief while the others do not.

Weaponization of Critical Thinking

- Critical Thinking, especially counterfactual reasoning, has historically been taught as a debiasing technique
  - Paradoxically, those who critically reason are more likely to fall prey to certain biases due to their inherent skepticism
- By not committing to one hypothesis (to avoid contextual spreading activation) and considering/storing both (i.e., all) positions (to avoid base-level influence), in theory you can be more ‘rational’
  - You are also told to be aware of the source and trust ‘reliable’ sources
- I argue that without proper metacognitive training (learning how to learn and be critical), this can be easily manipulated
  - Homophily and similar opinions: **Frequency -> Familiarity -> Trust/Reliability**

Implication of the Metacognitive Perspective

- The human mind wasn’t designed with strategies to process the availability of online information without using heuristics
    - Availability confuses our metacognitive ability to process effort and risk
    - Affective content exacerbates the need to engage

- ‘Echo chambers’ eliminate the confederate in conformity research
    - Recency, frequency, and order effects (and environment) all point towards crystallization and polarization
    - Flipping a few people won’t be enough to move the needle

- Consistent messaging from **a single reasonably trusted source is more important** than fewer sources or mixed-messaging

What’s Next?

- Systematic investigation of major cognitive biases computationally to show similar effects from basic mechanisms
  - What environmental features and/or rehearsal strategies are implicated?

- Operationalizing ‘optimal’ stance flipping behavior and (in/out-)group dynamics

- Validation against human data
  - Does anyone have any?

- Find collaborators ☺