Confirmation Bias and Illusory Truth in Epistemic Networks

Nathan Gabriel

Logic and Philosophy of Science, UC Irvine

Introduction

The following networks models investigate the effects of confirmation bias in conjunction with illusory truth on agents's ability to successfully form a consensus about the truth. The type of confirmation bias investigated [1] involves agents being less likely to update their beliefs based on evidence that they know disagrees with their prior beliefs. Extending a model from Zollman [2], O'Connor and Gabriel [3] have already explored confirmation bias with results showing that a moderate amount of confirmation bias can benefit group learning. However, one can still question whether this continues to hold in conjunction with other dynamics. Here I investigate confirmation bias with the addition of an illusory truth dynamic. Illusory truth [4][5][6] is a phenomenon in which an agent's credence in a statement increases with repeated exposure to that statement, even in cases in which the statement is a known falsehood. It is also easy to conceive of natural ways in which one might be repeatedly exposed to the same claim or piece of evidence; perhaps a particular journal article is often discussed in one's social circle. Similarly, we are well aware that the internet can be a vehicle for repeated exposure to fake news, a problem that Pennycook et al. [5] are explicitly concerned with.

Base Model

The base model in this poster is equivalent a model given by Zollman [2] (but with Edos-Renyi random network structures). Scientific inquiry is modeled as a two armed bandit problem. Agents in a network collect and share evidence about which of two arms has higher payout rates. At each time step in the model, agents pull from the arm they believe to have higher payouts and share the results of the action. Agents then use Bayes' theorem to update their beliefs (modeled as beta-binomial distributions) with their own data from arm pulls along with any results that they accept from other agents. For all simulations, arm A pays out with probability $p_a = 0.499$, arm B pays out with probability $p_b = 0.5$, and agents perform trials of 1000 pulls at each timestep.

Illusory Truth Dynamics

Illusory truth dynamics extend the base model behavior of collecting and sharing data. Agents have memory of up to M many previously accepted data points. At the beginning of a simulation agents have no data in memory. On each time step, agents randomly choose up to S many of data points from memory to share with agents to whom they are connected. If a confirmation bias dynamic is present, agents then choose to accept or reject data shared them according to that dynamic. Agents do not update their beliefs until they have made all decisions about accepting or rejecting data that must be made on the given time step. Finally, agents randomly choose at most D many data points from among those that were accepted in the current round to replace a random data point in memory. Parameters, M, S, and D can be varied across simulations.

Resharing Evidence Harms Beliefs

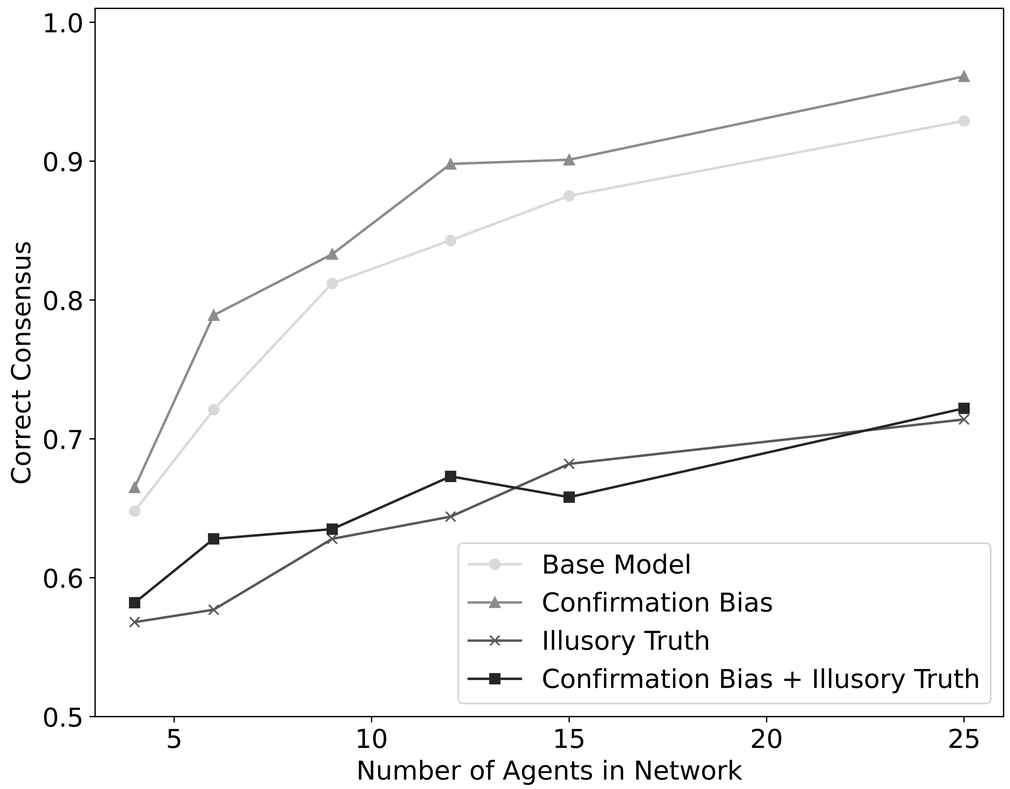

Figure 1:This figure shows the frequency of networks reaching correct consensus, with each data point representing 1000 simulations of the model, 30,000 time steps per simulation, and networks being generated with the probability of a connection between agents being $ER-prob = 0.5$. When confirmation bias is present, $t = 0.25$. When Illusory truth is present, $M = 50$, $S = 25$, and $D = 10$.

Confirmation Bias Dynamics

On each time step, after trial results are shared according to network connections, if shared results are from the bandit arm that the receiving agent believes to be worse, then she accepts the results and accordingly updates her beliefs with probability $p_{accept} = \gamma^t$ where:

$$\gamma = \sum_{i=0}^{999} \left(pmf_X(i, \alpha_X, \beta_X) * \sum_{j=i+1}^{1000} pmf_Y(j, \alpha_Y, \beta_Y) \right)$$

where $pmf_X(s)$, $pmf_Y(s)$ are the probability mass functions for the beta binomial distributions representing the arm believed to be better and the arm believed to be worse respectively. This can be computed analytically:

$$pmf(s, \alpha, \beta) = \binom{n}{s} \frac{B(s+\alpha, n-s+\beta)}{B(\alpha, \beta)}$$

where $B(\alpha, \beta) = \int_0^1 u^{(\alpha-1)}(1-u)^{(\beta-1)}du$.
t is a tuning parameter reflecting agents' degree of intolerance. Setting t to a low value models agents being more tolerant of results they consider unlikely and setting t to a high value models agents as more intolerant.

Takeaways

- Repeated sharing of old evidence, illusory truth, decreases the probability of a network reaching correct consensus.
- Confirmation bias can increase the probability of a network reaching correct consensus.
- Shown in Figure 2, illusory truth dynamics can allow polarization of agents beliefs to persist for long periods of time.
  - Confirmation bias can amplify the persistence of a polarized state in a network.
- Slowing the speed at which a network reaches consensus is not always beneficial.
  - O'Connor and Gabriel [3] note that confirmation bias slows the speed of reaching consensus effectively lowering the probability of premature consensus on a false belief. Zollman [2] finds a similar result favoring less connected networks over more connected networks.
  - However, illusory truth slows the speed of reaching consensus by allowing, potentially anomalous evidence to be shared and re-shared repeatedly nullifying any potential benefit from a slowed rate of consensus.

Effects on Consensus

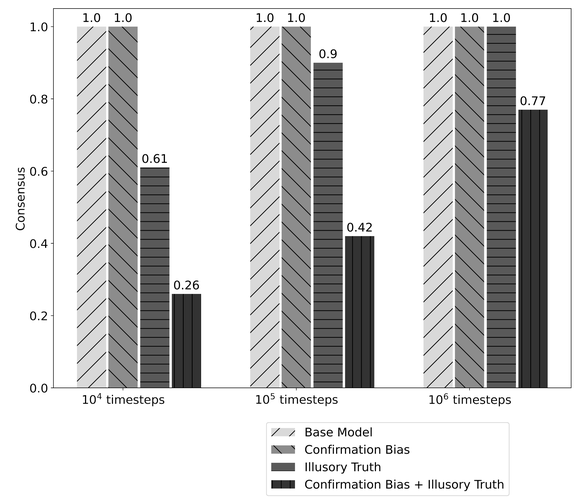

Figure 2:This figure shows the frequency of networks reaching consensus (correct or incorret) with each bar representing at least 100 simulations of the model for networks of 50 agents and the probability of a connection between agents being $ER-prob = 0.1$. When confirmation bias is present, $t = 1$. When illusory truth is present, $M = 50$, $S = 25$, and $D = 5$.

Note: In small networks, it is possible that illusory truth dynamics can speed up convergence to consensus. E.g. for 9 agents in a wheel network structure (a la Zollman [2]) 8,097 of 10,000 illusory truth simulations compared to 7,800 of 10,000 base model simulations were at consensus after just 100 timesteps (illusory truth parameters were $M = 10$, $S = 5$, and $D = 2$).

References

[1] William Hart, Dolores Albarracín, Alice H. Eagly, Inge Brechan, Matthew J. Lindberg, and Lisa Merrill.
Feeling validated versus being correct: A meta-analysis of selective exposure to information.
Psychological Bulletin, 135(4):555–588, 2009.

[2] Kevin Zollman.
The epistemic benefit of transient diversity.
Erkenntnis, 72(1):17–35, 2010.

[3] Cailin O'Connor and Nathan Gabriel.
Can confirmation bias improve group learning?
forthcomming.

[4] Lisa K. Fazio.
Repetition increases perceived truth even for known falsehoods.
Collabra: Psychology, 6(1), 07 2020.
38.

[5] Gordon Pennycook, Tyrone D Cannon, and David G. Rand.
Prior exposure increases perceived accuracy of fake news.
Journal of Experimental Psychology: General, 147(12):1865–1880, 2017.

[6] Alice Dechêne, Christoph Stahl, Jochim Hansen, and Michaela Wänke.
The truth about the truth: A meta-analytic review of the truth effect.
Personality and Social Psychology Review, 14(2):238–257, 2009.