



AI misinformation threatens the Wisdom of the crowd

By Emil Svenberg



Misinformation from AI

- Generative AI models are an increasingly common source of misinformation and disinformation
- Potentially one or multiple orders of magnitude more able to generate false content
- Hallucinates frequently
- Can remove an important assumption in mathematical “wisdom of the crowd” results (such as Condorcet’s Jury Theorem)

Condorcet's Jury Theorem (CJT)

Assumptions:

- Each voter votes independently of others (independence)
- Each voter has the same probability p of voting correctly, where $0.5 < p < 1$ (competence)
- The number of voters, n , is odd
- Only 2 choices, true or false
- By the law of large numbers, we expect the following results:
 - Crowd Infallibility: The vote approaches certainty as the number of individuals approach infinity
 - Increasing Reliability: As you add 2 voters, the reliability will increase

Very nice results, but can we really assume voter independence and competence?

So two of the main assumptions made in the theorem is that voters vote independently and have a greater than fair coin probability in voting correctly

But what if there's a common cause that affected the vote?

Well the conditional jury theorem steps in, and saves the results. However, if the common cause is a systematic source of misinformation, then voter competence is threatened

In fact, it reverses the result so the voters get it completely wrong in the limit and adding 2 voters decreases their reliability!

The Multi-Agent Simulation

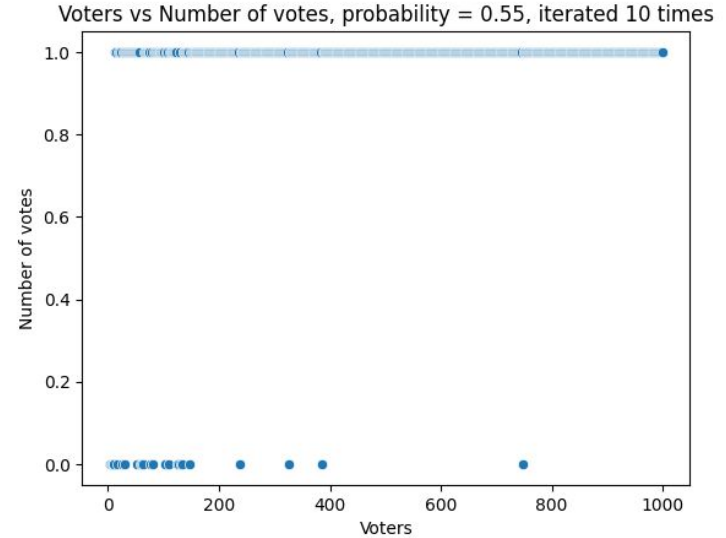
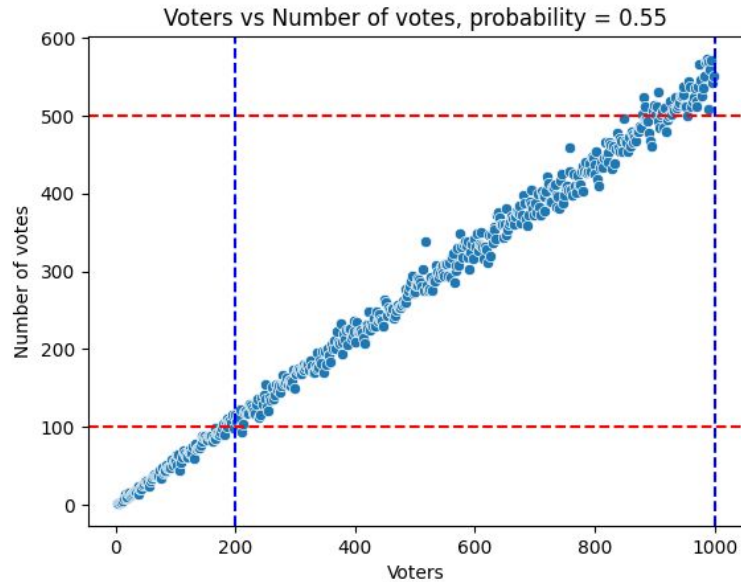
We create a basic, test simulation for the CJT in a program called “Mesa”. These agents will vote based on a fixed probability on initialization

Whenever they vote, they have some probability of getting it correct, affected by a common cause variable cc

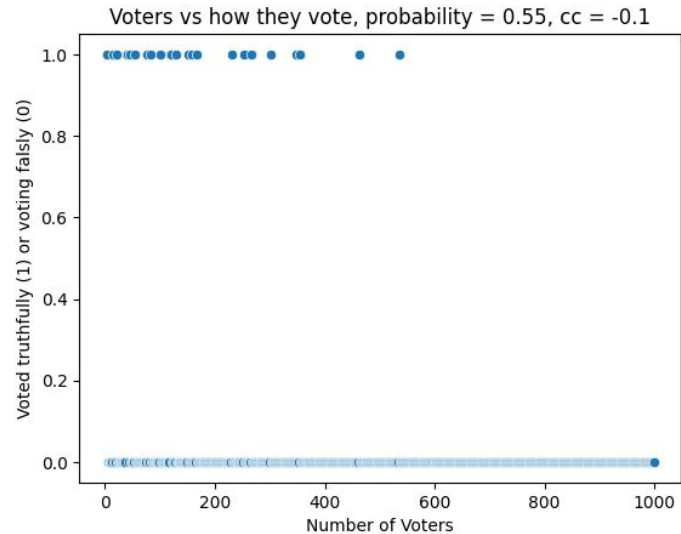
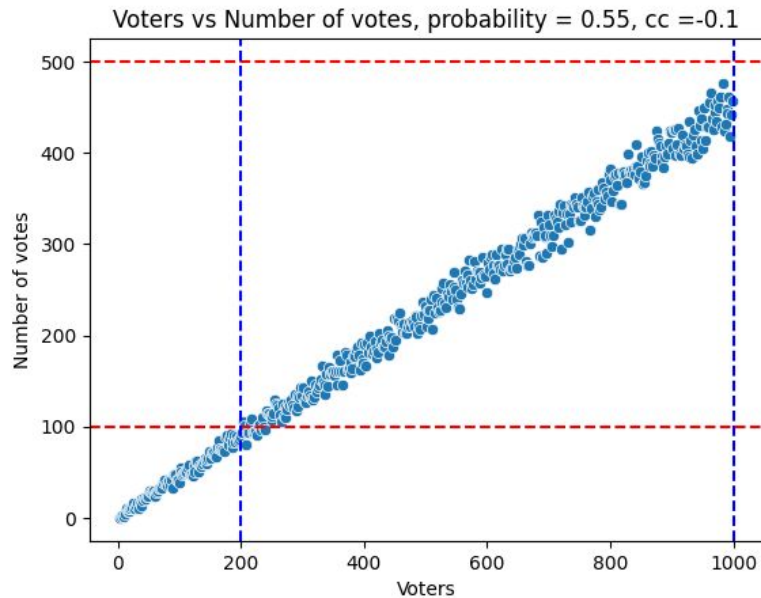
If $cc < 0$, then the common cause is a misinformation source and worsens the epistemic performance of each agent. If $p + cc < 0.5$, then we no longer have the voter competence assumption

We also take it that $cc = 0$ is the same as voter independence, which at $p > 0.5$ should give us the same result as in CJT

Results, $cc = 0$, $p = 0.55$



Results, $cc = -0.1$



Discussion

The results of $cc=0$ is how we expected them to be. But in the second scenario with a common cause of misinformation where $c < 0$ we saw how it severely impacted the epistemics of democracy

Future Research

- Work on expanding the complexity of the model, introducing a graph-network model where nodes represent individuals and edges represent information spread. We may be able to determine exactly how AI models can affect such spread and affect democratic epistemics
- Work empirically to test if AI generated misinformation/disinformation has caused a large-scale decrease in voter competence over time
- Work on implementing ways of preventing disinformation spread, such as prediction markets, or even using AI itself to check information