AI Misinformation Threatens Wisdom of the Crowd

The Philosophy Addict

ABSTRACT

This article explores generative AI models as a source of misinformation and can cause the epistemic performance of democracy to diminish by removing assumptions in well-established mathematical theorems regarding the wisdom of the crowd. More specifically, the theorem known as the Condorcet's Jury Theorem (CJT). We first present a theoretical examination of how generative AI can function as a systematic source of misinformation, potentially eroding the epistemic foundations of democracy by influencing voter independence and thus, compromising the collective decision-making process as outlined by CJT. Next, we demonstrate a Python-based multi-agent simulation framework, Mesa, to model voter behavior under two distinct scenarios: independent voting, and voting influenced by a source of misinformation. In the independent scenario, our results align with the CJT, showing that as the number of voters increases, the probability that the majority decision is correct approaches certainty, reaffirming crowd infallibility in ideal conditions. However, when introducing a common cause factor (simulating systematic misinformation) the independence assumption is removed, and the reliability of the majority decision diminishes. As misinformation increases, the probability of a correct majority decision diminishes. These findings underscore the vulnerability of democratic systems to AI-generated misinformation and highlight the necessity for robust mechanisms to mitigate its influence.

Key words: Democracy - Jury Theorem - Generative AI - AI misinformation

1 INTRODUCTION

The truthfulness of information circulating within public domains have become paramount to the functioning of democratic societies. The rise of generative artificial intelligence (AI) technologies has introduced sophisticated capabilities in creating and disseminating information, which, while beneficial in numerous contexts, poses significant challenges to the reliability of information ecosystems. These AI systems, capable of producing realistic texts, images, and videos, have become potent tools for generating both misinformation and disinformation at scale. [Monteith, 2023] Misinformation, defined as the sharing of false information, and disinformation, the deliberate creation and dissemination of falsehoods, are catalyzing uncertainties and manipulations that threaten the fabric of informed democracy. A concrete example of such an event happening was in the recent elections in Bangladesh where certain groups used deepfakes to spread disinformation about their opponents. [Parkin, 2023]

The implications of AI-driven misinformation can be examined with the help of Condorcet's Jury Theorem (CJT). CJT offers an optimistic view of collective decision-making, suggesting that a majority vote among independent, competent individuals gets more likely than not to arrive at the truth as the number of voters increases. [Franz Dietrich, 2021] This theorem presupposes that each voter has a better than random chance of discerning the truth for binary choice, thereby concluding that the probability that the majority decision reflects the truth approaches certainty as the group size approaches infinite size.

However, the foundational assumptions can be unsettled in the presence of generative AI technologies that spread misinformation. If these systems serve as a common cause of misinformation, affecting the judgments of a significant portion of the electorate, the premises of the CJT are undermined. The introduction of systematic bias or error into the decision-making process by AI can as we'll see, "reverse" the theorem to give a very pessimistic result instead.

2 THE CONDORCET'S JURY THEOREM

The CJT is an important theorem that illustrates the collective wisdom of making a correct epistemic guess under majority rule for binary choice. The theorem states that if each voter has an independent probability p of voting correctly, where $p > 0.5$ (better than a coin flip of a fair coin), and if the number of voters n is odd (to avoid ties) and tends to infinity, then the probability that the majority decision is correct approaches 1. This section will provide a mathematical proof of CJT using the Law of Large Numbers. [Franz Dietrich, 2021]

Assumptions:

- Each voter votes independently of others (independence)
- Each voter has the same probability p of voting correctly, where $0.5 < p \leq 1$ (competence)
- The number of voters, n, is odd
- Only 2 choices, true or false

2.1 Proof of CJT

Let's define a random variable X_i associated with the i-th voter, such that it takes on the value 1 whenever i votes congruent with ground truth, and 0 whenever i votes incongruently with ground truth (i.e whether or not i voted truthfully)

The X_i's are independent and identically distributed Bernoulli random variables with parameter p (the probability of voting correctly).

The sum of these n Bernoulli trials, S_n, represents the total number of correct votes:

$$S_n = \sum_{i=1}^{n} X_i$$

The expected value and variance of S_n are given by:

$$E(S_n) = np$$

$$\text{Var}(S_n) = np(1-p)$$

We then utilize the Law of Large Numbers, which states that the sample average $\overline{X}$ converges in probability to the expected value $E(X_i) = p$ as n approaches infinity:

$$\overline{X} = \frac{S_n}{n} \to p \quad \text{as } n \to \infty$$

For a majority to be correct, more than half of the voters must vote correctly, which mathematically can be stated as:

$$\frac{S_n}{n} > 0.5$$

Given that $p > 0.5$, the Law of Large Numbers ensures that $\overline{X}$ converges in probability to p, and thus:

$$P\left(\frac{S_n}{n} > 0.5\right) \to 1 \quad \text{as } n \to \infty$$

Q.E.D

This result shows that the probability of a correct majority decision tends towards 1 as the number of voters n becomes large, provided that each voter has an independent probability greater than 0.5 of voting correctly. As such we have mathematically validated the CJT, confirming that a larger, well-informed group is more likely to make correct decisions under majority rule and approaches infallibility in the limit. This property is known as "crowd infallibility".

A corollary that we won't prove here is called "increasing reliability", which states that for non-infallible voters, as the number of voters increase, the probability of being correct increases. So to remain an odd number of votes, as we add 2 voters with the same probability of being correct as everybody else, the probability of being correct still increases.

3 COMMON CAUSE PRINCIPLE

One of the assumptions in the last section was that each voter votes independently of others. This is an unrealistic assumption, and in this section we'll see what happens when voters have a common cause for their vote. Specifically, we'll see what happens when a misinformation common cause is presented, which may also threaten voter competence.

Just like how any two correlated variables can have a common cause, for example, the air-pressure causes both the barometer to change and the weather to change, so too can voters have a common cause for their choice of vote. We instead get the following: $P(A \cap B) > P(A)P(B)$ that is to say, it's more likely that they'll vote the same. [Christopher Hitchcock, 2020] Such a common cause could be the truth itself. However, it can also be because individuals are systematically misinformed by some common misinformation source. Such a common cause could be created by generative AI models that hallucinates (misinformation) or people setting up bot accounts for malicious spread of misinformation (disinformation).

If we consider the possibility that voters might not be entirely independent in their decision-making due to shared sources of information, cultural biases, or misinformation—the original assumptions of CJT are compromised. While that can be resolved by a reworked [Franz Dietrich, 2021] Conditional Jury theorem, if the common cause is not the truth itself then this still risks voter competence (i.e. that they have a greater than 50% probability of being correct). This common cause of misinformation can systematically skew the probability of each voter making a correct decision. If this misinformation is substantial, such that is pushed down below 50%, this inverts the theorem's result into decreasing reliability and the crowd always being wrong in the limit! [Franz Dietrich, 2021] This destroys the epistemic power of democracy, and wisdom of the crowd becomes stupidity of the crowd.

If generative AI systems is a common cause of misinformation that systematically misinforms voters, then this is a threat to the epistemic performance of democracy and the wisdom of the crowd. As a consequence, democracy becomes an increasingly epistemically worse system in comparison to various well-informed less democratic societies which are on a whole less desirable than a well-informed democratic system. While common causes of misinformation already exists, the automated process of AI systems can lead to a rapid increase in the propagation of misinformation as AI is much more efficient at spreading misinformation on a societal scale. This scale could be of perhaps one or multiple orders of magnitude. For instance, researchers have suggested that they are many times faster than humans in generating content with equal quality. [Nick Hajli, 2021] [Guillermo Sanchez Rosenberg, 2024]

4 THE MULTI-AGENT MODEL

The multi-agent simulation runs on python with a package named "Mesa". Mesa is an open-source Python library designed to enable the development and analysis of agent-based models (ABMs). ABMs are computational models that simulate the actions and interactions of autonomous agents (individuals or collective entities such as organizations) to assess their interactions with the system as a whole. Mesa provides a framework that facilitates the creation of complex agent-based simulations in a Pythonic, object-oriented environment. [Project-Mesa-Team, 2024]

The code will also involve typical data analysis libraries such as Pandas and Numpy, aswell as Seaborn which is used for visualization. Random is a default python package for pseudo-random number generation. The code will be available in a public, open-source Github page: https://github.com/Kattenelvis/AIxDemocracy

4.1 Multi-Agent Model With Independence Assumptions

We begin by importing all the above mentioned tools into our project:

```
import random
from mesa import Agent, Model, batch_run
from mesa.time import SimultaneousActivation
from mesa.datacollection import DataCollector
import pandas as pd
import seaborn as sns
import numpy as np
```

After which we define an VoterAgent which will contain the properties of the agent:

```
class VoterAgent(Agent):
    """ An agent with a belief given some
    probability p"""
```

```
    def __init__(self, unique_id, model, p, cc):
        super().__init__(unique_id, model)
        self.p = p
        self.cc = cc


    # Agents vote
    def step(self):
        self.model.voteIndependent(self.p, self.cc
    )
```

After which we define the multi-agent model. The model itself will hold two important simulation variables and two data gathering variables. The simulation variables being the probability of each agent being correct and the common cause factor. The data gathering variables are the votes in each model and "totalnumvotes" (being the variable for the total number of votes). The model makes all agents vote at once. Iterating over the model (i.e making the voters vote multiple times with the same probability as before) will reset the votes variable at each tick, but not "totalnumvotes" which accumulates.

```
class VotingModel(Model):
    """A model with some number of agents."""
    def __init__(self, N, p, cc):
        self.num_agents = N
        self.schedule = SimultaneousActivation(
    self)
        self.p = p
        self.cc = cc
        self.votes = []
        self.totalnumvotes = 0
        self.running = True


        # Create agents
        for i in range(self.num_agents + 1):
            a = VoterAgent(i, self, p, cc)
            self.schedule.add(a)


        # Set up data collector
        self.datacollector = DataCollector(
            model_reporters={"Majority_Vote":
    lambda m: sum(m.votes) > N/2,
                             "Total_Votes": "
    totalnumvotes"}
        )


    # Votes based on independence assumptions i.e
    P(A&B) = P(A)P(B)
    def voteIndependent(self, p):
        self.votes.append(1 if random.random() < p
     else 0)


    def step(self):
        self.schedule.step()  # Activate all
    agents, each agent votes once
        self.totalnumvotes += sum(self.votes)  #
    Add number of votes this step to total
        self.datacollector.collect(self)  #
    Collect data after votes
        self.votes = []  # Clear votes after
    collecting data for this step
```

Here's how to run it:

```
p = 0.51  # Probability to believe in ground truth
N = 53 # Number of individuals
cc = 0 # Common Cause factor
T = 23 # Number of times the population vote. Can
    be imagined that they recast their vote at
    each timestep, and this is how many timesteps
    long the simulation goes on for.


model = VotingModel(N, p, cc)
for i in range(T):
    model.step()


data = model.datacollector.
    get_model_vars_dataframe()
```

4.2 Multi-Agent Model with the Common Cause

To implement the common cause principle in our simulation, we adapted our voting function to include a common cause factor, denoted as cc. The modified voting function, voteCommonCause, adjusts the probability of an individual casting a 'yes'-vote, reflecting the influence of a common cause that affects all voters uniformly. The code snippet for this function is as follows:

```
class VotingModel(Model):
    # Votes based on the common cause principle i.
    e P(A&B) > P(A)P(B)
    # If cc = 0, then its equivalent with the
    independence assumption i.e P(A&B) P(A)P(B)
    def voteCommonCause(self, p, cc):
        self.votes.append(1 if random.random() -
    cc < p else 0)
```

The function voteCommonCause integrates the common cause effect directly into the voting decision of each agent. Here, p represents the baseline probability of a voter voting 'yes' under normal conditions (without any external influence). The common cause factor cc modifies this probability. Specifically, random.random() - cc shifts the uniform distribution generated by random.random() by the factor cc. If this adjusted value is less than p, the voter votes 'yes' (represented by 1), otherwise 'no' (represented by 0).

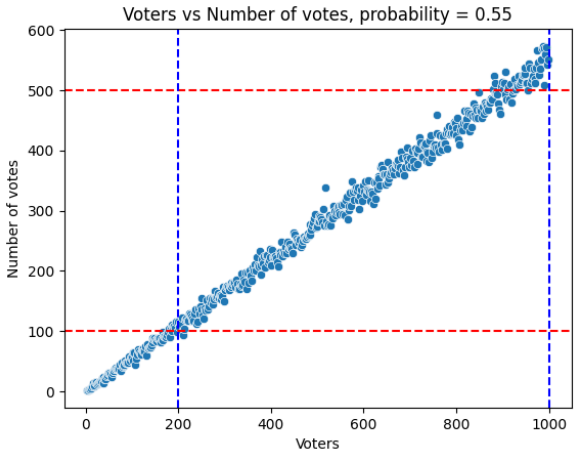
When cc = 0: The adjustment factor is zero, meaning that random.random() < p directly determines the vote. This scenario reverts to the independence assumption where each voter's decision is made independently of others, and the probability of a voter voting 'yes' is exactly p. When cc > 0: The common cause decreases the threshold below which a voter decides to vote 'yes'. This effectively increases the probability of a 'yes' vote, illustrating how a common cause can skew the voting behavior of the population towards a more uniform outcome.

5 RESULTS

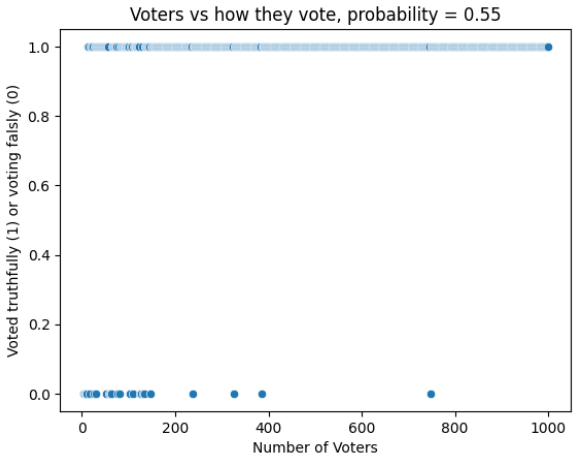
5.1 Multi-Agent Model With Independence Assumptions

We initiated our simulation under the framework of independence assumptions, specifically setting the common cause factor to zero. The outcomes of this simulation closely aligned with our initial hypotheses concerning voter behavior in an independent voting scenario. The results demonstrated the pattern we expected on the number of voters and how they would vote. Notably, with 1000 voters, the total number of affirmative votes consistently exceeded 500, thereby securing a majority 'yes' vote under majority rule. This outcome is

clearly depicted in the second image. For comparison, at an earlier stage with only 200 voters, a significant proportion of the results fell below the 100 votes necessary to achieve a majority.

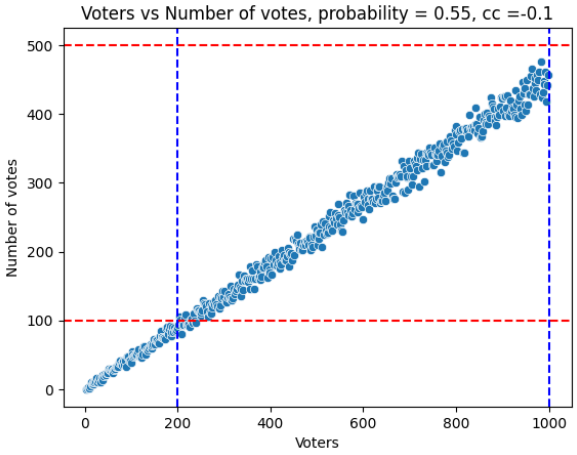


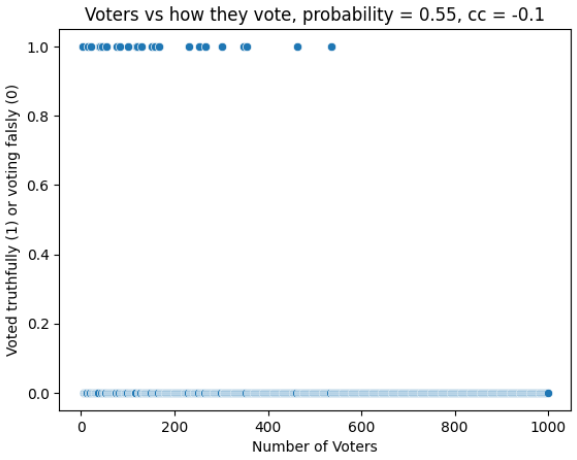
We can then see if they voted correctly. In the image below, we can see that as the number of voters increase, an ever increasing frequency of them voted based on ground truth.



5.2 Multi-Agent Model with the Common Cause

Secondly, we ran an simulation without independence assumptions i.e cc<0. Particularly, it has a negative common cause factor, which indicates that the common cause is creating a negative pressure on voting based on ground truth. Just like with the independence assumptions, the outcomes were again in accordance with our hypotheses regarding voter behavior when being subject to systematic misinformation from some common cause. Thus we end up with results where the wisdom of the crowd turns on its head, where the voters become increasingly unreliable over time, as can be seen in the following pictures.





6 CONCLUSION

The results from the simulation confirms that under the independence assumption, as the number of voters increases, the voters collective reliability improves. The data from our simulations align with the theoretical expectations from the CJT. However as the independence assumptions are reduced and a common cause of misinformation is introduced, we see that even when each voter has a over 50% probability of being correct, the common cause makes them sufficiently worse to threaten the wisdom of the crowd.

And as was discussed in section 3, if the reports regarding the rapid, automated production of misinformation from generative AI are accurate, then we may see a rapid increase in the cc factor in the real world. That is to say, it may go from a small factor of say -0.01 when only humans were producing misinformation, but now it may become -0.1. If humans are close to coin-flip regarding ground truth, this can seriously undermine the epistemics of democracy.

The simulation of what the mathematical theorems have already affirmed may seem like empirically testing Pythagorean theorem in the real world after it has been proved, and it is! However, such a simulation can be expanded to find more complex and interesting phenomena such that new results can potentially be found. Such research may include fitting the multi-agent models with more sophisticated versions of common causes, including making information spread on a network graph where individuals are on nodes and edges symbolize information paths between individuals, and having some special information nodes where information comes from (truth-sources and

AI generated sources) such that there may be multiple competing common causes (and perhaps also, multiple competing propositions rather than just one that is being judged by the voters).

We may also desire empirical testing on humans regarding these hypothesises, which includes doing a large scale test where people may be exposed to AI generated misinformation combined with information and made to choose what they think is true. This way one can empirically test the results in this study and replicate it for a human sample. For future research, prediction markets may be one way to diminish the risk, as there will be stronger incentives to be less misinformed.

There has also been research done on Predictive Liquid Democracy [Hagberg, 2023] utilizing a jury theorem in their proof. Perhaps prediction markets can incentivize truthful information, and that these prediction markets can inform voters on how to vote. It's worth looking into various incentive structures which may be created that could encourage believing in ground-truth and minimizing risk of misinformation. Predictive liquid democracy is an improved version of Robin Hanson's Futarchy [Loke Hagberg, ated].

Perhaps AI can be used to decrease risk of misinformation by analysing information and its source better than a human.

And lastly an important ethics declaration: This paper may not be used to justify anti-democratic viewpoint. Do not do that. This paper does not state that democratic systems can or should be abandoned in the face of increased misinformation.

BIBLIOGRAPHY

Christopher Hitchcock, M. R. (2020). Reichenbach's common cause principle. *Stanford Encyclopedia of Philosophy*.

Franz Dietrich, K. S. (2021). Jury theorems. *Stanford Encyclopedia of Philosophy*.

Guillermo Sanchez Rosenberg, Martin Magnéli, N. B. M. K. (2024). Chatgpt-4 generates orthopedic discharge documents faster than humans maintaining comparable quality: a pilot study of 6 cases. *ResearchGate*.

Hagberg, L. (2023). *Collected papers on finitist mathematics and phenomenalism: A digital phenomenology and predictive liquid democracy*.

Loke Hagberg, T. K. (undated). Predictive liquid democracy. *ResearchGate*.

Monteith (2023). Artificial intelligence and increasing misinformation. *The British Journal of Psychiatry*, 224:33–35.

Nick Hajli, Usman Saeed, M. T. F. S. (2021). Social bots and the spread of disinformation in social media: The challenges of artificial intelligence. *British Journal of Managment*.

Parkin, B. (2023). Bangladesh election emerges as testing ground for ai-created disinformation videos. south asia. *Financial Times*.

Project-Mesa-Team (2024). Mesa: Agent-based modeling in python.