# Determining the spread of scientific misinformation related to AI, climate change, and vaccines.

**Fiona Calcagnini, COL '23 ; Justin Goldstein, COL '24; Rich Pihlstrom, COL '24**

Georgetown University, Massive Data Institute (MDI), Center for Security and Emerging Technology (CSET)

## Introduction

The objective of this research project is to understand how peer-reviewed, published, scientific articles are used (or misused) as evidence for misinformation on social media. In our research, we analyzed three different fields of scientific information: artificial intelligence, climate change, and vaccines. This semester explored the first part of a two part research process: developing an ethnography of scientific misinformation, and performing a quantitative analysis.

We are interested in determining if certain subareas of each field are more susceptible to being used as evidence for misinformation. Therefore, for each field of scientific information we used the following methods:

- We computationally developed a list of topics that capture subcategories within each respective field.
- Within each topic, we then produced a dictionary of keywords that corresponds to each subcategory.

We then determined the popularity of the scientific articles using Altmetrics and collected tweets mentioning each article:

- We filtered this list to only have tweets which contained phrases about misinformation.
- Once we identify these tweets and their associated article, we determined the level of spread associated with these misinformation clusters.

**Branch 1**

## Article Topic Analysis

We used CSET's "Digital Science Dimensions" dataset, which contains the *doi, title, abstract, year,* and *country of origin* of published papers, to computationally develop topic-dictionaries of keywords for each scientific field. The dictionary distributions illustrate the popularity of different keywords across variables like year, country, and topic choice.

| | | | | | |
|---|---|---|---|---|---|
| **YIELD** | crop | food security | agriculture | sustainable | cultivation |
| **EMISSION** | atmosphere | pollution | greenhouse | waste | carbon |
| **PLANT** | genes | biomass | genotypes | soils | decomposition |
| **ENERGY** | generation | power | fuel | electricity | green |
| **BIODIVERSITY** | conservation | ecosystem | restoration | invasive | extinction |
| **LAKES** | salinity | microbial | nutrients | coral | watershed |
| **GLOBAL SOCIETY** | governance | cooperation | education | media | tourism |
| **HEALTH** | disease | public health | exposure | transmission | pathogens |
| **MARKET** | investment | companies | costs | economy | industry |
| **RISKS** | vulnerability | disasters | scarcity | hazards | extreme events |

**Figure 1.** An example table of topic-dictionary entries for climate change articles. We repeatedly ran a **Guided Topic-Noise Model (GTM)** algorithm on the *titles* and *abstracts* of every paper, which produced the seed dictionary for the next iteration. After a few runs, we labelled each grouping with an apt title.

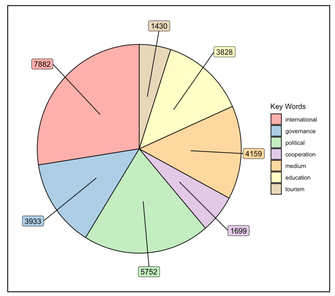


**Figure 2.** A pie chart of counts for each keyword in the "Global Society" topic of climate change. From the chart, we see that universal terms like "international" and "political" appear most often, while words like "tourism," which may only be relevant for select countries with robust tourist industries, are less popular.

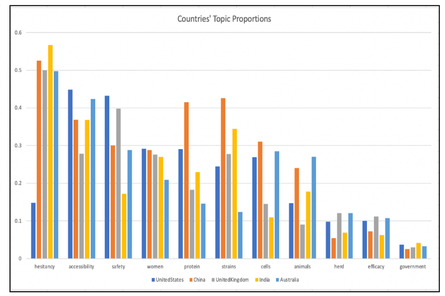


**Figure 3.** A nested bar chart illustrating the proportion of vaccine papers from five countries with at least one keyword from a topic for all vaccine topics. This graph reveals how different countries focus on different aspects of vaccines. Countries like China and India focus on "hesitancy" while the US emphasizes "safety."

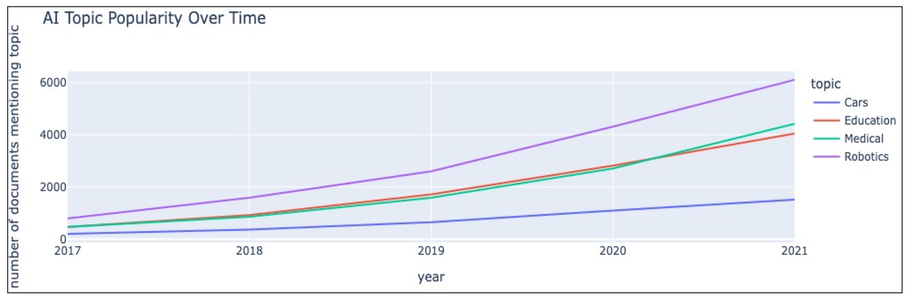


**Figure 4.** A line graph illustrating the change in the number of AI papers that contain at least one keyword from a topic for four AI topics (cars, education, medical, robots) from 2017 to 2021. While the trend of steady increase is consistent for every topic, features like the steeper increase of medical keywords during COVID years offer insight about the change in popularity of various topics.

**Branch 2**

## Tweet Collection

We used CSET's "Altmetrics" dataset, which contains *ids* for tweets associated to published papers, alongside Python's Tweepy Twitter API to collect information about original tweets related to articles in each scientific field. Tweet information included *user id, time of creation, follower count, text,* and a list of r*etweet ids*.

## Finding Misinformation Online

To identify which papers were used to propagate misinformation online, we filtered the tweet information to find tweets that contained phrases indicating misinformation (e.g. "climate scam"). For papers at the center of misinformed discussion, our visualization shows the network of tweets referring to the paper and their associated retweets.

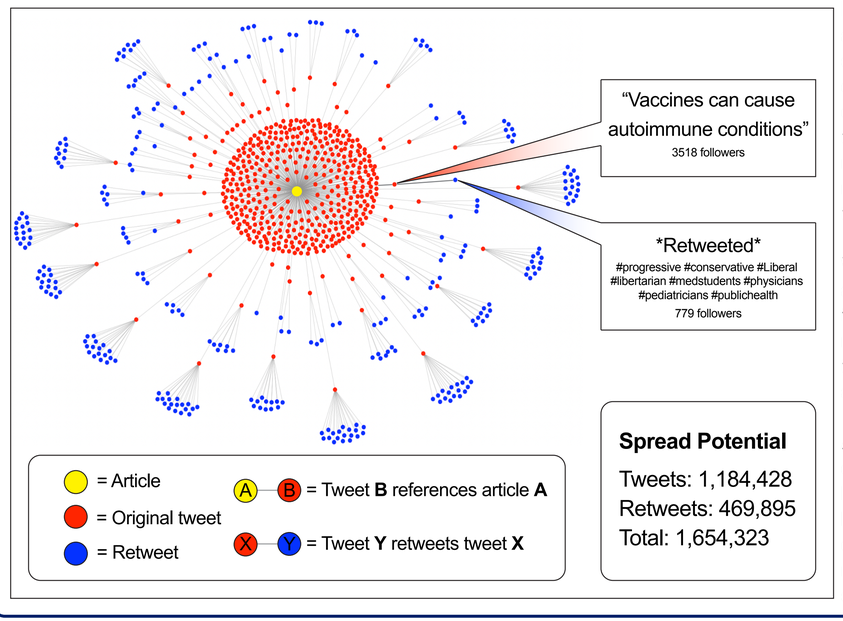


**Figure 5.** A network graph centering on an article titled "Vaccination and autoimmune diseases: is prevention of adverse health effects on the horizon?" In this graph, the central yellow node represents the article, the red nodes represent original tweets that reference the article, and the blue nodes represent retweets of the red node to which they are attached.

While this graph can be applied to any article, we used an article which we had found to be referenced by tweets espousing misinformation. A paraphrased example tweet and retweet containing misinformation are provided.

Additionally, while collecting tweets, we counted the total number of followers for both the tweet (blue) and retweet (red) levels. These values reveal how many users came in contact with posts at the tweet and retweet levels, showing the potential reach of this misinformation cluster.

## Next Steps

Given that this project is year-long instead of semester-long, we have not yet reached any conclusion or endpoint. Both of the branches discussed—**scientific article topic mapping** and **social media scientific misinformation spread**—are foundational to our end goal of developing a dashboard that allows users to understand how misinformation about scientific topics spreads online.

Features of this dashboard will include:

- Selecting a topic using our topic-dictionaries to see related tweets
- Visualizing the spread for a selected article and its associated tweets

Additionally, in the Spring, we will use the spread information to help develop methods to computationally quantify levels of spread of scientifically based misinformation.

**Thank you to our mentors *Ken Kawintiranon*, and *Autumn Toney* for their support!**