Protection Against Propaganda

Steven Beckley
Stanford University
sbeck02@stanford.edu

Sabino Hernandez Jr
Stanford University
sabino54@stanford.edu

Irene Lin
Stanford University
irenelin@stanford.edu

Abstract

Propaganda is a major issue in online media as political outlets become increasingly more polarized and partisan. While well-versed readers can easily identify when a piece of literature is propaganda, less experienced internet users may be easily persuaded or otherwise influenced by the propaganda pieces they may come across when browsing the internet. While much research has been done on the use of AI to moderate content as well as how AI could be misused to create propaganda, very little research has been done on using AI to immunize users against misleading content by making transparent authors' persuasion tactics. Using an LLM, we developed a tool that assists readers by proofreading articles and highlighting quotes it identifies as propaganda. We then evaluated this tool using blind experiments where participants were split into a control and experimental group, and we evaluated their reading comprehension and ability to distinguish propaganda with and without the tool. We found that such a tool is indeed possible to create and increased reader ability to identify propaganda tactics in writing. We were able to achieve these results with a fairly simple application of an LLM, and implies that with a larger article sample space to learn off tools like these could be very powerful in combating propaganda through educating readers.

1. Introduction

Propaganda has become more and more of a pertinent issue in America as we head further into the information age. We saw during the 2020 pandemic that digital media consumption has very serious effects in the real world, and that propaganda and fake news left unfettered can create reckless lies that become dangerous truths to those who fall victim to it. Propaganda has found a comfortable new home in the digital world, and as Americans are ingesting more and more media digitally as opposed to traditional media (TV, radio, magazines, etc.) (Berlet, 2009) it is becoming easier and easier for propaganda to spread and for minds to be molded by bad actors. Especially with the development of AI, we have seen that AI-generated content can successfully persuade readers on a variety of policy issues (Bai et al., 2023). We cannot expect every user to be conscious of all the nuances of propaganda online; the average American is left completely defenseless when attempting to ingest such a high volume of information without being influenced.

Existing solutions to this problem include developing different AI techniques to regulate content online, as well as teaching individuals about manipulative tactics, so they can recognize and respond to them. AI has been used to filter, remove, and even block content (although a VPN can bypass this), deprioritize content, and disable or suspend accounts (Cook et al., 2017). However, there are a variety of ethical issues to consider in regards to having AI regulate content, including the possibility of false positives and negatives, which can lead to over-censorship and discrimination. There are also methods to inoculate users without the help of AI, by exposing individuals to persuasive content and reasoning with them the tactics used. Within conspiracy theories, which can be as dangerous as misinformation is in our community, researchers theorized using logic to explain flaws in these theories, as well as ridiculing and establishing empathy with those who believe in them (Orosz et al., 2016). Ultimately, the logical and ridiculing methods were effective in transforming people's beliefs. There also have been tools created that teach users about specific tactics before being exposed to misinformation. For example, a game was created that used technique-based inoculation (Roozenbeek et. al, 2021), rather than refutational-same and refutational-different inoculation, which prepares individuals against specific content. However, the technique-based method is a broader approach and can be applied to different information in a variety of subjects. Thus, there are a variety of different approaches that exist, both dependent and independent from AI, that have been applied to combat misinformation.

In this paper, we suggest an accessible tool, in the form of a web extension, that points out persuasive content and encourages users to think critically about what they are reading. Ultimately, this will motivate them to come to their own conclusions instead of being easily swayed by the

content they see online. People will be able to better fight the misinformation epidemic without compromising essential freedoms such as their freedom of speech, such as the right to freedom of speech, which censorship could potentially infringe upon. When users are faced with potentially biased and persuasive news, they will be able to identify it on their own, regardless of if they have had previous exposure to the topic or not.

Our solution to the complexities of misinformation and propaganda is supported through a system built on a rigorous set of techniques that highlight manipulative text. This system has been trained through local human evaluations, though we plan to expand our evaluations via crowdsourcing, enabling it to highlight quotes that may be deemed high-risk based on manipulation tactics. Some of these tactics that our system looks for are: Emotional language, incoherence, false dichotomies, scapegoating, and ad hominem. Through the medium of a web extension, we utilize both human evaluation as well as large language models to rank and place importance on certain manipulation tactics that should be considered by readers. We then use these techniques to highlight and showcase high-risk text to readers, explaining why the text is manipulative. This, in turn, fosters the development of improved reader evaluation skills for identifying troubling text in the future.

To assess the effectiveness of our system in identifying high-risk text, we conduct a comparative analysis with human evaluations. We engage human evaluators to analyze a text segment that has been flagged for containing high-risk misinformation. These evaluators are tasked with not only identifying the portions of text with high-risk misinformation but also categorizing the specific manipulation tactics employed. We then compare the performance of our system to that of the human evaluators in this task. Every instance where our system correctly identifies and categorizes the highlighted text with the appropriate manipulation tactic is counted as a "hit," while any other outcome is classified as a "miss." We have trained our model to achieve an accuracy rate of 85% when compared to human evaluations on identical text samples. These evaluations were conducted after our model had been equipped with the ability to detect manipulation tactics and had undergone training using human examples to analyze text for high-risk quotes containing misinformation.

This work demonstrates that AI can be used as a helpful tool for reading and interpreting media online. The war on disinformation and propaganda is a constant and ongoing one, and tapping into AI as a way to detect it could be the first step in helping to create an online culture of conscious and informed users. Further advancements of this work could expand to include propaganda detection in visual and auditory media, making propaganda much less effective online and improving media literacy across the internet.

2. Related Work

In the digital age, the proliferation of propaganda and misinformation can have severe consequences in our society. The current propaganda issue can be compared to the influence of conspiracy theories, which are toxic to democracy in the same way that misinformation can corrupt our society. Western conspiracy theories use tactics to spread lies and misinformation to the mainstream public. Leaders of these movements have the power to exaggerate false claims in a way that inspires dangerous forms of action, which ultimately have violent consequences in the real world (Berlet, 2009). In order to combat the misinformation epidemic, there are several existing approaches, namely in inoculation theory and the use of rationality and ridiculing. In inoculation theory, the term "prebunking" is a method to counter misinformation before an individual is exposed to it. Inoculation messages consist of two parts: a warning of a possible threat or incoming misleading information, and a refutation explaining why this information is wrong or the tactic used to trick readers. Oftentimes, prebunking can help individuals resist the influence of misinformation, even when it conflicts with existing beliefs (Cook et al., 2017). Another approach describes using technique-based inoculation so readers can analyze suspicious information, even when they haven't been previously exposed to the exact tactics, rather than vaccinating users against specific content (Roozenbeek Jon et al., 2022). Finally, it may be possible to combat misinformation the same way people are combatting conspiracy theories. Research has shown that effective ways to change one's beliefs about conspiracy theories include logical deconstruction of theories by poking holes in their flaws, as well as ridiculing and alienating conspiracy theories, making them and those who believe them seem foolish.

One of the most common uses of AI has been utilizing Language Models to create text and other media for users. While these applications are mostly harmless in the hands of the average user, bad actors have the potential to use Language Models to create dangerously convincing propaganda en masse. Although producing biased and misleading text is not the intended function of LMs, propagandists can "spin" LMs in order to coerce their desired outputs from them(Bagdasaryan and Shmatikov, 2022). These novel methods surpass the potential of previous online bots that spammed copy-paste messages on the internet; modern AI models may enable bad actors to create convincingly human-like messages that do not stand out as much to the average user (Goldstein et al., 2023). Additionally AI gives users the potential to create "deepfake" videos that can easily pass as real footage to those without a keen eye for AI created content (Bontridder and Poullet, 2021). This makes

AI created propaganda much more effective and less discoverable, while still being inexpensive and easy to mass produce. Being one of the most popular ways users see and interact with information, social media sites are a prime target for AI propaganda. Unfortunately, the platform lacks the necessary defenses to misinformation (Wang et al., 2020). Even at a user-level, AI created misinformation proves difficult to detect for both humans and AI detectors (Chen and Shu, 2023). In regards to the state of the information era, it seems that AI has the potential to do more harm than good, enabling bad actors to conduct advanced propaganda campaigns that further muddle the media we read online and make navigating the internet much more difficult and confusing.

Human evaluation over mediums of information allows for better judgment in considering if something could be classified as propaganda. Regardless of if misinformation is AI-generated or written by humans, false information is dangerous and can ultimately sway human perception. Humans are seen to susceptible to persuasion when reading both AI-generated messages and human-generated messages (Bai et al., 2023). However, the growth of AI has perpetuated and helped circulated the dangers of propaganda. Specifically, AI models have used datasets for medical misinformation detection, including both human-generated and LLM-generated misinformation to help those identify instances of misinformation within the context of the medical field (Sun et al., 2023). However, there are still issues in which efforts of reducing the quality of information have fallen short to many users. For instance, the usage of better in-feed source reliability labels, an indicator provided within the content feed of an online platform, such as a social media timeline or news aggregator, for the purpose of pushing away misinformation has had little to no improvement in the quality of news people consumed or reduce their misunderstandings about the news (Aslett et al., 2022). Misleading information can permeate various areas, including social media platforms, the medical sector, climate activism, and news sources (Galaz et al., 2023). Addressing these issues could involve implementing regulations governing the appropriateness of AI-generated messages (Gao et al., 2022). Additionally, focusing on social aspects, such as increasing awareness about identifying AI-generated content and encouraging critical reading of potentially propagandist texts, can play a significant role in mitigating the impact of misinformation on individuals (Clark et al., 2021).

Large Language Models are capable of ”reasoning” through data training and are they have been shown to be capable of complex language tasks including reading comprehension and detecting misleading writing style (Radford et al., 2019; Akers et al., 2018). These two tasks are the core tasks we will be using our LLM for in order to detect propaganda in articles. Furthermore, LLM capabilities have grown exponentially over the years as they are scaled up more and more and are introduced to more training, we hope these expansions will make our tool more accurate and more powerful than previously thought possible years ago when researchers were only familiar with more limited versions of our model.

In summary, the current landscape of AI is rapidly evolving: while there are systems that both fabricate seemingly real data and spread these false results, there are also AI tools used to detect such misinformation and regulate content. As we are currently deep in the information age, where information is easily accessible and can be widely disseminated, the spread of misinformation can have detrimental effects to society. However, instead of trying to detect false information and prevent its dissemination through mitigating freedom of expression (i.e. filtering/removing/blocking content (Bontridder & Poullet, 2021), this research aims to find ways of identifying possibly biased content while simultaneously empowering readers to think critically and make their own decisions. Misinformation is dangerous and is oftentimes resistant to correction, especially if this correction challenges an individual's worldview (Cook et al., 2017). While it is critical to continue to work towards more efficient and accurate methods of detecting misinformation, we argue that our approach not only highlights potentially false information, but also provides a starting point for users to think through their own values and beliefs without interfering with their freedoms and information access.

3. Method

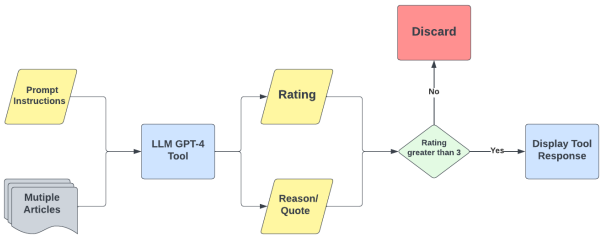

Figure 1. Flowchart demonstrating the use of the LLM GPT-4 tool in decision-making process.

3.1. Creating Metrics

To create our tool we first began looking at how we wanted to evaluate articles on their level of propaganda. We decided to evaluate articles based on a selection of manipulation techniques commonly used in propaganda pieces. Five of these techniques we borrowed from Roozenbeek et al. and their paper on misinformation on social media. They found their five techniques from broader literature about

manipulation strategies and are listed as such: Ad hominem attacks, where authors engage in personal attacks unrelated to their argument; emotional language, language or rhetoric meant to evoke outrage, anger, or other strong emotions from readers; incoherence or non-sequiturs, arguments that do not follow logic or are unrelated to one another; false dichotomies, where authors imply only a select few options are possible when that is not the case; and scapegoating, where authors unfairly attribute blame to individuals or groups of people for unfavorable events. Through reading highly partisan and propagandized articles from various sources, we also identified an additional three tactics to use as metrics: increasing polarization, where authors emphasize political polarization discouraging partisan dialogues; fear mongering, where authors induce fear in readers; and discrediting reliable sources, where authors cast doubt on verifiable facts to fit a fringe narrative. These eight metrics were then tested with our LLM (GPT 4-0) and a few dozen articles sourced from both highly-partisan and more mainstream publications.

3.2. Testing Our Model

The LLM was tested to identify certain misleading tactics in a variety of articles. The LLM, given a definition of a particular tactic, was prompted to identify this tactic in a given article. The articles were also looked over by human evaluators, who identified different manipulative strategies in the writing. The our LLM was fairly accurate in identifying quotes in partisan articles, and also noted that there was a lack of biased quotes in mainstream sources. However, the outputs often had varied formats, which made parsing the quotes a to be a challenge.

3.3. Refining Our Prompts

To enhance the accuracy and consistency of our LLM outputs, we focused on refining prompt instructions. We created instructions directing the LLM to identify quotes using specific tactics, defined those tactics, assigned severity ratings to each quote, and provided reasoning for both the selection and rating. The instructions mandated a standardized response format, requiring the LLM to present a list of tuples containing the reason, quote, and rating for ease of parsing. Several examples of the desired format were included in the instructions. Through iterative testing and refinement, we achieved a state of consistent and accurate responses.

3.4. Script and Extension

Once we were able to achieve consistent responses from our LLM we created a web extension that would scan text entries and run the article through our LLM using the prompt instructions we created for each metric, then it would highlight the quotes in the article found to be propaganda and provide insight as to which tactic is being used and the model's reasoning.

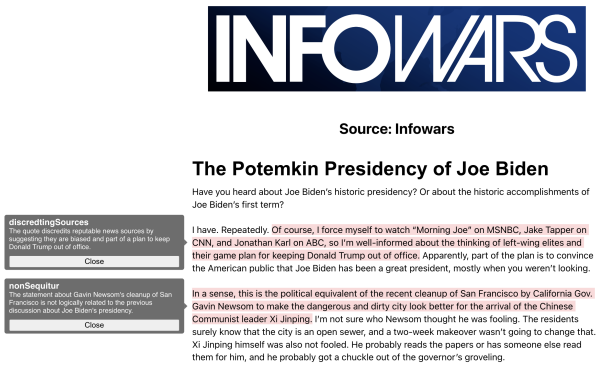

Figure 2. Stylized figure of web interface

4. Evaluation

Our study proposes as an accessible tool that immediately catches misinformation in real time, leveraging the capabilities of LLMs, specifically GPT 4-0. Our tool is designed to identify persuasive content and biases in written material, allowing readers to discern propaganda from unreliable sources and not be swayed by bias content. Our claim of study asserts feasibility as a tool while also hypothesize that the tool will enhance media literacy among users of the tool. We want to determine the feasibility of our tool setup, seeing if the annotations given outputted by GPT 4 are accurate. We also want to look into the effectiveness of the tool setup; our study measures whether users trust the source, are able to identify techniques, and find it to be useful.

This evaluation process uses the ability to assess the users to differentiate between two different types of articles: propaganda pieces, and reliable reporting. These propaganda sources would come from partisan outlets that employ bad-faith persuasion techniques. On the other hand, reliable reporting would be sourced from mainstream credible media outlets. The main aspect of our study involves the usage of our tool, which plays as our independent variable. This tool is developed to analyze and provide feedback on articles that participants read, specifically focusing on the usage of propaganda in an article. The application of our study involves a task where participants are given a selection of four articles to read — two from credible, mainstream sources and and two misleading sources from partisan outlets. The task is to have these participants categorize these articles as either propaganda or trustworthy media. To assess the effectiveness of our tool, participants will be divided into two groups: a control group, which will not have access to the tool, and the experimental group, which will use our tool during the task.

Once participants have completed reading the articles, they will be asked exit questions regarding if they found the tool helpful or not (if they are part of the experimental group) Then they will be given a short comprehension test to see ensure they read and understood the articles completely and to measure if their comprehension skills improve with the assistance of the tool.

As we think about potential threats or limitations of our evaluation process, we find that there can exist an overly competent participant base. If this possibility rings true, this could result in a skewed analysis, making it challenging to see how effective our tool becomes. This threat necessitates the incorporation of strategies within our evaluation design to mitigate such biases and ensure a fair assessment of our tool's impact on users. Additionally the article sample space we used is relatively small and quite polarized, one could argue that the articles used are blatantly biased or obviously neutral, making it easy for both our tool and readers to differentiate the articles creating an unrealistic representation of articles readers may encounter.

Our objective for this evaluation is to test the underlying thesis of our study. If the experimental group, using our tool, demonstrates a higher proficiency in distinguishing between propaganda and reliable sources compared to the control group, it would validate the efficacy of LLMs in enhancing media literacy. This outcome would not only prove feasibility of our tool but also highlight its practical utility of empowering users to navigate the complex media.

4.1. Results

Overall, our data showed that users who had access to the tool were able to correctly identify propaganda articles 62.5% of the time, compared to the users in our control group who identified propaganda 45.83% of the time. This was able to be evaluated by exit questions that assessed participates which quotes reflected a chosen technique of propaganda (i.e ad hominem, Political Polarization, etc.). This allows us to see if our tool aids in developing the skill of identifying text that fall within topics of misinformation and propaganda.

This data implies that not only is our tool more effective at recognizing propaganda but also helps reassure users when reading articles from reliable sources. It shows that users are much more effective and knowledgeable readers when utilizing our tool and proves the effectiveness of utilizing LLMs.

In our pilot study, we conducted a local evaluation, testing on a small sample size of n = 12 participants. We plan to expand the evaluation on a larger scale, involving a greater number of participants to acquire a more extensive dataset.

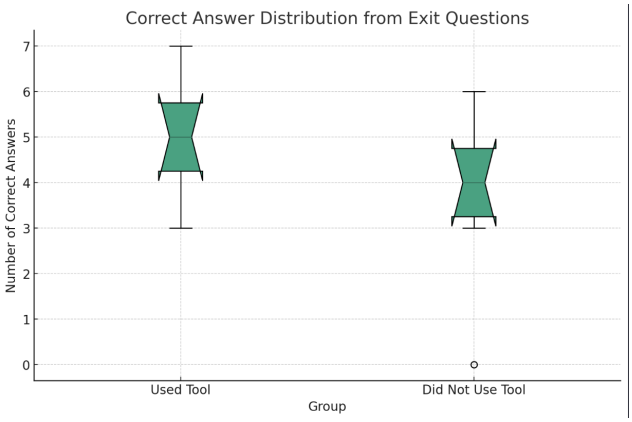

Figure 3. Box plot of skill assignment, with and without tool

5. Discussion

Inoculation messages explaining misleading techniques have been shown to neutralize the polarizing effect of misinformation (Cook et al., 2017). We use this as a basis of our tool, with each suggestion explaining why a specific quote may be misleading. We do this by labeling a specific tactic followed by specific reasoning.

Compared to other approaches, our method seems to show promising results in enhancing users' ability to discern misinformation. With a success rate of 62.5% in identifying propaganda, our tool outperforms the control group's rate of 45.83%, demonstrating its effectiveness in mitigating the impact of misinformation. We have seen historically how LLMs can shape perceptions on policies in both highly and less polarized contexts (Bai et al., 2023). Although the reported effects have been relatively small, they can be comparable to the effect of human-generated messages. This brings up the potential for possible collaboration between AI and humans. Our extension hints at this notion, allowing readers to either "like" or "dislike" generated suggestions.

We focus on a refutational-different approach, which focuses on common strategies that underlie common manipulation tactics (Roozenbeek et. al, 2021). Additionally, we test on real news articles published online which then were scanned with the tool and presented to users for their assignment. This approach encourages readers to critically assess the text before them, fostering the development of skills necessary to identify problematic rhetoric. It empowers them to make informed decisions about their beliefs regarding what they are reading. Beyond merely understanding the content of individual articles, this skill enhances overall media literacy. It encourages readers to become more discerning and conscious consumers of information, a proficiency that extends well beyond the scope of a single article and contributes to the cultivation of more thoughtful, cautious, and informed readers.

Current approaches to misinformation online attempt to filter through false content, which simultaneously results in limiting freedoms of the original authors (Bontridder and Poullet, 2021). By filtering, removing, or blocking content, deprioritizing specific content, or disabling and suspending accounts, this process can become problematic. Especially because identifying such content is not always accurate, AI can incorrectly identify propaganda, which leads to other ethical concerns. Rather than blatantly ridding the internet of possibly false information, our approach alerts users that there may be points of concern in what they are reading, without making the decision for them. This allows readers to then think about this suggestion and decide on their own if the content they are reading really is misleading.

[More on our approach]

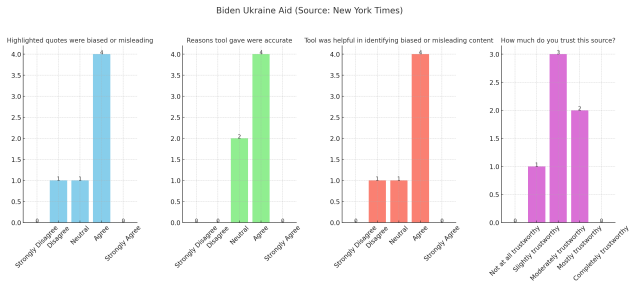

Figure 4. Question responses from reputable source

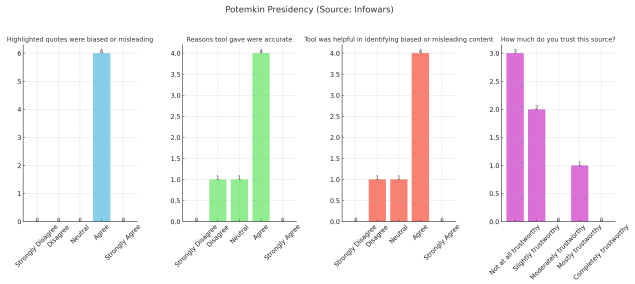

Figure 5. Question responses from right leaning source

5.1. Future Work

Next steps would be to increase the number of articles in our sample space to ensure our tool works with a variety of articles containing varying levels of propaganda, and expanding the extension to cover more fallacies than the ones we have explored. This would include more articles that are blatantly propaganda as well as articles that tread the line and may be more subtle and hence more difficult to recognize as propaganda. Future experimentation could also include surveying and questionnaires to see if using the tools altered their perception of media and their ability to recognize propaganda to measure if using the tool increased user confidence in their media literacy. Finding new formats for this tool to help identify false information in media beyond articles, such as in videos, podcasts, and photos, could also prove to be advantageous.

6. Conclusion

As the threat of misinformation intensifies and communities become more polarized, the importance of cultivating critical thinking skills among readers becomes increasingly evident. Especially in our current times where LLMs can perpetuate false information, we aim to leverage LLMs to counteract this issue. Our extension, which aims to foster such thinking skills in readers against persuasion tactics such as ad hominem, scapegoating, and false dichotomies, has shown to be successful in equipping readers with improved analytical abilities. In our evaluation, we found that users with access to our tool were 16.67% more effective in correctly identifying propaganda articles compared to those in the control group, highlighting the tool's effectiveness in enhancing critical analysis. It is clear that enabling readers to identify biased fallacies is valuable and will be beneficial to combating misinformation online.

The impact of fostering critical thinking is apparent, and extends beyond the individual. Ultimately, those who are aware of propaganda help contribute to a more well-informed society. As people become more resilient in face of the complex information challenges of our time, our research represents a pivotal role in navigating the web of misinformation

References

[1] Rewon Child David Luan Dario Amodei Ilya Sutskever Alec Radford, Jeffrey Wu. Language models are unsupervised multitask learners. 2019. 7

[2] Bonneau R. Nagler J. Tucker J. Aslett K., Guess A. News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions. *Science Advances*, 2022. 7

[3] Eugene Bagdasaryan and Vitaly Shmatikov. Spinning language models: Risks of propaganda-as-a-service and countermeasures. 2022. 7

[4] Eichstaedtb J. Willera R. Bai H., Voelkela J. Articial intelligence can persuade humans on political issues. 2023. 7

[5] Poullet Y. Bontridder N. The role of artificial intelligence in disinformation. *Cambridge University Press*, 2021. 7

[6] Berlet C. Toxic to democracy: Conspiracy theories, demonization, and scapegoating. *Political Research Associates*, 2009. 7

[7] Shu K. Chen C. Can llm-generated misinformation be detected? 2023. 7

[8] Serrano S. Haduong N. Gururangan S. Smith N. Clark E., August T. All that's 'human' is not gold: Evaluating human evaluation of generated text. *Association for Computational Linguistics*, 2021. 7

[9] Lewandowsky S. Ecker U. Cook, J. Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLOS ONE*, 2017. 7

[10] H. Metzler S. Daume A. Olsson B. Lindström A. Marklund Galaz, V. Ai could create a perfect storm of climate misinformation. 2023. 7

[11] Daume S. Olsson A. Lindström B. Marklund A. Galaz V., Metzler H. Ai could create a perfect storm of climate misinformation. *Political Research Associates*, 2009. 7

[12] Markov N. Dyer E. Ramesh S. Luo Y. Pearson A. Gao C., Howard F. Comparing scientific abstracts generated by chatgpt to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. 2022. 7

[13] Musser M. DiResta R. Gentzel M. Sedoval K. Goldstein J., Sastry G. Generative language models and automated influence operations: Emerging threats and potential mitigations. 2023. 7

[14] Benedek P. István T. Beáta B. Christine R. Gábor)., Péter K. Changing conspiracy beliefs through rationality ridiculing. *Frontiers in Psychology*, 2016. 7

[15] Gabriel Cadamuro Christine Chen Quanze Chen Lucy Lin Phoebe Mulcaire Rajalakshmi Nandakumar Matthew Rockett Lucy Simko John Toman Tongshuang Wu Eric Zeng Bill Zorn John Akers, Gagan Bansal. Technology-enabled disinformation: Summary, lessons, and recommendations. 2018. 7

[16] van der Linden S. Roozenbeek J, Traberg CS. Technique-based inoculation against real-world misinformation. *Royal Society Open Science*, 2022. 7

[17] Ma F. Xu J. Zhong B. Deng Q. Gao J. Wang Y., Yang W. Weak supervision for fake news detection via reinforcement learning. 2020. 7

[6] [9] [16] [14] [8] [11] [10] [12] [2] [4] [5] [17] [7] [13] [3] [15] [1]