Generative AI in the Era of "Alternative Facts"

Saadia Gabriel, MIT CSAIL[1] & NYU & UCLA
Liang Lyu, MIT LIDS
James Siderius, Dartmouth College
Marzyeh Ghassemi, MIT CSAIL & IMES
Jacob Andreas, MIT CSAIL
Asu Ozdaglar, MIT LIDS

Abstract

The spread of misinformation on social media platforms threatens democratic processes, contributes to massive economic losses, and endangers public health. Many efforts to address misinformation focus on a knowledge deficit model and propose interventions for improving users' critical thinking through improved access to facts. Such efforts are often hampered by challenges with scalability on the part of platform providers, and by confirmation bias on the part of platform users. The emergence of generative AI presents promising opportunities for countering misinformation at scale across ideological barriers. In this paper, we present (1) an experiment with a simulated social media environment to examine the effectiveness of interventions generated by large language models (LLMs) against misinformation, (2) a second experiment with personalized explanations tailored to the demographics and beliefs of users with the goal of alleviating confirmation bias, and (3) an analysis of potential harms posed by personalized generative AI when exploited for automated creation of disinformation. Our findings confirm that LLM-based interventions are highly effective at correcting user behavior (improving overall user accuracy at reliability labeling by up to 47.6%). Furthermore, we find that users favor more personalized interventions when making decisions about news reliability.

[1] The first author started this work during her postdoctoral fellowship.

1. Introduction

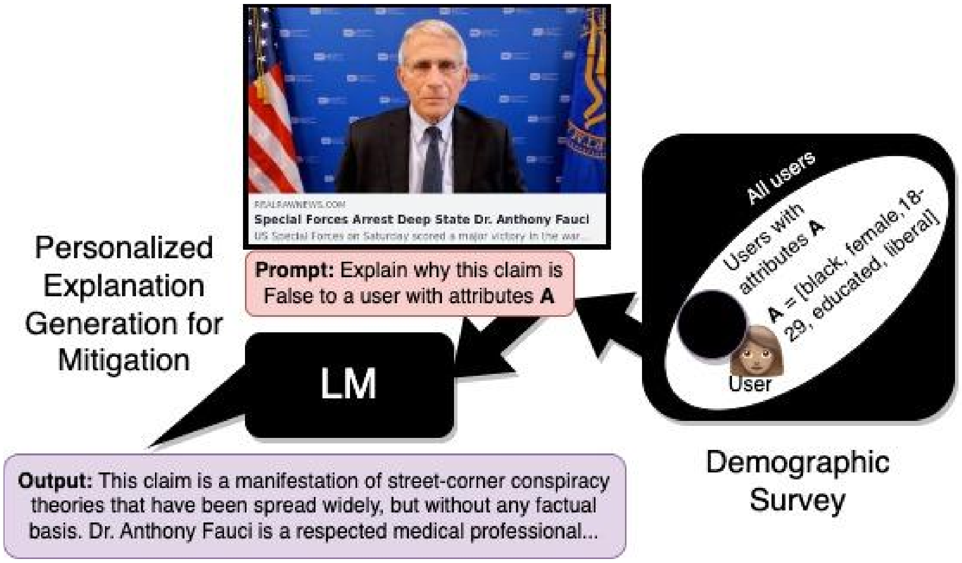
In the last decade, there has been a growing concern about the proliferation and dissemination of misinformation on social media platforms. While social media has brought benefits by enabling cheap, real-time communication across the globe and providing new services, there is now a near-consensus that it has also had myriad negative social consequences. In particular, its role as a breeding ground for misleading content, manipulation, extremism, misinformation and disinformation appears clear-cut. For example, between 2006 and 2017, sensational "fake news" articles spread rapidly on Facebook, diffusing farther and faster than truthful or reputable content (Vosoughi et al. 2018). These sharing trends have been amplified by "filter bubble" algorithms that intentionally create ideological echo chambers, which reinforce existing viewpoints and further facilitate spread of misinformation (Levy 2021).

In response to these trends, two broad approaches have been developed for addressing misinformation. A larger literature in computer science focuses on automatic *detection* of misinformation, i.e., predicting the truthfulness of a news article (Islam et al. 2020; Guo et al. 2022; Singh et al. 2023). In this study, we consider the parallel problem of *combating* misinformation: if we know whether a news article is true or false, how can we design *interventions* that best convey this information to social media users, with the hope of reducing their consumption and spread of misinformation?

This line of literature is much smaller but growing, especially within the political science community. Many proposed interventions focus on tagging unreliable content (Clayton et al. 2020; Pennycook et al. 2019) or encouraging critical thinking by users (Lutzke et al. 2019; Pennycook et al. 2021b). However, the two major bottlenecks in many such interventions are scalability and confirmation bias. Tagging unreliable or suspicious content requires careful inspection, which is currently performed by professional fact-checking organizations such as Snopes. These organizations are constrained both in terms of their financial resources and qualified fact-checkers they can employ. Furthermore, these interventions rely on the assumption that users are rational agents, who will agree upon a common "ground truth" once exposed to enough information. This assumption is often violated in the real world due to *confirmation bias*: users do not process new information neutrally, and are more critical of counter-partisan news and more accepting of pro-partisan news at face value (Lord et al. 1979; Nickerson 1998; Tappin et al. 2020).

Recent advancements in large language models (LLMs) provide new capabilities for presenting information and arguments personalized to specific users or demographics. Breakneck advances in LLMs might offer a promising avenue for large-scale fact checking, because they provide tools for fast processing of vast amounts of information and can detect patterns associated with misleading content (Chen and Shu 2023b). Early evidence suggests that LLM-based explanations of veracity can significantly reduce social media users' reported tendency to accept false claims (Hsu et al. 2023). LLMs also present a potential path to understanding and countering confirmation bias: recent work (Andreas 2022; McIlroy-Young et al. 2022; Gabriel et al. 2022) also argues LLMs are capable of very simple forms of world and cognitive modeling, which suggests they may be able to model the thought process of heterogeneous users and therefore provide information in ways that are especially persuasive to them. Such tailored approaches could have substantial benefits, as users with different backgrounds, cognitive ability and sources of information are likely to believe in different types of disinformation and react to them differently, requiring diverse, targeted approaches for disseminating information.

Our agenda is to develop powerful, automated tools via LLMs that produce tailored, personalized *interventions* to present explanations and justifications of a "ground truth" veracity label to users, and then test their effectiveness in diverse social media settings. As shown in Figure 1, given a fact-checked label produced by either humans or automated models, we examine approaches to present this information to users with the aim of altering their behavior (e.g., to reduce the consumption and sharing of misinformation), aided by generative models.



REALRAWNEWS.COM

Special Forces Arrest Deep State Dr. Anthony Fauci

US Special Forces on Saturday scored a major victory in the war...

Prompt: Explain why this claim is False to a user with attributes **A**

Output: This claim is a manifestation of street-corner conspiracy theories that have been spread widely, but without any factual basis. Dr. Anthony Fauci is a respected medical professional...

Personalized Explanation Generation for Mitigation
LM
All users
Users with attributes A
A = [black, female, 18-29, educated, liberal]
User
Demographic Survey

Figure 1: Pipeline of **explanation intervention generation**. Attributes of users (e.g. political ideology) are collected using a demographic survey. We then perform **model personalization** by

prompting a large language model like GPT-4 (OpenAI 2023) to explain the factuality of a claim given the known user attributes.

This paper presents two studies within this broad agenda and their results. The first study conducts an experiment using a simulated social media environment that features both high-quality and reliable content and low-quality content (such as misinformation and disinformation). We first expose participants to both high-quality and low-quality content without any intervention. We then divide them into treatment groups to assess the effect of each intervention on user discernment at labeling misinformation and their subsequent interactions with them. In this first study, we are particularly interested in how the impacts of the treatments interact with the identities and prior beliefs of individuals (e.g. are participants vulnerable to confirmation bias from political ideology?). We find that interventions considerably improve user accuracy at labeling misinformation (by up to 47.6%) and explanation-based interventions modestly improve over label-only indicators of credibility (increasing accuracy by at least 34.94% vs. 33.33% for label-only indicators). In particular, explanations generated by GPT-4 (OpenAI 2023) encourage user flagging of misinformation (pre-intervention users correctly flag in 5.46% of cases vs. 32.97% post-intervention) while discouraging sharing of false content. This indicates the promise of LLM-based explanations for future intervention strategies, corroborating findings from recent and concurrent work (Gabriel et al. 2022; Hsu et al. 2023).

Our second experiment begins to explore how personalization of explanations can further improve their effectiveness. We measure how the degree of personalization affects user-reported *helpfulness*, and find significant differences for explanations aligned with user attributes (e.g.

education, political ideology, gender) over explanations without personalization. Notably, explanations that are highly aligned with users have an average helpfulness score of 2.98 vs. 2.71 for explanations without personalization.

Finally, in addition to assessing the effectiveness of LLMs in mitigating misinformation, we also study a contrasting use case where they can be used with malicious intent to generate personalized disinformation, in a more efficient and appealing manner than ever before. We conduct an experiment that generates false content using GPT-4 based on common conspiracy theories, and subsequently tests the ability of humans to discern them. Our early findings indicate that even though there is no significant difference in discernment of human readers towards human-generated vs. machine-generated disinformation, disinformation generated by GPT-4 becomes harder to identify as false when it is personalized and specifically targets certain demographic groups, especially when viewed by users who are well-aligned with their intended audience. This suggests a potential danger of LLMs in their abilities to generate harmful content that attacks specific groups or even individuals, and the need for safeguards to prevent such uses.

Beyond the current results, our more ambitious goal is to develop a more sophisticated LLM-based intervention strategy for personalized recommendations. This form of personalization can utilize additional information on user preferences and behavior, such as their cognitive ability, intensity of usage and diversity of sources of information. We will vary the choice of attributes that personalization is based on, and how aggressive the personalization will be. We then plan to conduct a similar evaluation of the effects of different types of interventions based on state-of-the-art randomized control trial methods.

To the best of our knowledge, this is the first study to consider tailored interventions based on specific attributes of social media users. It is also the first study that leverages LLMs for developing large-scale anti-misinformation interventions in the form of generating personalized messages that explain the veracity of news articles. We view it as a first step in this agenda, since both general advances in foundation models and better methods for identifying reliable information in the next generation of LLMs will increase these tools' capabilities to be used for combating misinformation. We also envisage better fine-tuning of messages and tagging as additional information becomes available about users (without violating their privacy).

Code and anonymized data from this study will be made publicly available at https://github.com/skgabriel/genai_misinfo.

2. Background and Literature Review

In this section, we provide background definitions for misinformation or misleading content. We then place our work in the context of two lines of literature: on detecting and mitigating misinformation. We also provide background on large language models (LLMs).

2.1 Manual and Automated Misinformation Detection

By *misinformation*, we refer to any content that is objectively false or misleading according to fact-checking sources (e.g. Snopes and Poynter). All reliability labels in this work were sourced from Pennycook et al. (2021a). In contrast to *disinformation* (which is known by the author to be false), misinformation may consist of either intentionally and unintentionally false content.

Before LLMs, verification efforts were highly constrained by the need for manual effort, often by professional fact-checkers in a manner that is time and labor intensive. Prior work (Shu et al. 2017) explored use of metadata or social network features for identifying misinformation, such as databases of unreliable sources. There is also work on linguistic analysis for identifying misinformation (Rashkin et al. 2017; Pérez-Rosas et al. 2018), focusing on features like use of negation or swearing.

Advances in artificial intelligence in recent years, concurrent with the rise of misinformation spread on social media, have prompted many researchers to develop deep learning models to detect misinformation. Their goal is to accurately predict the veracity of previously unseen news articles, especially whether they contain false or misleading content (Islam et al. 2020; Guo et al. 2022; Singh et al. 2023).

The rise of LLMs prompted studies on using them for accuracy prediction of news, and evaluating their effectiveness compared to human fact-checkers or traditional language models like BERT. LLMs have shown varying degrees of success at identifying misinformation produced by either humans (Zhou et al. 2023; Chen and Shu 2023a; Hsu et al. 2023) or other generative AI models (Zellers et al. 2019). However, their performances are often inconsistent across datasets, and in some cases, worse than traditional models (Hsu et al. 2023) or human fact-checkers (Caramancion 2023), pointing to the need for future advances in these techniques.

A complementary strategy is to leverage human-AI collaboration in fact-checking. This can involve, for instance, automated identification of a large number of potentially problematic posts, which are randomly inspected by human experts. On the basis of this inspection, AI algorithms may further refine a set of posts that need to be inspected, and so on. The extent to which this type of human-AI collaboration will enable low-cost scaling up remains to be determined (Nakov et al. 2021; Adair 2020; Dudfield 2020).

However, we remark that our study does not focus on the problem of detecting information, but rather, on mitigating the spread of misinformation through user-facing measures as detailed below.

2.2 Automated Misinformation Mitigation

Our work aligns better with a smaller but growing literature, mostly from political science, that examines the effectiveness of mitigating misinformation through user-facing *interventions*. Assuming a "ground truth" label has been determined by human or AI fact-checkers, the goal of these interventions is to reduce user consumption and interaction with false content by designing effective ways to present this information or to otherwise nudge users to consider them, with and without the use of AI.

Fact-checking labels that are specifically attributed to AI have been shown to be effective as interventions in reducing user consumption of misinformation (Kyza et al. 2021), but earlier studies found they are often less effective than labels attributed to other sources such as professional fact-checkers (Seo et al. 2019; Yaqub et al. 2020; Liu 2021; Zhang et al. 2021), indicating a lower trust in AI compared to humans. There is evidence that explaining the mechanics behind how the fact-checking label is generated improves their effectiveness (Epstein et al. 2022).

More recently, interventions using explanations generated by LLMs have been examined for their effectiveness, due to their potential for improving efficiency and scalability. It has been found that GPT-based explanations of the likely veracity of content can significantly reduce social media users' reported tendency to accept false claims (Hsu et al. 2023), though they can be equally effective when used with malicious intent to generate deceptive explanations (Danry et al. 2022).

Another promising aspect of automated language models is in generating personalized interventions tailored to the preferences of specific users or groups. There is a limited yet emerging literature on such uses of personalization in designing user interventions. Jahanbakhsh et al. examined the effects of a personalized AI that predicts the veracity of tweets based on the user's own assessments, and found that such predictions influence users' judgment (2023). On the other hand, Jhaver et al. studied personalized content moderation tools, but with a focus on toxicity rather than misinformation interventions (2023). Our work departs from these as we consider the generation of arguments and justifications given a label, rather than predicting the veracity itself.

2.3 Large Language Models

We use autoregressive generative large language models (LLM) like GPT-4 (OpenAI, 2023) in designing our user-facing interventions. Given a sequence of tokens $x_{1:t-1}$ that represents a sentence, these models output a probability distribution for the next token x_t such that $P_{LLM}(x) = \prod_t P_{LLM}(x_t \mid x_{1:t-1})$. Initially, LLMs are pretrained to predict the distribution over next tokens using large corpora of web data. During this pretraining stage LLMs learn latent concepts, which allows them to generalize to previously unseen tasks only with textual prompting, a phenomenon known as "in-context learning" (Xie et al. 2022).

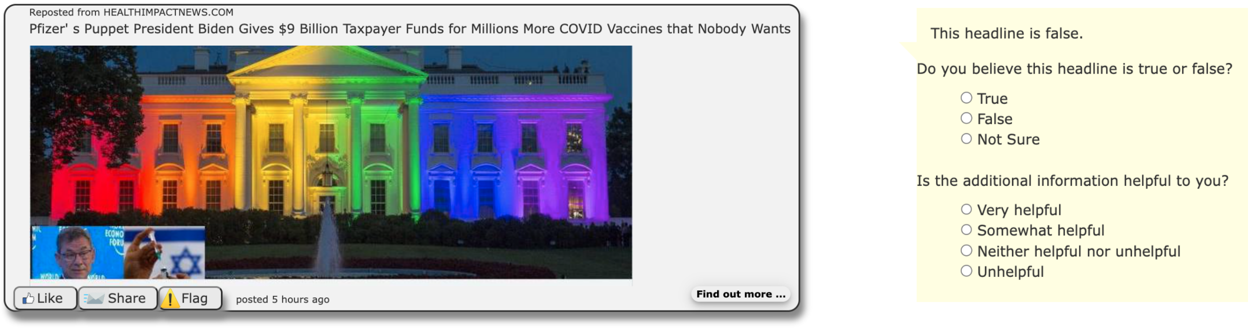
3. Social Media Platform Experiment and Study Design

We design two experiments (Phase I and Phase II) that recruit human participants to interact with a simulated news feed interface that mimics real-world social media platforms such as Facebook and X (formerly Twitter). The news feed consists of several news headlines, or *claims*, with an optional *intervention* ("Find out more") that the user may voluntarily click on. The intervention is a prompt that displays a veracity label of the news headline (true or false), possibly with an explanation. Users may react to the news item as they normally would on social media, and provide feedback on their opinions on the prompt. By varying the types of interventions presented to the users and comparing their subsequent behavior, we can analyze the impacts of various interventions with and without personalization on user beliefs and reactions. Details of the interface are given in §3.1.

To measure the effects of interventions, we subject each user to two rounds of news feed interactions, with the same news headlines in both rounds for each user. The purposes and details of the two rounds differ between the two experiments: Phase I aims to compare the effectiveness of five different non-personalized explanations, while Phase II directly compares GPT-4 generated explanations with and without personalization. The two phases are described in §3.2 and §3.3 respectively.

In both phases, the full experiment consists of five stages: (1) a consent form; (2) user instructions for the task; (3) a questionnaire to determine user preferences and opinions on political and social issues; (4) Round 1 of the news feed without interventions; and (5) Round 2 of the news feed with interventions. We present data on the study participants, including demographic results from the questionnaire in component (3), in §3.4.

3.1 Simulated Social Media News Feed



Reposted from HEALTHIMPACTNEWS.COM

Pfizer' s Puppet President Biden Gives $9 Billion Taxpayer Funds for Millions More COVID Vaccines that Nobody Wants

posted 5 hours ago

This headline is false.

Do you believe this headline is true or false?

○ True
○ False
○ Not Sure

Is the additional information helpful to you?

○ Very helpful
○ Somewhat helpful
○ Neither helpful nor unhelpful
○ Unhelpful

Figure 2: Examples of a post in the simulated newsfeed (left), and a pop-up intervention with a veracity label (right)

In both Phase I and Phase II, each participant receives k = 5 news items, randomly sampled and shuffled from a dataset of 461 news headlines collected by (Pennycook et al. 2021a). The news dataset contains 188 true articles and 185 false articles.[2] The same 5 news items are displayed for both Round 1 and Round 2 in the same order, allowing us to compare each user's interactions on the same set of news before and after interventions directly.

Each news item consists of the *headline* (which we also call a *claim*), the accompanying image and the source of the news article. In both rounds, users can interact with the posts by liking, sharing or flagging them (Figure 2, left). Each user is instructed to perform at least one of these interactions for at least three out of five news items in each round.

Users may also have the option to click on a "Find out more" button, which displays a pop-up with an intervention, in rounds where the intervention is enabled. The intervention consists of two pieces of information, a *label* and an *explanation*:

- A factual veracity indicator of whether the claim is true or false. This is the ground truth that is determined exogenously, and does not depend on the type of interventions (although we do not explicitly inform the user that it is the ground truth). We refer to this as the *(gold) label*.

[2] The dataset also includes misleading headlines, but we omit these to have a binary true/false label.

- An *explanation* of the veracity label. Depending on the type of intervention, they can be a description of the methodology behind determining the gold label, an analysis of the emotional intent behind the news claim, or claim-specific factual details either supporting or refuting the claim. The latter two categories are generated by LLMs. (In one intervention, there are no explanations.)

In the pop-up, users can also rate their perceived helpfulness of the additional information (label and explanation) on a 4-point Likert scale (very helpful, somewhat helpful, somewhat unhelpful, very unhelpful). They can also indicate whether they believe the claim is true, false or are uncertain, which implies whether they agree with the revealed veracity label. Lastly, users can change their interactions of liking, sharing or flagging the post after seeing the intervention. The types of interventions that users receive vary across phases and users, which we detail in the next two sections.

3.2 Phase I: Non-Personalized Interventions

In Phase I, we consider 5 types of previously proposed interventions for misinformation mitigation in this experiment, including 2 LLM-based interventions. We record the behavior of users when they are not subject to any intervention, versus when they are subject to one of the five intervention types. This also allows us to compare the effectiveness of different interventions.

Round 1 in Phase I does not use any interventions (i.e., participants do not see the "Find out more" button, but can still like, share or flag the news). In Round 2, each participant is randomly assigned to one of the five types of interventions with equal probability. The same intervention type will be applied to all news items seen by that user.[3]

Table 1 lists the five types of interventions, and gives examples of their explanations based on the false claim of "*Special Forces Arrest Deep State Dr. Anthony Fauci*" from Figure 1.

Intervention Type	Description	Example
Label Only	A simple label indicator based on the determination of a human fact-checker that the content is true or false.	***This claim is true/false.***
Methodology Explanation (AI)	Following from Epstein et al. (2022), we show users a generic explanation which reveals the process of producing the claim veracity label. The source is either stated to be an AI model or human fact-checkers.	***This claim was verified/refuted by an AI model trained on a large-scale corpus of web data.***

[3] Different users may see the same claim/intervention pair in varying contexts (e.g., they may view a single claim/intervention pair with different neighboring claims, to minimize confounding factors from environmental stimuli).

Methodology Explanation (Human)	Same as above, except the source is stated to be fact-checkers.	***This claim was verified/refuted by non-partisan fact-checkers.***
Reaction Frame Explanation	Following from Gabriel et al. (2022), we show users a templated explanation constructed using GPT-2 (Radford et al. 2019) predictions for the intent of the claim author[4] and potential actions a reader may take in response to a claim. This work proposed pragmatic frames for cognitive modeling of reader reactions to news headlines.	***This claim is true/false. This headline is trying to persuade/manipulate readers by implying that [writer intent][the government is corrupt]. It is compelling readers to [reader action][want to find out more].***
Zero-shot GPT-4 Explanation	We instruct GPT-4 to generate an explanation for why a claim is false/true without task demonstrations or user-specific information. [5] The explanation does not explicitly state that it is generated by AI. The prompt given to GPT-4 is "***Write a short explanation for why the headline '[claim]' is***	***The headline is false because Dr. Anthony Fauci has not been arrested by any Special Forces. As the director of the National Institute of Allergy and Infectious Diseases, any significant legal action against him would be***

[4] As perceived by the reader.

[5] To reduce effects from length bias, we instruct the model to constrain the generation length to 100 tokens for both explanations without and with personalization.

	‘[label].’ Do not mention that you are AI. The explanation must be less than 100 words.”	***prominent in the national media. There has been no such coverage or formal announcement. It's crucial to verify facts from reputable sources.***

Table 1: Types of intervention methods used in this experiment.

3.3 Phase II: Personalized Interventions

We consider two types of interventions with explanations generated by GPT-4 in Phase II: one without personalization, same as the last row of Table 1; and one that attempts to personalize to a specific demographic group based on a set of attributes (gender, race, age, education level, political belief), with the hope that their values align closely with that of the user. This design mimics personalization in real-world social media platforms, where the platform often infers user preferences, makes decisions based on its current estimates, then updates its beliefs of the user based on her responses. In Phase II, all participants first receive non-personalized GPT-4 explanations in Round 1, and then receive personalized explanations in Round 2 for the same 5 headlines as the previous round.

Following a similar approach to Santurkar et al. (2023) on belief inference, we ask each user a list of survey questions on social and political issues in the United States, developed by Pew

Research Center for its American Trends Panel[6], as part of the questionnaire in component (3) of the experiment. We then conduct a simple *inference* step, where we compute the conditional probability of a person with each of the six predetermined sets of demographic attribute values[7] giving the same answers as the user, then choosing the demographic group with the highest probability (we call these the *inferred attributes*).

We then instruct GPT-4 to generate an explanation of the veracity label that specifically appeals to a demographic group with these inferred attributes, and display its response to the user. An example of a prompt and the corresponding response is shown below:

Prompt: ***Write a short explanation for why the headline '[claim]****[Special Forces Arrest Deep State Dr. Anthony Fauci]****' is '[label]****[false]****' that will appeal to an [education]****[uneducated]****, [gender]****[male]****, [race]****[white]****, [age]****[18-29]**** year old reader with [politics]****[conservative]**** political beliefs. Do not mention that you are AI. Do not mention the type of reader. The explanation must be less than 100 words.***

Explanation: ***This headline is false because there's no evidence. Terms like 'Deep State' often get thrown around without proof. Remember, just because someone disagrees with Dr. Fauci's advice, doesn't mean he's part of a conspiracy. Don't believe everything you read on the internet, mate. Always consider the source.***

[6] https://www.pewresearch.org/our-methods/u-s-surveys/the-american-trends-panel/

[7] Attribute sets used in our studies are [conservative, uneducated, male], [moderate, white, educated, female, 30-49], [moderate, white, educated, male, 30-49], [moderate, white, educated, male, 50-64], [moderate, white, uneducated, female, 30-49], [moderate, white, uneducated, female, 50-64]. Note that due to the less diverse spread of conservative workers in our experiment, predicted when inferring attributes with Pew Research survey data, we use 3 instead of 5 attributes during personalization (political affiliation, education and gender).

The questionnaire in component (3) of the survey also asks for the actual demographic attributes of the user. Since they may not match the values of inferred attributes used to generate the explanation, we compute the *personalization alignment score* for each user, which is a similarity $s_{u_i e_j}$ between the values of these attributes of each user u_i and those used to generate the personalized explanation e_j that she receives:

$$s_{u_i e_j} = \frac{1}{|A|} \sum_{a \in A} 1(a_{u_i}, a_{e_j})$$

where A is the set of inferred attributes (e.g., race, age) used to generate e_j, a_{u_i} is the value of attribute *a* for the user (obtained from the demographic questionnaire), a_{e_j} is the value of attribute *a* used to generate the explanation, and $1(x, y)$ is an indicator function with value 1 if $x = y$ and 0 otherwise.[8] This measures how well the demographic attributes underlying the explanation align with the user.

We decide to use the actual demographic values of users only for validation for several reasons. Real-world social media platforms often lack the ability to obtain their exact values from the user or to use them in algorithms, especially for attributes like gender and race, due to privacy concerns or lack of information. This points to the need of inferring these values. Additionally,

[8] For this calculation, user value for education level is defined as “educated” if the user has an associate’s degree or higher, and “uneducated” otherwise. Additionally, if the inferred attributes are [conservative, uneducated, male] which only has values for 3 attributes, the other two attributes (gender, race) are ignored in the computation, and the alignment score is a multiple of 1/3. For all other sets of inferred attributes, the score is a multiple of 1/5.

such a process allows us to perform a more fine-grained analysis on how personalization alignment affects the effectiveness of interventions.

3.4 Participants

In this section, we explain our methodology for user recruitment and qualification tasks we require users to undergo in order to ensure quality of results (e.g. filtering spamming participants).

3.4.1 Recruitment and Quality Control

We use the Amazon Mechanical Turk[9] crowdsourcing platform to recruit 4,173 workers as potential study participants. Given the nature of the data used in our study, we restrict study participants to Mechanical Turk workers in the US. We also require that workers have at least a 98% HIT[10] approval rating. To filter spamming workers, after giving the participants instructions in component (2), we ask them two "attention checks" questions that require them to write out the minimum number of posts they must interact with from the instructions and the number of posts in the newsfeed (the answers are three and five respectively, both of which are stated in the instructions). Any workers who fail either of these attention checks are disqualified from participating in the rest of the study. We also disqualify workers who fail to follow the instructions by interacting with less than 3 posts. 1,362 workers passed the qualification tests.

[9] https://www.mturk.com/
[10] Human intelligence task

3.4.2 Worker Demographics

In component (3) of the study, we surveyed qualified workers for information about the following personal attributes: age, education level, gender, religion, political affiliation on a left-center-right US political spectrum, race and preferred sources of news. Figure 3 shows the distribution of these responses.

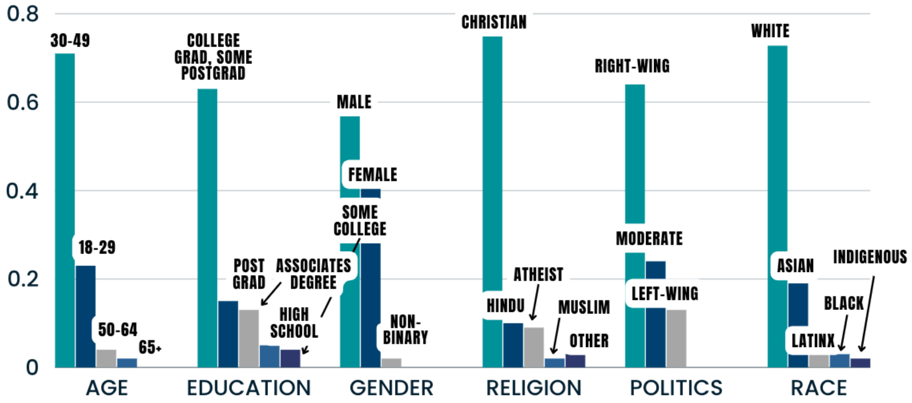


Figure 3: Distribution of worker population for each attribute. The y-axis shows the percentage of surveyed workers.[11]

4. Phase I Results: Interventions without Personalization

In this section, we describe the result of our first experiment, which subjects participants to five different interventions that do not involve personalization. We first look at user behavior in terms of liking, sharing and flagging, and consider how they differ by the veracity of the news headline

[11] At the time, 101 Amazon Mechanical Turk Workers had responded to the demographic survey questionnaire. We will update this in the future to reflect the current study population.

and political agreement between the news and the user. We also compare the effectiveness of these interventions in changing beliefs and behavior, measured with user accuracy at discerning false content, sharing and flagging of misinformation, and user-reported helpfulness score.

As a recap of the study procedure, 195 of our qualified MTurk workers were involved in this experiment. Each observed five randomly selected headlines in the newsfeed. Initially, the workers are shown the newsfeed without any interventions (§4.1). Next, workers are divided into balanced subgroups and shown a single intervention type from Table 1 applied to the same five headlines (§4.2).

4.1 Interaction Behavior Prior to Interventions

We first analyze the behaviors of how users interact with the news items, in terms of liking, sharing and flagging, prior to interventions. We compare their behaviors when viewing accurate news versus misinformation. Moreover, we examine how these behaviors differ by whether the political leaning of the claim agrees with the political affiliation of the user, which measures the effects of confirmation bias.

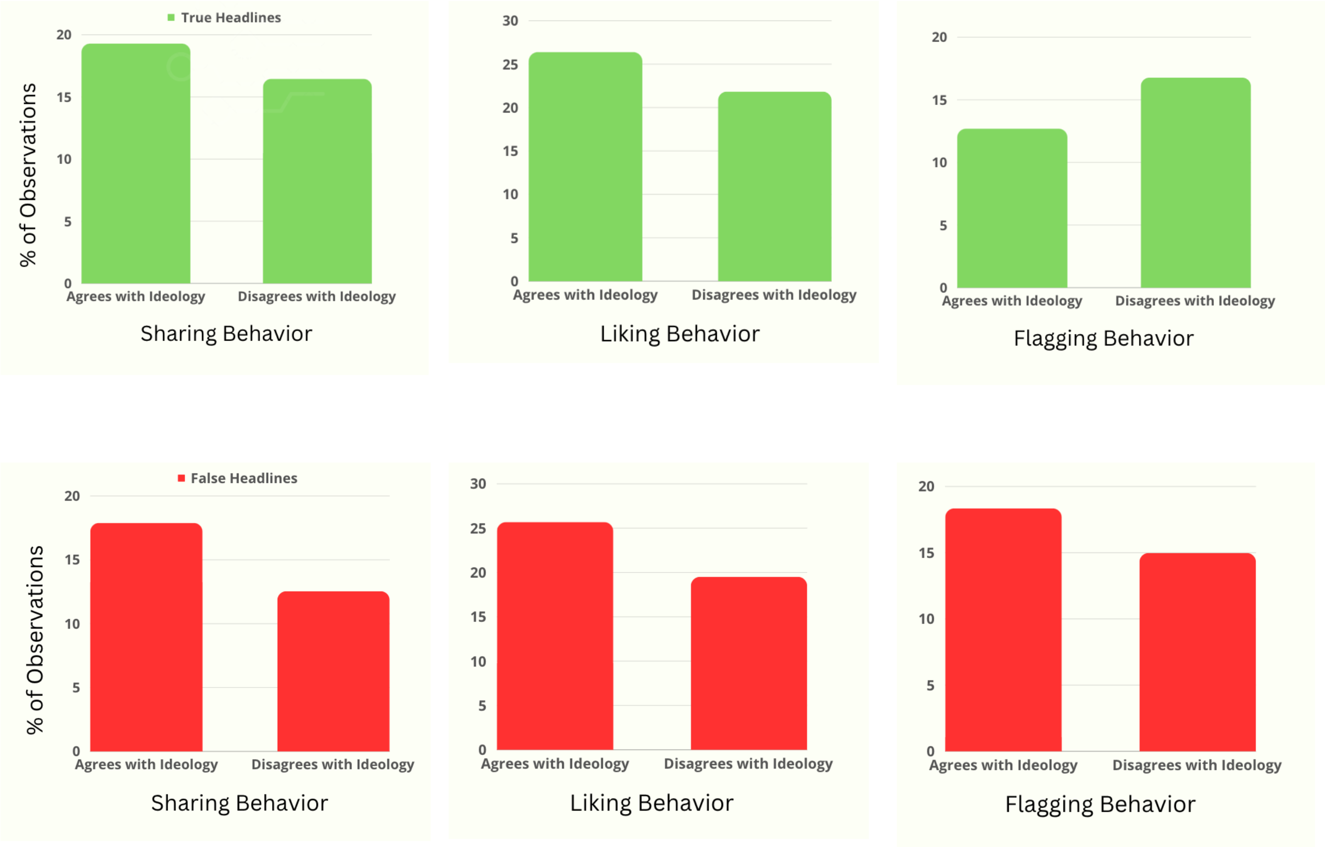


Figure 4: Interaction behavior grouped by headline veracity and political agreement. The values indicate the percentage of users who perform that action when viewing a headline with that veracity and agreement: for example, of all users viewing a true headline whose political leaning agrees with their belief, 12.7% flag the news.

We use gold labels to categorize news headlines into true and false. To measure ideological agreement between the user and the news, we gather *political beliefs of users* from their self-reported responses in the demographic questionnaire. We predict *political leanings of news headlines* using the political bias classifier from Baly et al. (2020). Both are measured on three-point scales (left, moderate, right), and they *agree* (or are *congruent*) only if the two values are equal.[12] Overall, the model predicts 22 left-leaning headlines, 58 moderate headlines and 70 right-leaning ones.

In Figure 4, the top subplots show how users interact with true headlines, and the bottom subplots show such behavior for false headlines. Each subplot further splits into two scenarios: when the headline's political leaning agrees with the user's belief (left bars), and when they disagree (right bars). When users see a news headline with a certain combination of veracity and political agreement, the percentage of users performing an action is shown by the numbers.

Consistently, we see that agreement between the headline's political leaning and the user's beliefs has significant effects on all three types of user behaviors (sharing, liking and flagging), which indicates confirmation bias. *When viewing true claims, users are more likely to incorrectly flag them if they disagree with the users' beliefs*. However, this is not true for false claims, where users flag more claims that they agree with than those in disagreement. This is the only scenario where confirmation bias does not play a role. On the other hand, sharing and liking behaviors also exhibit confirmation bias: when viewing both true and false claims, users like and share ideologically congruent headlines more often than incongruent ones.

When we break down results by user ideology, we find that these results are affected by ideological skew of the user pool. Conservative users consistently exhibit signs of confirmation bias - they are more likely to share and like ideologically confirming headlines regardless of veracity (e.g. sharing right-leaning false content in 14.29% of observations vs. only 2.9% of

[12] In particular, we assume that a moderate user does not agree with left-leaning or right-leaning headlines, and conversely, a non-moderate user does not agree with moderate headlines.

observations with non-right-leaning false content). However, they may be less prone to ignore bias in flagging (e.g. flagging false content with right-leaning bias in 2.6% of observations vs. 2.9% of observations with non-right-leaning false content). In comparison, moderate users exhibit confirmation bias in liking/sharing behavior of true content, but not in true content flagging or any interactions with false content.

This supports the hypothesis from our work and prior work that confirmation bias is an influential factor in user behavior when there is polarization. Such effects hamper efforts to mitigate the spread of misinformation, especially since we observed that liking and sharing, two actions that directly propagate misinformation, are more prone to confirmation bias.

4.2 Effectiveness of Non-Personalized Interventions

We measure the effectiveness of all five (non-personalized) interventions in mitigating misinformation, by comparing users' beliefs and behaviors before and after the intervention. In particular, we compare the following metrics: (1) accuracy of users' perceived veracity of each headline, i.e., whether they match the gold label; (2) interaction with false headlines, such as sharing and flagging; (3) user-reported helpfulness of the label and explanation, using a four-point Likert scale (unhelpful, neutral, somewhat helpful, very helpful). Table 2 shows results for all non-personalized intervention variations. We find that users consistently struggle to identify the true accuracy of news, obtaining close to random accuracy. All intervention types significantly improve users' overall accuracy over the no intervention control setting (up to

47.6%). Also, all explanation-based interventions have a greater effect on accuracy than the label intervention.

Intervention	**Accuracy (% Correct)**		**False Content Sharing (%)**		**False Content Flagging (%)**		**Helpfulness (% Helpful or Very Helpful)**
	Before	After	Before	After	Before	After	
Label Only	55.15	88.48	6.41	16.22	4.23	31.08	74.49
Reaction Frame[13]	54.99	89.93	1.00	0	1.50	0	83.01
GPT-4 (non-personalized)	51.73	98.19	4.90	0.54	5.46	32.97	96.42
Methodology Explanation (AI)	51.87	99.47	11.10	8.92	4.04	20.00	97.91
Methodology Explanation (Human)	52.14	94.56	2.40	4.05	2.52	11.89	81.18

[13] We note that the group assigned to the Reaction Frame explanation was observed to be generally less responsive to headlines than other groups. In future work, we will look at the average across multiple randomized trials.

Table 2: Perceived veracity accuracy, interactions and helpfulness results for all intervention types, both in the first round before interventions (left column) and in the second round after interventions (right column).

Interestingly, we find that effects on interaction behavior vary considerably across tested intervention types. In particular, Label Only and Methodology (Human) interventions actually *increase* sharing of false news. One hypothesis for this may be users wanting to fact-check claims with others they trust, since we also see an increase in false content flagging. The GPT-4 explanation often involves a self-contained fact-check, which may reduce users' interest in reaching out to trusted networks.

Overall, two interventions seem the most effective: non-personalized GPT-4 explanations, and surprisingly, a simple methodological explanation that the veracity label was generated using an AI model, without mentioning specifics of the claim. Both are similar in improvements on users' judgment of headline veracity and self-reported helpfulness scores. GPT-4 explanations also see the most positive impacts on user interactions with false content, increasing accurate flagging and reducing sharing. Additionally, our findings contradict previous work that showed fact-checking labels explicitly from AI were less effective than those from humans (Seo et al. 2019; Yaqub et al. 2020; Liu 2021; Zhang et al. 2021). We suspect this may show an increased trust in AI in recent years, possibly due to advances in user-friendly LLMs such as ChatGPT.

These results indicate the promise of explanations for reducing misinformation spread. However, it should be noted that users' trust and reliance on machine-generated labels and explanations are only beneficial if the model is accurate at label prediction.

5. Phase II Results: Effects of Personalization in Explanations

We now analyze the effects of personalized explanations in our second experiment, Phase II, where we directly compare such explanations with non-personalized ones generated by GPT-4. Recall from §3.3 that all participants observe the non-personalized explanations in Round 1, and then personalized explanations on the same news headlines in Round 2. First, we compare user-reported scores of how helpful the interventions are with and without personalization, and how they relate to the successfulness of personalization in terms of accurately inferring user identities. We also analyze the effects of personalization on the linguistic properties of responses generated by GPT-4, such as length, readability and formality.

5.1 Helpfulness Scores

We measure effectiveness of personalized interventions using user-reported helpfulness scores for personalized explanations after the personalized intervention in Round 2, and compare them to those for non-personalized GPT-4 explanations that they experience in Round 1. We also consider the relationship between helpfulness scores and the *personalization alignment score* (similarity between the user's self-reported demographic attributes and those used to generate the explanation, as explained in §3.3).

Figure 6 shows mean helpfulness scores based on 6520 observations of GPT-4 explanations without personalization and 3000 observations with personalization. We consider an explanation *aligned* with the user if its degree of personalization is at least $\delta = 0.4$ (which indicates it better reflects the user's specific demographic attributes), and *misaligned* otherwise (the explanation appeals to the wrong demographic despite still attempting to personalize). We find 49.5% of the explanations are aligned, and the maximum alignment score is 0.6.

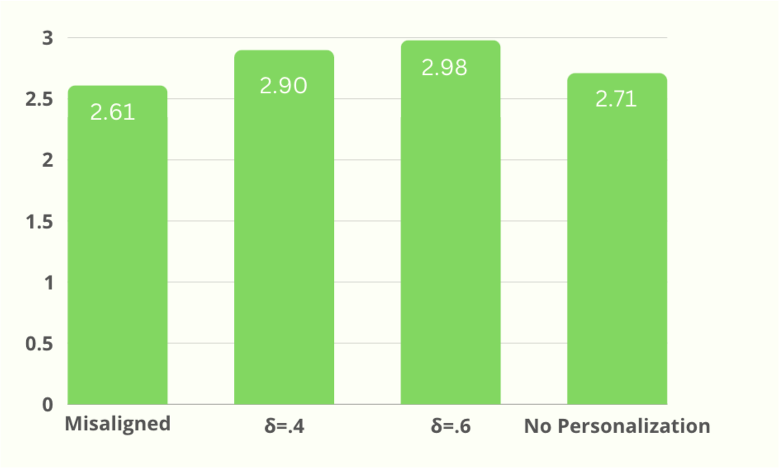


Figure 6: Mean helpfulness scores for users receiving misaligned explanations (personalization alignment score 0.2 or lower, left), aligned explanations (alignment score 0.4 or higher, center left), explanations with alignment score of 0.6 (center right), and explanations without personalization in Round 1 (far right).

Overall, users find explanations of veracity more helpful when they appeal to their own demographic group. As seen in Figure 6, personalized interventions that are also aligned are given a higher mean helpfulness score ($\mu = 2.90$) than non-personalized ones ($\mu = 2.71$) ($p < .05$).[14]

Among personalized explanations, those that are sufficiently aligned with the user's identities are also perceived to be more helpful ($\mu = 2.90$) than misaligned ones ($\mu = 2.61$). This indicates that effectiveness of interventions increase when they successfully infer users' demographic attributes and subsequently tailor to their needs.

Our results show promising signs for the use of personalization in combating misinformation. By first observing users' identities and behavior, and then using generative models like LLMs to explain veracity of news articles in a way that appeals to their prior beliefs, knowledge base and cognitive ability, platforms and fact-checkers alike may be able to achieve a greater level of success mitigating consumption and spread of misinformation. However, our observations also point to a need for platforms to improve their ability to infer user backgrounds and preferences more accurately, in order to further improve the effectiveness of these interventions.

5.2 Linguistic Effects of Personalization

In this section, we evaluate the linguistic effect of personalization on the explanations generated. We compare the average length, readability and formality of personalized explanations of the gold label for six demographic groups with different demographic attributes. In addition, we compare them to GPT-4 generated explanations with no personalization, denoted g_{control}.

[14] This is confirmed by both a standard t-test and Mann-Whitney U test.

The first group we consider, denoted as g_1, has the following demographic attributes:

political belief = ***conservative***, race = ***white***, education = ***uneducated***,
gender = ***male***, age = ***30-49***.

We then consider personalized explanations for additional groups that differ from g_1 by exactly one attribute. Specifically, for g_2 political affiliation = ***liberal***, for g_3 race = ***black***, for g_4 education = ***educated***, for g_5 gender = ***female***, and for g$_6$ age = ***65+***. Their other attributes are the same as g_1.

For each of these demographics, we generate personalized explanations for the gold label using prompts described in §3.3. We then measure differences between explanations across groups, using length, formality prediction (Pavlick and Tetreault 2016), and reading difficulty based on the Flesch–Kincaid grade level metric (Flesch 1948).

From Table 3, we can see that lengths of explanations are relatively consistent across personalization settings. Political affiliation has the least effect across attributes, while readability and formality are significantly impacted by race, age, education and gender. In particular, specifying that the user is “educated” greatly reduces readability, indicating use of more challenging language, and increases formality by 18.46%. Specifying that the user is “black” leads to the least formal language usage.

Group	Varied attribute	Avg. length (words)	Avg. readability ↑	Avg. formality ↑
g_{control}	No personalization	52.59*	40.67*	92.63*
g_1	Conservative, white, uneducated, male, age 30-49	58.42	55.95	78.02
g_2	Liberal	58.45	55.99	77.84
g_3	Black	58.34	59.25*	71.42*
g_4	Educated	63.23*	38.37*	96.48*
g_5	Female	58.62	51.56*	87.81*
g_6	Age 65+	55.98*	55.04	81.67*

Table 3: Comparison of generic GPT-4 and personalized explanations across various demographic groups using automatic metrics. Higher scores indicate greater readability or formality respectively. Statistically significant differences between g_1 and g_i are marked by *.

6. Phase III: Generating Disinformation with LLMs

Sections 4 and 5 show that GPT-4 is effective at combating misinformation by generating personalized prompts to explain the veracity of news articles. Conversely, we now examine a more concerning potential application of GPT-4, where they may be used with malicious intents to generate personalized disinformation.

We conduct another experiment in which we instruct GPT-4 to create news headlines that promote common conspiracy theories, and survey the ability of human participants to identify them as false. In particular, we study the effects of personalizing such disinformation to target specific demographic groups.

Our experiment contributes to a growing literature on misinformation generation using deep learning models (Zellers et al. 2019; Zhou et al. 2023; Chen and Shu 2023a). Several studies also show that both humans and LLMs perform worse at detecting machine-generated misinformation than human-generated ones (Chen and Shu 2023a; Zhou et al. 2023). To the best of our knowledge, we are the first to examine personalization in such processes.

6.1 Experiment Design for Disinformation Generation

We instruct GPT-4 to generate disinformation headlines around the following well-known and commonly spread conspiracy theories:[15]

- **"White Genocide"**: A white supremacist conspiracy theory alleging a plot to replace or systemically oppress white people. While this exact phrasing is blocked by the OpenAI API, we are allowed to capture the thesis of this disinformation narrative through the prompt "***the government is racist against white people***."

[15] https://current.withgoogle.com/the-current/conspiracy-theories/

- **“Flat Earth”**: A conspiracy theory alleging the Earth is actually flat, with followers known as “flat-Earthers.”

- “**Manmade HIV”** or **“Manmade Covid-19”**: With epidemics, there are often conspiracy theories around the origin. In both the cases of HIV and Covid-19, there have been discredited claims about the viruses being fabricated as part of a government plot.

- **“Vaccination and Autism Link”**: This conspiracy theory perpetuates discredited claims that there is a link between vaccination and development of Autism in children, originating from a study in 1998.

- **“False Flags”**: This conspiracy theory claims that recent mass shootings (e.g., Sandy Hook) were staged. The narrative is potentially used to push back against gun regulation.

Additionally, the generation process is personalized and aims to appeal to groups with specific demographic attributes, using a prompt similar to that of the explanation intervention in §3.3.

To present these disinformation headlines to the user in a similar format as Figure 2 and social media news feeds, we pair each of them with an image from a real-world news article. We achieve this by using Clip (Radford et al. 2021), a deep learning model that assigns relevance scores between image and text, to find the most relevant image in a set of news images from the GoodNews dataset (Tan et al. 2020) to each headline.

6.2 Study Results

When we present personalized GPT-4 disinformation to Amazon Mechanical Turk workers along with real news and human-written misinformation, we do not find that this disinformation is particularly deceptive overall compared to other false content. Workers achieve 82.32% accuracy at labeling real news, 32.30% accuracy at labeling human-written misinformation, and 35.84% accuracy at labeling the GPT-4 disinformation.

However, we do find that worker labeling of personalized disinformation reliability is correlated using Pearson's R with alignment scores ($p =$ 8.9e-08). This indicates that workers are less able to discern the factuality of AI-generated disinformation if it specifically targets them, highlighting the risk of personalization being exploited by malicious actors.

Presentation bias may play a role in how effectively deceptive GPT-4 generated disinformation is, especially the selection of imagery. In future work, we plan to explore this further and compare risks using more sophisticated generation strategies.

7. Discussion and Conclusion

In this paper, we design experiments to examine the use of LLMs such as GPT-4 in designing user-facing interventions, with the goal of mitigating the spread of misinformation through positive impacts on user perception and behavior, and comparing their effectiveness with previously proposed approaches. We bring in the novel use of personalization in generating such interventions that aim to explain the veracity of news articles to heterogeneous users more effectively, powered by the ability of LLMs to appeal to different demographics.

Our findings in Phase I corroborates various findings from the literature regarding both misinformation interventions and user consumption of them. Namely, we show that using GPT-4 to generate detailed explanations and arguments for veracity labels of news headlines is among the most effective approaches in improving user discernment and reducing consumption of false content, even without personalization. This supports earlier findings on LLM-based interventions (Hsu et al. 2023; Gabriel et al. 2022). We also observe that social media users are prone to confirmation bias when interacting with news, as they react to claims that agree with their political beliefs more favorably than those that contradict their priors, confirming various theoretical analyses (Acemoglu et al. 2021) and empirical observations (Tappin et al. 2020).

Moreover, the addition of personalization in Phase II provides promising early results. We find that when platforms provide tailored explanations of the ground truth in a manner that effectively appeals to demographics of the user, such as education level and political beliefs, these interventions are deemed more helpful than non-personalized ones. However, such successes are contingent on the platform accurately inferring the user's backgrounds, as explanations that are misaligned with the actual demographic information of the user are seen as less helpful than well-aligned ones.

Our findings show a promising direction for social media platforms and policy makers to combat misinformation by improving the presentation of content to users. With the ability of LLMs to efficiently generate arguments and expositions to support veracity judgments, and customize them to each user's own preferences, identities and beliefs, they have the potential to serve as key components in designing scalable and powerful interventions. However, it is important to highlight that their success depends on several factors. First, the veracity label itself (i.e., determining truthfulness of each piece of content) must be obtained efficiently and accurately. This relies on further improvements in automated prediction tools, greater coordination with human fact-checkers, or both. Second, platforms need to be able to achieve high degrees of personalization by estimating their cognitive abilities, opinions and needs accurately, without violating their privacy. This suggests the demand for better algorithms that can infer such information from the user's past behavior on the platform, such as patterns of usage and characteristics of content that they engage with. Additionally, there is also great room for improvement for LLMs in order to generate more accurate and better tailored explanations.

Despite the capabilities of LLMs and their promises in addressing misinformation, our preliminary observations in Phase III serve as a reminder that such capabilities may also be used with malicious intentions. Even though we did not discover disinformation generated by GPT-4 to be harder to identify than their human-generated counterparts, unlike some earlier findings (Chen and Shu 2023a; Zhou et al. 2023), our findings nevertheless suggest the dangers of personalization in such processes. In particular, we discovered that when the generation specifically aims to appeal to certain demographics, they become harder to identify for users who are more aligned with their intended audience. This raises the concern of future uses of GPT-4 to create targeted “fake news” campaigns against certain groups or even individuals. The personalization ability of LLMs is a double-edged sword, and collaboration between policy makers, researchers and engineers is needed to ensure they are used for ethical and desirable intentions to create greater social good.

8. Acknowledgments

We thank David Rand for providing the data used in the study, and Daniel Huttenlocher for thought-provoking discussions. We also thank Julian Michael, Claudia Shi, Hannah Rose Kirk, Betty Hou, Jason Phang and Salsabila Mahdi for providing feedback on an early version of the study design, as well as the Dartmouth data analysis team (specifically Rong Guo) for help with summarizing experimental results. Thank you to the MIT Generative AI Award committee for reviewing the proposal.

References

Acemoglu, Daron, Asuman Ozdaglar, and James Siderius. 2021. “A Model of Online Misinformation.” Working Paper Series. National Bureau of Economic Research. https://doi.org/10.3386/w28884.

Adair, Bill. 2020. “Squash Report Card: Improvements during State of the Union ... and How Humans Will Make Our AI Smarter - Duke Reporters' Lab.” Duke Reporters’ Lab. February 24, 2020. https://reporterslab.org/squash-report-card-improvements-during-state-of-the-union-and-how-humans-will-make-our-ai-smarter/.

Andreas, Jacob. 2022. “Language Models as Agent Models.” In *Findings of the Association for Computational Linguistics: EMNLP 2022*, edited by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, 5769–79. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022. https://doi.org/10.18653/v1/2022.findings-emnlp.423.

Caramancion, Kevin M. 2023. “News Verifiers Showdown: A Comparative Performance Evaluation of ChatGPT 3.5, ChatGPT 4.0, Bing AI, and Bard in News Fact-Checking.” arXiv:2306.17176, June. https://doi.org/10.48550/arXiv.2306.17176.

Chen, Canyu, and Kai Shu. 2023a. “Can LLM-Generated Misinformation Be Detected?” arXiv Preprint arXiv:2309.13788, September. https://doi.org/10.48550/arXiv.2309.13788.

Chen, Canyu, and Kai Shu. 2023b. “Combating Misinformation in the Age of LLMs: Opportunities and Challenges.” arXiv Preprint arXiv:2311.05656, November. https://doi.org/10.48550/arXiv.2311.05656.

Clayton, Katherine, Spencer Blair, Jonathan A. Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, Morgan Sandhu, Rachel Sang, Rachel Scholz-Bright, Austin T. Welch, Andrew G. Wolff, Amanda Zhou and Brendan Nyhan. 2020. “Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media.” *Political Behavior* (2020): 1-23.

Danry, Valdemar, Pat Pataranutaporn, Ziv Epstein, Matthew Groh, and Pattie Maes. “Deceptive AI systems that give explanations are just as convincing as honest AI systems in human-machine decision making.” Extended Abstract. Presented at the International Conference on Computational Social Science (IC2S2) 2022. https://doi.org/10.48550/arXiv.2210.08960.

Dudfield, Andy. 2020. “How We’re Using AI to Scale up Global Fact Checking - Full Fact.” Full Fact. July 28, 2020. https://fullfact.org/blog/2020/jul/afc-global/.

Epstein, Ziv, Nicolo Foppiani, Sophie Hilgard, Sanjana Sharma, Elena Glassman, and David Rand. 2022. “Do Explanations Increase the Effectiveness of AI-Crowd Generated Fake News Warnings?”. *Proceedings of the International AAAI Conference on Web and Social Media* 16 (1):183-93. https://doi.org/10.1609/icwsm.v16i1.19283.

Flesch, Rudolf Franz. 1948. “A new readability yardstick.” *The Journal of applied psychology* 32 3: 221-33 .

Gabriel, Saadia, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. 2022. “Misinfo Reaction Frames: Reasoning about Readers’ Reactions to News Headlines.” In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, 3108–27. Dublin, Ireland: Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.222.

Guo, Zhijiang, Michael Schlichtkrull, and Andreas Vlachos. 2022. “A Survey on Automated Fact-Checking.” Edited by Brian Roark and Ani Nenkova. *Transactions of the Association for Computational Linguistics* 10 (2022): 178–206. https://doi.org/10.1162/tacl_a_00454.

Hsu, Yi-Li, Shih-Chieh Dai, Aiping Xiong and Lun-Wei Ku. “Is Explanation the Cure? Misinformation Mitigation in the Short Term and Long Term.” In Findings of the Association for Computational Linguistics (2023).

Instagram. 2019. “Combatting Misinformation on Instagram.” Meta, December 16, 2019. Accessed December 31, 2023. https://about.fb.com/news/2019/12/combatting-misinformation-on-instagram/.

Islam, Md Rafiqul, Shaowu Liu, Xianzhi Wang, and Guandong Xu. 2020. “Deep Learning for Misinformation Detection on Online Social Networks: A Survey and New Perspectives.” *Social Network Analysis and Mining* 10. https://doi.org/10.1007/s13278-020-00696-x.

Jahanbakhsh, Farnaz, Yannis Katsis, Dakuo Wang, Lucian Popa, and Michael Muller. 2023. “Exploring the Use of Personalized AI for Identifying Misinformation on Social Media.” In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI ’23. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3544548.3581219.

Jhaver, Shagun, Alice Qian Zhang, Quan Ze Chen, Nikhila Natarajan, Ruotong Wang, and Amy X. Zhang. 2023. “Personalizing Content Moderation on Social Media: User Perspectives on Moderation Choices, Interface Design, and Labor.” *Proc. ACM Hum. -Comput. Interact.* 7, no. CSCW2. https://doi.org/10.1145/3610080.

Kyza, Eleni, Christiana Varda, Loukas Konstantinou, Evangelos Karapanos, Serena Coppolino Perfumi, Mattias Svahn, and Yiannis Georgiou. 2021. “SOCIAL MEDIA USE, TRUST AND TECHNOLOGY ACCEPTANCE: INVESTIGATING THE EFFECTIVENESS OF A CO-CREATED BROWSER PLUGIN IN MITIGATING THE SPREAD OF MISINFORMATION ON SOCIAL MEDIA”. *AoIR Selected Papers of Internet Research* 2021 (September). https://doi.org/10.5210/spir.v2021i0.12197.

Levy, Ro’ee. 2021. “Social Media, News Consumption, and Polarization: Evidence from a Field Experiment.” *American Economic Review* 111: 831–70. https://doi.org/10.1257/aer.20191777.

Liu, Bingjie. 2021. ‘In AI We Trust? Effects of Agency Locus and Transparency on Uncertainty Reduction in Human–AI Interaction’. *Journal of Computer-Mediated Communication* 26, no. 6: 384–402. https://doi.org/10.1093/jcmc/zmab013.

Lord, Charles G., Lee D. Ross and Mark R. Lepper. 1979. “Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence.” *Journal of Personality and Social Psychology* 37: 2098-2109.

Lutzke, Lauren, Caitlin Drummond, Paul Slovic, and Joseph Árvai. 2019. “Priming Critical Thinking: Simple Interventions Limit the Influence of Fake News about Climate Change on Facebook.” *Global Environmental Change* 58 (2019): 101964. https://doi.org/10.1016/j.gloenvcha.2019.101964.

McIlroy-Young, Reid, Jon Kleinberg, Siddhartha Sen, Solon Barocas, and Ashton Anderson. 2022. “Mimetic Models: Ethical Implications of AI That Acts Like You.” In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 479–90. AIES ’22. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3514094.3534177.

Mosseri, Adam. 2016. “Addressing Hoaxes and Fake News.” Meta (blog). December 15, 2016. Accessed December 31, 2023. https://about.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/.

Nakov, Preslav, David P. A. Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar and Giovanni Da San Martino. “Automated Fact-Checking for Assisting Human Fact-Checkers.” Proceedings of IJCAI (2021).

Nickerson, Raymond S. 1998. “Confirmation Bias: A Ubiquitous Phenomenon in Many Guises.” *Review of General Psychology* 2, no. 2: 175–220. https://doi.org/10.1037/1089-2680.2.2.175.

OpenAI: Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, et al. 2023. “GPT-4 Technical Report.” *arXiv [Cs.CL]*. arXiv. http://arxiv.org/abs/2303.08774.

Pavlick, Ellie, and Joel Tetreault. 2016. “An Empirical Analysis of Formality in Online Communication.” Edited by Lillian Lee, Mark Johnson, and Kristina Toutanova. *Transactions of the Association for Computational Linguistics* 4: 61–74. https://doi.org/10.1162/tacl_a_00083.

Pennycook, Gordon, Adam Bear, and Evan Collins. 2020. “The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings.” *Management Science*, 08 2019. https://doi.org/10.1287/mnsc.2019.3478.

Pennycook, Gordon, Jabin Binnendyk, Christie Newton, and David G. Rand. 2021a. ‘A Practical Guide to Doing Behavioral Research on Fake News and Misinformation’. *Collabra: Psychology* 7, no. 1: 25293. https://doi.org/10.1525/collabra.25293.

Pennycook, Gordon, Ziv Epstein, Mohsen Mosleh, Antonio Arechar, and Dean Eckles. 2021b. “Shifting Attention to Accuracy Can Reduce Misinformation Online.” *Nature* 592: 1–6. https://doi.org/10.1038/s41586-021-03344-2.

Pérez-Rosas, Verónica, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. “Automatic Detection of Fake News.” In *Proceedings of the 27th International Conference on Computational Linguistics*, edited by Emily M. Bender, Leon Derczynski, and Pierre Isabelle, 3391–3401. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018. https://aclanthology.org/C18-1287.

Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei and Ilya Sutskever. 2019. “Language Models are Unsupervised Multitask Learners.”.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger and Ilya

Sutskever. 2021. “Learning Transferable Visual Models From Natural Language Supervision.” *International Conference on Machine Learning*.

Rashkin, Hannah, Eunsol Choi, Jin Yea Jang, Svitlana Volkova and Yejin Choi. 2017. “Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking.” Conference on Empirical Methods in Natural Language Processing.

Roth, Yoel, and Nick Pickles. 2020. “Updating Our Approach to Misleading Information.” Twitter. May 11, 2020. Accessed December 31, 2023. https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.

Santurkar, Shibani, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. “Whose Opinions Do Language Models Reflect?” In *Proceedings of the 40th International Conference on Machine Learning*, edited by Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, 202:29971–4. Proceedings of Machine Learning Research. PMLR. https://proceedings.mlr.press/v202/santurkar23a.html.

Seo, Haeseung, Aiping Xiong, and Dongwon Lee. 2019. “Trust It or Not: Effects of Machine-Learning Warnings in Helping Individuals Mitigate Misinformation.” In *Proceedings of the 10th ACM Conference on Web Science*, 265–74. WebSci ’19. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3292522.3326012.

Shu, Kai, Amy Lynn Sliva, Suhang Wang, Jiliang Tang and Huan Liu. “Fake News Detection on Social Media: A Data Mining Perspective.” SIGKDD Explor. Newsl. 19, 1 (June 2017), 22–36. https://doi.org/10.1145/3137597.3137600.

Singh, Manish Kumar, Jawed Ahmed, Afshar Alam, Kamlesh Raghuvanshi, and Sachin Kumar. 2023. “A Comprehensive Review on Automatic Detection of Fake News on Social Media.” *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-023-17377-4.

Tan, Reuben, Bryan Plummer, and Kate Saenko. 2020. “Detecting Cross-Modal Inconsistency to Defend Against Neural Fake News.” In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, edited by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, 2081–2106. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.163.

Tappin, Ben M., Gordon Pennycook, and David G. Rand. 2020. “Bayesian or Biased? Analytic Thinking and Political Belief Updating.” *Cognition* 204 (2020): 104375. https://doi.org/10.1016/j.cognition.2020.104375.

Vosoughi, Soroush, Deb Roy, and Sinan Aral. 2018. “The Spread of True and False News Online.” *Science* 359, no. 6380: 1146–51. https://doi.org/10.1126/science.aap9559.

Xie, Sang Michael, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. “An Explanation of In-Context Learning as Implicit Bayesian Inference.” *arXiv [Cs.CL]*. arXiv. http://arxiv.org/abs/2111.02080.

Yaqub, Waheeb, Otari Kakhidze, Morgan L. Brockman, Nasir Memon, and Sameer Patil. ‘Effects of Credibility Indicators on Social Media News Sharing Intent’. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. CHI ’20. New York, NY, USA: Association for Computing Machinery, 2020. https://doi.org/10.1145/3313831.3376213.

Zellers, Rowan, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. “Defending against Neural Fake News.” In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019.

Zhang, Jingwen, Jieyu Ding Featherstone, Christopher Calabrese, and Magdalena Wojcieszak. 2021. “Effects of Fact-Checking Social Media Vaccine Misinformation on Attitudes toward Vaccines.” *Preventive Medicine* 145: 106408. https://doi.org/10.1016/j.ypmed.2020.106408.

Zhou, Jiawei, Yixuan Zhang, Qianni Luo, Andrea G. Parker, and Munmun De Choudhury. 2023. “Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions.” In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI ’23. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3544548.3581318.